The style characteristic of China's stock market: an application to PCA for interval symbolic data

Dingmu Cao*,**, Wen Long*,**,***

*Graduate University of Chinese Academy of Sciences, Beijing, 100190, China
**Research Center on Fictitious Economy & Data Science, Chinese Academy of Sciences, Beijing 100190, China
***Corresponding author: longwen@ucas.ac.cn

Abstract. By applying the symbolic principal component analysis (SPCA) on the empirical data of the CITIC style indices in six years (2005-2010), we studied the characteristics of Chinese stock market from multiple perspectives. Two components are extracted from five variables–P/E ratio, NMC, turnover rate, return rate, and volatility–and are defined as the market performance factor and the size factor. Further, drawing the run track of the six stock style portfolios and combining with the zoom-star plots of symbolic data, we find that the Chinese stock market is excessive speculated and bounded rational.

1 Introduction

Since the beginning of China's stock market in 1990, it is developing at an alarming rate. The total market capitalization [1] of Chinese mainland stock market was beyond the total GDP [2] for the first time on August 9, 2007. As the main representative of Chinese capital market, the development of China's stock market is important for China's economic reform. There is an urgent need in an in-depth analysis of history and current situation of China's stock market development, and an exploration of the inherent regularity of the development. It is practically necessary not only for the scientific understanding in recent years, but also for decision-making of building healthier and more mature stock markets.

Researches on the stock market have already been done by both domestic and foreign scholars. Cheng (2003) studied the Chinese stock market through the classification of CITIC style index [3], discuss the risk-benefit asymmetry, and pointed out the deficiencies in the system of the Chinese stock market were the underlying causes for excessive speculation.Shi and Xu (2003) performed the technological and economic analysis on the stock market behavior, and quantified the contribution of stock investment and speculation in Chinese stock market.Huang and Sun (2008) made a literature review on stock market bubbles, and pointed out that the real

1. Market capitalization (or simply market cap) is the total value of the negotiable shares of a publicly traded company. It is equal to the share price times the number of shares outstanding.
2. Gross domestic product (GDP) is the market value of all officially recognized final goods and services produced within a country in a given period.
3. We will detail the concept of CITIC style index in the third paragraph of Part 2.

purpose of the stock market bubble was a scientific understanding of the law of accumulation of the stock market bubble and the stock market collapse mechanism, and then designed the market supervision mechanism from the perspective of investor behavior research to prevent the adverse effects of the stock market crash on the socio-economic.

Since first proposed by Diday (1988), symbolic data analysis (SDA) has acted as an effective tool for huge and complex data analysis, gaining its development in theory and wide application in empirical researches. Wang et al. (2005) made a canonical correlation analysis of the symbolic data on Chinese stock market, and proved the validity of the analytical techniques to simplify the multidimensional dynamic data system. Long et al. (2009) applied symbolic principal component analysis (SPCA) to 10 years (1996-2005) of interquartile data, finding the underlying reasons for the seemingly peculiar negative beta/return relation. They also found a trend of high speculation on growth stocks through the empirical research.

On the basis of previous research, this article will apply SPCA to six years (2005-2010) of interquartile data, confirms the over-speculation and bounded rationality of the market behavior characteristics in recent years, the same characters that founded in Cheng (2003) and Long et al. (2009). But the stock market demonstrates new features.

This paper is organized as follows. In section 2, we introduce the CITIC style indices and the advantage of them. In Section 3, we make a description of the history trends and risk-return characteristics of the stocks. In Section 4, we state the methodology of SPCA and five variables used in the following part. Section 5 presents the empirical findings, and we conclude the paper in Section 6.

2 The CITIC stock style portfolios

Since 1990, the number of listed companies in China has increased from 10 to more than 2000. During this process, it often happens that listed company merger, restructuring, rename, listing or delisting, which causes many problems to data analysis, such as the amount of data in a sample change, name of the individual change, content of the individual change, and so on. For example, if a stock corresponds to an individual, the stock name change means it's a different individual, which will cause confusion in analysis.

For these reasons, the data of Chinese stock market are huge and the data structures are complex. In order to better grasp the intrinsic relationship of the data attributes and get access to implicit knowledge in the data ocean, we analyze the Chinese stock market through the CITIC style indices.

The CITIC style indices are constructed by the Chinese International Trust and Investment Company (CITIC). They draw on the latest research results of the capital asset pricing model and practical experience of international investment, and made the necessary adjustments according to the actual situation in China. The application of CITIC style index will help to capture the behavior characteristics and dynamic law of the stock market from an overall perspective, and greatly improve the efficiency of data analysis.

The CITIC style index system consists of three scale style indices and six sub-indices. Three scale style indices are large-cap, mid-cap, and small-cap index. Further, the three scale style indices are categorized as large-cap growth index, large-cap value index, mid-cap growth index, mid-cap value index, small-cap growth index and small-cap value index.

The specific construction process of the CITIC style indices is as follows. First, determine the stocks involved in the grouping for this year; Second, calculate the average market capitalization of negotiable shares (NMC) of A-shares [4] within the sampling stocks from the first trading day of the year in January to April 30; Third, sort the average NMC of every stock in descending order, and then categorize the stocks into large-cap, mid-cap, and small-cap group; Fourth, rank the stocks in the respective large-cap, mid-cap, and small-cap category again according to the log of book-to-price ratio [5], ln(B/P). Remove the top and bottom 5% of each category, and we can calculate the mean ln(B/P). Then, the stocks with a larger ln(B/P) than the mean ln(B/P) are classified as value stocks, and those with a lower ln(B/P) are classified as growth stocks.

3 The history trends and risk-return characteristics of the stocks

We compare the trends of CITIC style indices and CSI 300 Index to get an overall view of the historical performance of the CITIC style indices. CSI 300 Index is constructed by China Securities Index Company, Ltd. It is a capitalization-weighted stock market index designed to replicate the performance of 300 stocks traded in the Shanghai and Shenzhen stock exchanges. The FIG. 1 is the running track of CSI 300 index, daily data from April 8, 2005 to December 16, 2011.

FIG. 1 – *Trends of CSI 300 Index (2005.04.08-2011.12.16).*

4. A-shares are specialized shares of the Renminbi currency that are purchased and traded on the Shanghai and Shenzhen stock exchanges.

5. Book-to-price ratio(B/P) measures a ratio to compare the current closing price of the stock by the latest quarter's book value per share.

3.1 Performance of large-cap, mid-cap, and small-cap stocks

In order to observe the movements of CITIC style indices, we convert all style indices to the same base point as the CSI 300 index, which was 1003.45 on April 8, 2005; then, we calculate the difference of each corresponding point in time series, and get the relative trend graph in FIG. 2. The specific steps are as follows. Denote $A_1, A_2, \cdots, A_{1630}$ as the daily data from April 8, 2005 to December 16, 2011 of the CSI 300 index. Denote, for example, $B_1, B_2, \cdots, B_{1630}$ as the original data of the large-cap index. To make the same base point of the data, we transform the large-cap index by multiplying A_1/B_1 , and get $B_1 * A_1/B_1, B_2 * A_1/B_1, \cdots, B_{1630} * A_1/B_1$ as the new serial data of the large-cap index. Then, we calculate the difference in each corresponding point in time series, and get $B_1 * A_1/B_1 - A_1, B_2 * A_1/B_1 - A_2, \cdots, B_{1630} * A_1/B_1 - A_{1630}$ as the serial data of the large-cap index in FIG. 2.

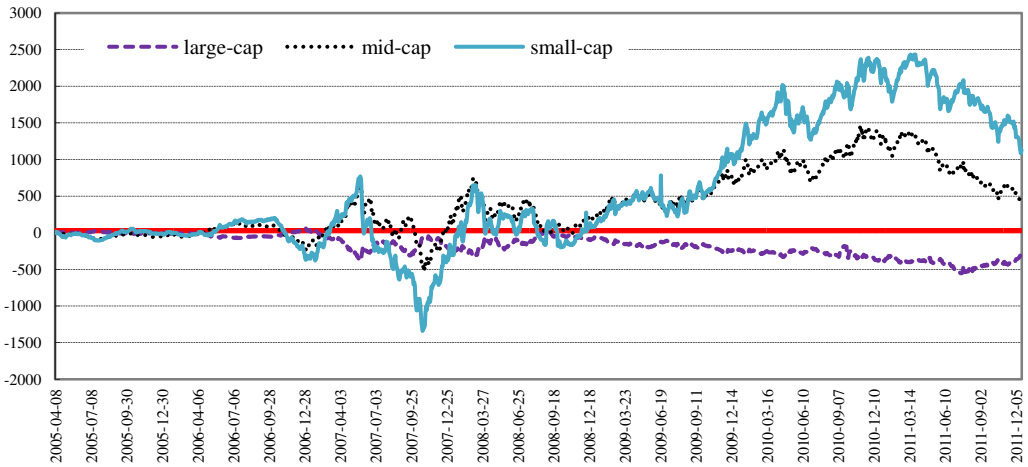

FIG. 2 – *Trends of CITIC scale style index relative to CSI 300 Index (2005.04.08-2011.12.16). NOTE: The CSI 300 index is displayed as the zero line (the red bold line in this figure) because of the transformation of the same base point and the difference calculation. With this benchmark, the other three lines represent the relative movements of the large-cap, the mid-cap, and the small-cap.*

The trends of large-cap, mid-cap, and small-cap stocks indices are basically the same as the CSI 300 index. They experienced an overall increase from 2005 to 2007, and got to the bull market peak between October 2007 and early 2008. Then, influenced by the financial crisis in 2008, the indices were all downward to the market bottom in late 2008. After that, the stock market picked up in 2009. The indices fluctuated in 2010 and continued to decline in 2011.

The CITIC scale style indices were basically the same as the CSI 300 index from 2005 to 2006, and none had the obvious advantage over others. During the bull market in 2007, the mid-cap portfolios had the best performance, followed by the large-cap index. However, the small-cap index was significantly better than other indices after August 2009, when other indices turned to parallel adjustment, and the gap between the indices became increasingly

apparent. The mid-cap, especially the small-cap portfolios showed a good defensive ability in the market crash environment during 2008, their performance is obviously better than the market's overall performance.

3.2 Risk-return characteristics of stock market

The risk-return symmetry is the cornerstone of modern financial theory. Stock investment risk comes from the uncertainty of stock price. Wang (2003) made an empirical research on China's stock market from 1998 to 2000 and pointed out that the unusual risk-return asymmetry in the Chinese stock market. In this paper, we use the data of the CITIC style indices and the CSI 300 index from 2005 to 2010 to see the risk-return characteristics of the stock market in recent years.

We draw the diagram of the risk-return characteristics as shown in FIG. 3, with the average daily return rate as the horizontal axis and volatility representing risk as the vertical axis.

From an overall perspective, we find the higher-yielding portfolios had a higher risk (such as the small-cap value stocks), and the lower-yielding had lower risk (such as the large-cap value stocks) (see FIG. 3). But CSI 300 had the lowest risk, resulting from the risk diversification effect of portfolio. In other words, the Chinese stock market has been transformed from the risk-return asymmetry before 2000 to the risk-return symmetric market recently, which reflects Chinese stock market more market-oriented and rational. However, there is a certain risk-return asymmetry in the stock market of 2005 and 2006. The small-cap stocks had the lower return and higher risk, while the large-cap stocks the other way. But after 2007, this risk-return asymmetry is reversed. In the case of the same high-risk, the return rate of the small-cap value stocks and small-cap growth stocks has increased a lot. This is also reflected in the following part 5.3, confirming the over-speculation and bounded rationality of the stock market.

4 Method and data indicators

4.1 The Symbolic Principal Component Analysis (SPCA)

The basic concept of the symbolic data is first proposed by E.Diday in 1988, in the First Conference of the International Federation of Classification Societies (IFCS). The data unit of the symbolic data can be a multi-value data, an interval data, or a distribution data. The symbolic data can greatly enrich the information content of the individual data, thereby reducing the size of the entire data system. A symbolic data consisting of $n * p$-dimensional data table is called a symbolic data table. Interval data is a common symbolic data. If $\underline{x}, \overline{x} \in R$, and $\underline{x} \leq \overline{x}$,then $[\underline{x}, \overline{x}]$ is an interval data.

The most common way to generate interval data sets establishes the sample classification and generates the range data using the concept of "data package". In accordance with certain principles of classification, a new sample collection is generated from the sample points in the raw data. The new data table in the interval data form reflects the characteristics of the performance indicators of the original data table. This method can simplify the sample space, reduce the workload of the data analysis, and improve the efficiency of data analysis. What's more, analysts can grasp more information these data contained.

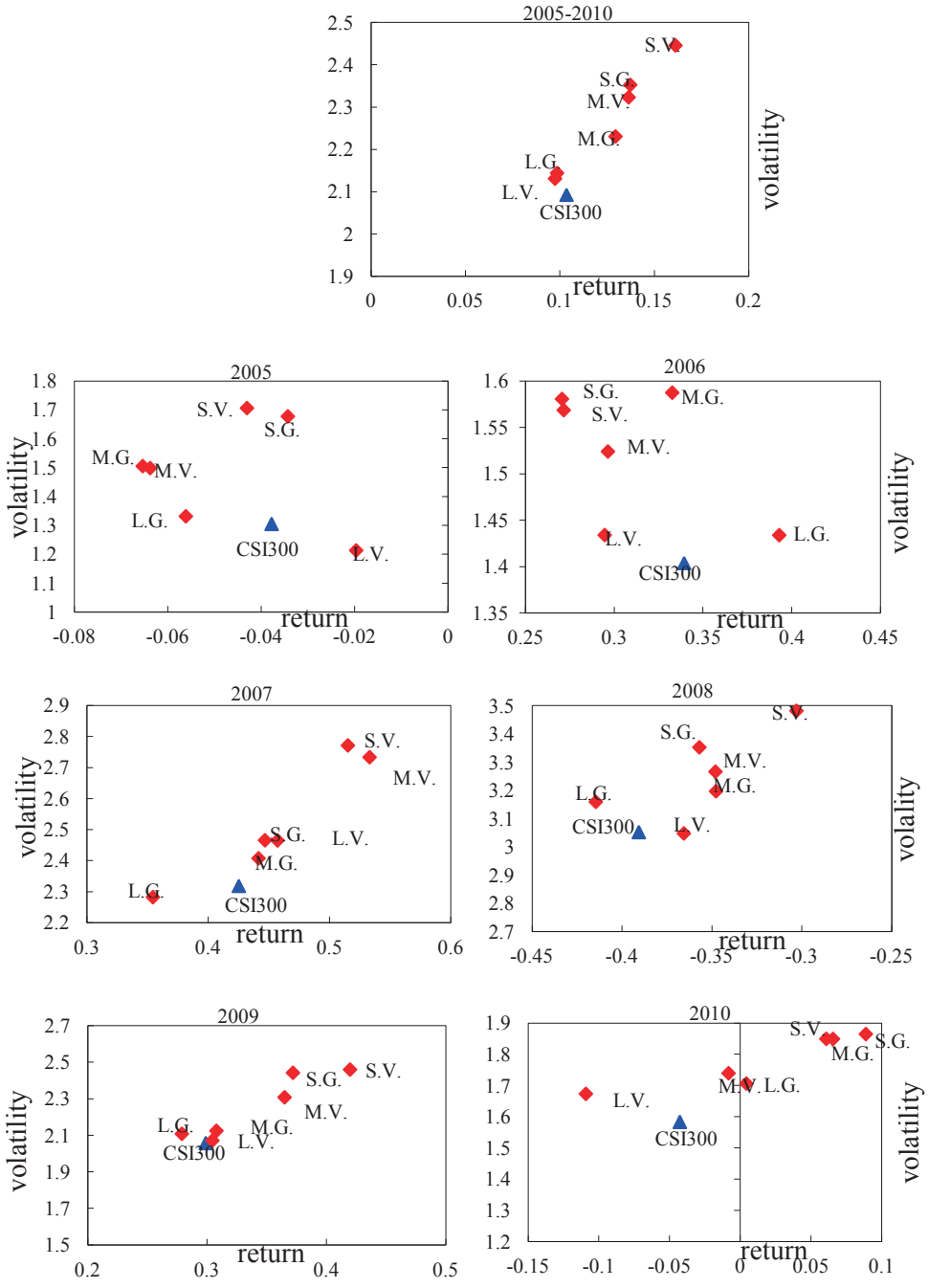

FIG. 3 – *Risk-return characteristics of Chinese stock market (2005-2010). Note: L.G. is an acronym for the large-cap growth stocks. L.V., M.G., M.V., S.G., and S.V. are the acronyms for large-cap value stocks, middle-cap growth stocks, middle-cap value stocks, small-cap growth stocks, and small-cap value stocks, respectively.*

Principal component analysis (PCA) in factor analysis is a multivariate statistical technique. Via PCA, a small set of components are extracted, and the complex data still remains the original character. Classical PCA starts with n data points $x_1, \cdots, x_n \in R^p$ in p-dimensional Euclidean space R^p. Cazes et al. (1997) and Chouakria (1998) first proposed the extensions of PCA to interval data , widely known as the centers and the vertices methods. The centers method transforms the interval data matrix (see Equation1) into a classical matrix of the interval centers, and then perform the classical PCA is on the matrix of centers. This method ignores the difference of the interval data with the same center. For example, [0, 2012] and [1005, 1007] would be treated exactly the same observations for they have the same mean, resulting in the loss of important information. Lauro and Palumbo (2000) treat the midpoint and the interval range as two separate variables, represent in a space of reduced dimensions images of interval data, and point out differences and similarities according to their structural features. The vertices method by Chouakria et al. (2000) accounts for the internal variation by using the vertices of the hyper-rectangles instead of the centers. Douzal-Chouakria et al. (2011) showed that the variance of the vertices in fact includes some but not all of the internal variation. Le-Rademacher and Billard (2012) use the so-called symbolic covariance to determine the principal component (PC) space to reflect the total variation of interval-valued data. Giordani and Kiers (2004) propose an extension of classical PCA which deals with fuzzy data (in short PCAF). However, while the fuzzy data can be regarded as a special case of interval data, they are generally different in the symbolic data; examples of the differences see Billard and Diday (2006).

We extend the data of classical PCA to the interquartile data (typical interval data), and call the new multivariate statistical technique "symbolic principal component analysis (SPCA)". Based on the vertices method, SPCA is extended as follows.

Let $i = 1, \cdots, n$ denote n objects described by p variables $Y_1, \cdots, Y_p$ of interval type. Denote $\underline{x_{ij}}, \overline{x_{ij}}$ as the upper quartile and lower quartile of the possible values of variable j for the object i, respectively. Then $\varepsilon_{ij} = [\underline{x_{ij}}, \overline{x_{ij}}]$ is the interquartile data of the possible values of variable j for the object i, and the resulting symbolic data matrix $\underline{X}$ is given by:

$$\underline{X} = \begin{pmatrix} x_1^{'} \\ \vdots \\ x_n^{'} \end{pmatrix} = \begin{pmatrix} \varepsilon_{11} & \cdots & \varepsilon_{1p} \\ \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \cdots & \varepsilon_{np} \end{pmatrix} = \begin{pmatrix} \left[\underline{x_{11}}, \overline{x_{11}}\right] & \cdots & \left[\underline{x_{1p}}, \overline{x_{1p}}\right] \\ \vdots & \ddots & \vdots \\ \left[\underline{x_{n1}}, \overline{x_{n1}}\right] & \cdots & \left[\underline{x_{np}}, \overline{x_{np}}\right] \end{pmatrix} \quad (1)$$

Similar to the classical PCA, the purpose of SPCA is to describe the object i , and the data x_i, by a reduced number $s < p$ of the new interval variables.

Denote $x_1^{'} = (\varepsilon_{i1}, \cdots, \varepsilon_{ip}) = \left(\left[\underline{x_{i1}}, \overline{x_{i1}}\right], \cdots, \left[\underline{x_{ip}}, \overline{x_{ip}}\right]\right)$ as the symbolic data vector obtained for object i. This data point is visualized in the description space R^p by a hyperrectangle R_i with 2^p vertices. Similarly, a hyperrectangle in the p-dimensional space can be described by a matrix with 2^p rows and p columns where each row contains the coordinates of one vertex of the hyperrectangle in R^p. Take $p = 2$ as an example, we can describe the data of object i by the matrix:

$$M_i = \begin{bmatrix} \underline{x_{i1}} & \underline{x_{i2}} \\ \underline{x_{i1}} & \overline{x_{i2}} \\ \overline{x_{i1}} & \underline{x_{i2}} \\ \overline{x_{i1}} & \overline{x_{i2}} \end{bmatrix} \tag{2}$$

For the general case of a dimension p, the original matrix characterized by Eq.(1) can be described by the following numerical matrix M with $n * 2^p$ rows and p columns:

$$M = \begin{pmatrix} M_1 \\ \vdots \\ M_n \end{pmatrix} = \begin{pmatrix} \begin{bmatrix} \underline{x_{11}} & \cdots & \underline{x_{1p}} \\ \vdots & \ddots & \vdots \\ \overline{x_{11}} & \cdots & \overline{x_{1p}} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} \underline{x_{n1}} & \cdots & \underline{x_{np}} \\ \vdots & \ddots & \vdots \\ \overline{x_{n1}} & \cdots & \overline{x_{np}} \end{bmatrix} \end{pmatrix} \tag{3}$$

Apply the classical CPA method to the new data matrix M with $n*2^p$ rows, and we can find a suitable dimension $s \leq p$. Let $Y_1^*, \cdots, Y_S^*$ denote the first s numerical principal components.

Then we construct the principal components of interval type $Y_1^I, \cdots, Y_S^I$ from the components $Y_1^*, \cdots, Y_S^*$ as follows.

Let L_i be the set of row indices in matrix M which refers to the matrix M_i corresponding to i^{th} symbolic object x_i. For $k \in L_i$, let y_{kv} be the value of the numerical principal component Y_v^* with row index k. The value of the interval-type principal component Y_v^I for the i^{th} object is then denoted by $y_{iv} = \left[\underline{y_{iv}}, \overline{y_{iv}}\right]$ where:

$$\underline{y_{iv}} = \min_{k \in L_i} (y_{kv}) \,\&\, \overline{y_{iv}} = \max_{k \in L_i} (y_{kv}) \tag{4}$$

4.2 Five financial variables

We use five variables to examine the health of the stock market from the perspective of expectation, market size, activity, returns and risks. The variables are price-earnings ratio (P/E ratio), negotiable market capitalization (NMC), turnover rate (monthly average level), return rate (monthly average level) and volatility (standard deviation of return).

According to the construction process mentioned in Part 2, there are usually 100 stocks in the large-cap portfolios, 300 or so in the middle-cap portfolios, and around 1000 stocks in the small-cap portfolios. Since every stock has 5 financial value in one year, we have almost 70 thousand original data in one year, that is 420 thousand in six years (2005-2010). We extract the interquartile data of the five variables for the six CITIC style portfolios in every year, and get a $6 * 5$ dimensional data table for every year (see an example in table 1).

In the form of interval data, the data can not only describe the central tendency of the original data, but can also describe the discrete range, taking full advantage of the information in observed data sets.

Considering to perform global principal component analysis of the symbolic data, the original data table was converted into a timing interval of $6 * 5 * 6$ dimensional interquartile data

	PE	NMC (billions)	Turnover (%)	Return (%)	volatility
L.G.	[20.93, 50.80]	[25.31, 49.41]	[19.89, 42.92]	[-1.84, 1.96]	[10.27, 16.88]
L.V.	[9.22, 18.92]	[27.08, 122.13]	[12.77, 23.84]	[-3.46, -2.19]	[6.51, 11.53]
M.G.	[31.23, 61.61]	[7.29, 14.61]	[30.12, 63.97]	[-0.15, 3.39]	[10.46, 15.37]
M.V.	[16.45, 37.70]	[6.92, 14.00]	[29.29, 52.96]	[-2.85, 0.70]	[9.37, 13.63]
S.G.	[36.74, 109.33]	[1.91, 4.43]	[50.76, 91.83]	[-0.34, 3.25]	[9.81, 15.09]
S.V.	[22.27, 71.24]	[2.16, 4.43]	[45.39, 75.04]	[-1.43, 1.79]	[9.70, 13.52]

TAB. 1 – *An example of the interquartile data table (year of 2010).*

table, which has eliminated the impact of outliers and kept the robustness of every category. Such transformation simplifies the original sample data into six interval samples and greatly facilitates the analysis.

5 Characteristics of the stock market behavior

5.1 Results of the symbolic principal component analysis (SPCA)

According to SPCA in Section 4.1, the interquartile data table was extended by vertices method, for every year, to a matrix M with $6 * 2^5$ rows and 5 columns. Since aiming to study the characteristics of the stock market behavior and find the running track in recent years, we employ a global principal component analysis in this study. In this way, the matrix that we use in SPCA is a combined matrix with $6 * 6 * 2^5$ rows and 5 columns. Two factors were extracted to make the result more interpretable (Factoring loading plot see FIG. 4). The cumulative contribution of the two factors is 63.44%, which is considered acceptable.

The first component has a strong positive relationship with P/E ratio, turnover rate, and return rate (see table 2). The correlation coefficients are 0.75, 0.85, and 0.74 respectively. The

	component 1	component 2
PE	0.749	0.018
NMC	-0.264	0.912
Turnover	0.845	0.022
Return	0.743	0.395
Volatility	0.508	-0.167

TAB. 2 – *Component Matrix of the global principal components analysis.*

Notes:
P/E: price earnings ratio
NMC: the market capitalization of negotiable shares
Turnover: the activity of trading of stocks
Return: the monthly return rate
Volatility: the standard deviation of the monthly return rate

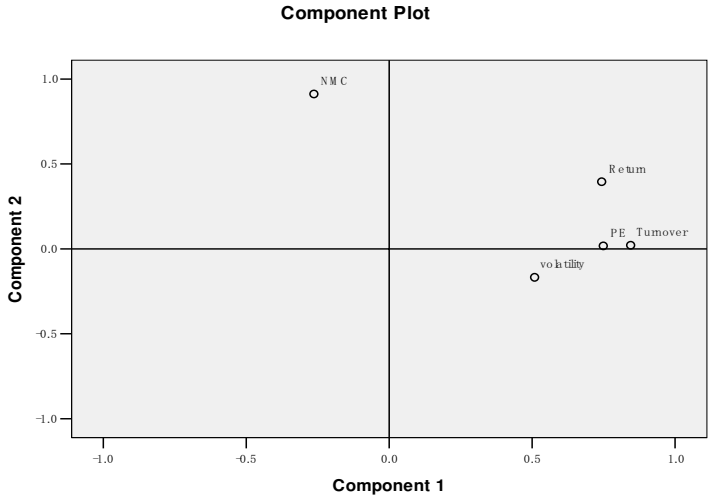

FIG. 4 – *Factoring loading plot of the global principal components analysis.*

first component and volatility have a moderate positive relationship, their correlation coefficient is 0.51. In addition, component 1 shows a weak negative relationship with NMC. Component 1 can be explained by the P/E ratio, turnover rate, return rate, and volatility. We can define the first component as the market performance factor. The second component demonstrates a strong positive correlation with NMC, and weak correlations with the other four variables. Thus we define the second component as the size factor.

5.2 Running track of the 2005-2010 stock market

All the first and second component factor scores can be derived from the global principal component analysis of empirical data from 2005-2010. Take the mean of the maximum and minimum scores for each year, six data points are produced that represent the basic situation of the stock market. Describe them in a global principal plan combined by Component 1 and Component 2, and we get the running track of the stock market as shown in FIG. 5.

The diagram reproduce the stock market's ups and downs in recent years, from the downturn market in 2005 to the prosperity in 2007, to the bear market in 2008 financial crisis, then the stock market picked up in 2009 and shocks to adjust in 2010.

In the stock market rally from 2005 to 2007, the market performance factor and the size factor are in continuous improvement, indicating that the stock market is expanding, that the market valuation of the returns and risks are improving, and that transactions become increasingly active.

In 2008, affected by financial crisis, both the performance and the size factor are in decline. It is worth mentioning that the unlock of non-tradable shares reaches a climax in 2008. The

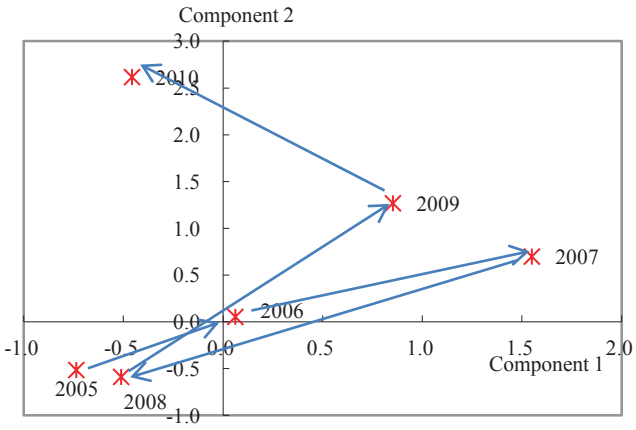

FIG. 5 – *Running track of Chinese stock market during 2005-2010.*

annual total desterilization amounts to 80 billion shares, 40% more than the number in 2007. There are also about 12 billion shares of the IPO[6] in 2008. The market capacity is actually increased greatly in 2008, indicating a sharp decline of the share prices. This can also be seen from FIG. 1.

The stock market in 2009 gradually goes up. There is a substantial increase in the size factor, and a moderate increase in the market performance factor. The data indicates that the increase in the performance factor mainly due to the improvement of the turnover rate, namely, the improvement of trading activity. The reason may be that investors are generally optimistic about the market in 2008, but they become much more optimistic in 2009 thanks to the government bailouts and other favorable factors.

The size factor in 2010 continued to climb, but the performance factor dropped a lot at the same time. The phenomenon can be explained by the running track diagram (see FIG. 6) and by the zoom-star plots in the later part of this section.

The diagram depicts the running track of three scale style stocks. It can be seen that the overall trend of each style is consistent. However, in the same year, the performance of small-cap stocks is much better than the others, the large-cap stocks performs the worst though they get the highest size scores. This reflects that the negative correlation between the size factor and the performance factor, and that the investors prefer trading in smaller stocks, leading to a cold transaction of large-cap stocks who have better financial position (empirical research by Long et al. (2009)). It is worth mentioning that the large-cap stocks in 2010 showed a significant increase in the size factor and a decline in performance factor. While the mid-cap and small-cap stocks had a slight decrease in size factor, and their decline in performance factor is more moderate than that in large-cap.

6. IPO (Initial Public Offering) is a type of public offering where shares of stock in a company are sold to the general public, on a securities exchange, for the first time.

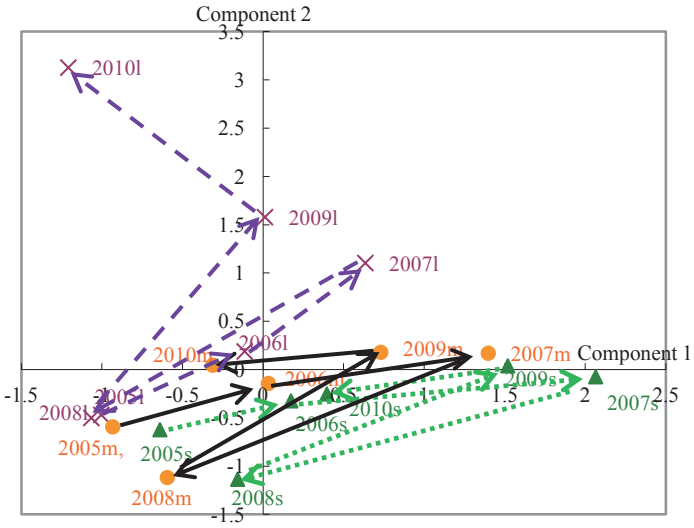

FIG. 6 – *Running track of the large, middle, and small cap stock market.*

5.3 Sample distribution of the style stocks

Take the maximum and minimum of factor scores for each style stocks in each year, we get the interval data sets of the factor scores. The distribution of the six style stocks can be drawn from the interval data sets in a globe principle plane constructed by Component 1 and 2.

From the sample distribution of style stocks in 2007 (a typical bull market) and in 2008 (a typical bear market), the typical characteristics of the stock market and its trading behavior can be seen, offering a better understanding of the dynamic changes in the operation of the market.

As shown in FIG. 7, size is an important determinant of the market performance. The small-cap stocks always perform the best, followed by the mid-cap stocks, depicting a typical negative correlation between the scale and the market performance. Small-cap stocks are actively traded and easy to resell, it also led to small cap stocks vulnerable to manipulation by minority investors. This resulted in the common pursuit of the market bid-ask spread, and constantly pushed the high small-cap market price. It also brought excessive speculation, increasing the risk to the stock market. At the same time, growth stocks in the market always performed slightly better than value stocks, reflecting the rational characteristics of the Chinese stock market trading. Therefore, the overall market is still limited rational and of excessive speculation.

As a typical bull market and bear market, the sample distribution in 2007 and 2008 are very different. The difference of the style stocks in 2007 is mainly reflected in the size aspect. The average size and market performance of value stocks and growth stocks are basically the same, but the dispersion degree of the market performance of growth stocks is much higher than that of value stocks. Investors have a good expectation of the whole market, and there is a good performance in both value stocks and growth stocks. In 2008 the overall market performance of

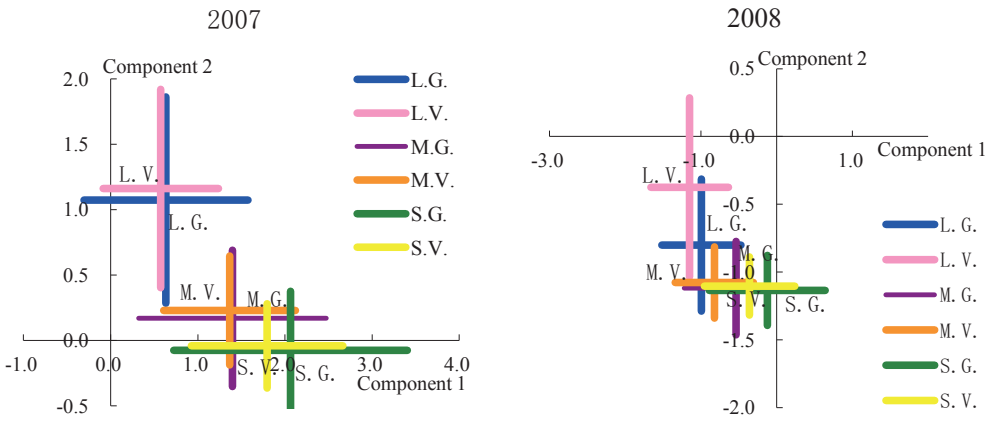

FIG. 7 – *Distribution of the six stock style portfolios in 2007 and in 2008.*

value stocks is significantly worse than growth stocks, while the size factor difference between the style stocks is reduced.

5.4 Zoom-star plots

Zoom-star plot is a descriptive statistics of symbolic data, it can visually compare the characteristics between the different indicators or different sample points, and the dynamic changes of these characteristics in the time series. Take the zoom-star plots of the return rate and the turnover rate as examples, we can find the differences of the two variables between the six style stocks.

5.4.1 Dynamic changes of the return rate

The Zoom-star diagram of return depicts the average and dispersion of the return rate of the six style stocks.

As shown in FIG. 8, the return rates in 2005 were negative, and they turned positive in 2006, and continued to grow in 2007. The maximum level in 2007 was the rate of mid-cap growth, which was nearly 13.76%. The minimum is 5.93% of large-cap growth. Then the rate plunged in 2008, when all style stocks are at loss, between which the best condition is -5.12% of small-cap value. After the rebound of the stock market in 2009, the return rate of large-cap value dropped dramatically in 2010.

Generally speaking, the return rate of value stocks is lower than that of growth stocks, and the dispersion of the value stocks is also smaller than that of the latter ones.

5.4.2 Dynamic changes of the turnover rate

The turnover rate reflect the activity level of transactions, the market expectations, and the preference of investors toward different stocks.

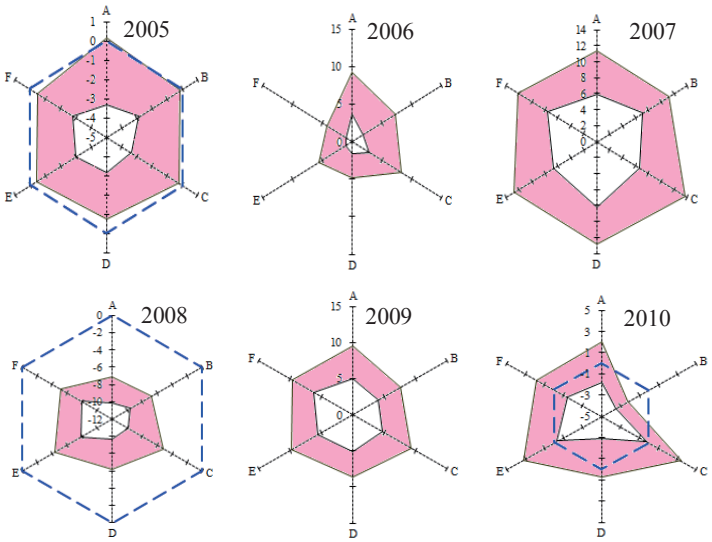

FIG. 8 – *Zoom-star plots of return (2005-2010). Note: The radial axes of each Zoom-plot diagram are, in a clockwise direction from 12 o'clock , (A) large-cap growth,(B) large-cap value, (C) mid-cap growth, (D) mid-cap value, (E) small-cap growth, and (F) small-cap value stocks. The interval area is filled to indicate the presence of interquartile values.*

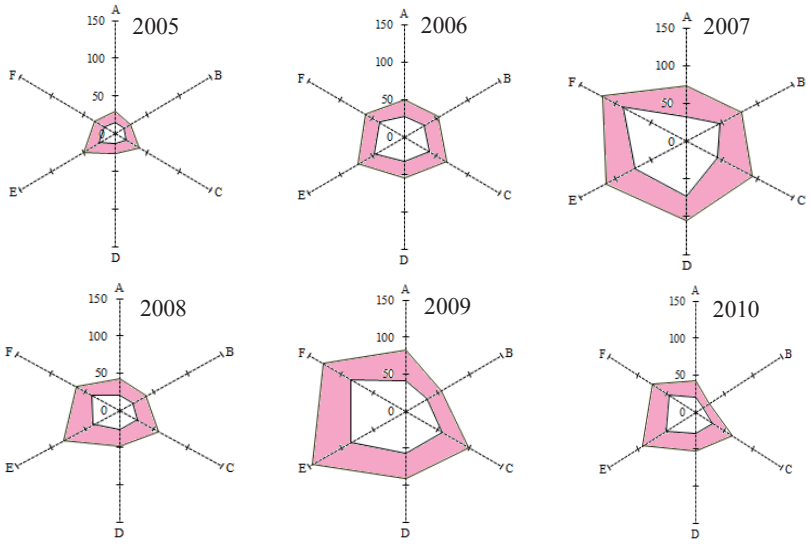

FIG. 9 – *Zoom-star plots of turnover (2005-2010). Note: The meaning of the radial axes are the same as that in Fig 8*

It can be seen from FIG. 9 that there are very active transactions in 2007 and 2009, especially the turnover rate of small-cap growth in 2009 when the very high level is 145.07%. In contrast, the turnover in 2010 dropped significantly.

The turnover of growth stocks is higher than that of value stocks in 2005 and 2006, indicating that the investors place great importance on the growth of the stocks. Since 2007, the turnover difference is reflected in the scale rather than the value or growth factors of the stocks. The small-cap stocks, regardless of growth stocks and value stocks, had a higher turnover rate than the large-cap and mid-cap stocks. The investors prefer the small-cap stock market to the growth of the shares.

From the above analysis, it is known that there was a very low return and turnover rate of large-cap value stocks in 2010. At the same time, the NMC of large-cap value stocks had a 104% year-on-year growth (see FIG. 10). All these lead to a substantial increase of the large-cap portfolios and value stocks in the market performance factor, and a substantial decrease of them in the size factor.

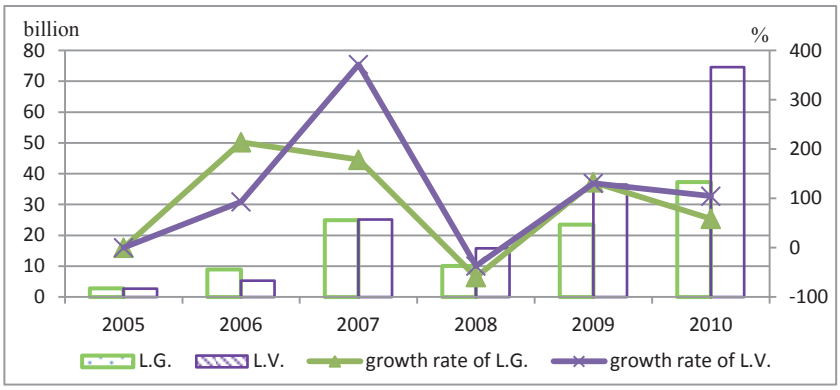

FIG. 10 – *The NMC of large-cap portfolios and its year-on-year growth rate (2005-2010).*

6 Conclusion

We use the symbolic principal component analysis on the empirical data of the CITIC style indices in six years (2005-2010), and try to find the characteristics of Chinese stock market from multiple perspectives. Two components are extracted from five variables–P/E ratio, NMC, turnover rate, return rate, and volatility. They are defined, according to the relationship between them and the variables, as the market performance factor and the size factor. We draw the run track and distribution maps of the six stock style portfolios by the two components, combine with the Zoom-star plots of symbolic data, and find that the Chinese stock market is of excessive speculation and of bounded rationality. That is, though the risk-benefit asymmetry has reversed after 2007, the investment behavior of the market still exhibits paying little attention to the stock's intrinsic value but investing more in the small-cap stock market speculation.

SPCA has been proved as an effective tool of analyzing the multi-dimensional data with huge number and complex structure (stock market as a typical example). Affected by many factors, such as the data lag, we cannot analyze the stock market in 2011 and 2012 when there is a spread of sovereign debt crises. Using the same method, we will trace China's stock market after 2010 to explore the new characteristic of the market.

Acknowledgment

This research was supported in part by National Natural Science Foundation of China (No.71101146, 71241014, 70921061) and the President Fund of GUCAS.

References

Billard, L. and E. Diday (2006). *Symbolic data analysis: conceptual statistics and data mining*, Volume 636. Wiley.

Cazes, P., A. Chouakria, E. Diday, and Y. Schektman (1997). Extension de l'analyse en composantes principales a des domnees de type intervalle. *Revue de Statistique appliquée 45*(3), 5–24.

Cheng, S. (2003). *Diagnosis and treatment: to Reveal the Chinese Stock Market*. Beijing: Economic Science Press.

Chouakria, A. (1998). *Extension des méthodes d'analyse factorielle à des données de type intervalle*. Ph. D. thesis, Universit ąäe Paris-Dauphine.

Chouakria, A., P. Cazes, and E. Diday (2000). *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data.* Berlin: Springer-Verlag.

Diday, E. (1988). The symbolic approach in clustering and related methods of data analysis: the basic choices. *In: H.-H. Bock (Ed.), Classification and Related Methods of Data Analysis, Proc. of IFCS 87*, 673–684.

Douzal-Chouakria, A., L. Billard, and E. Diday (2011). Principal component analysis for interval-valued observations. *Statistical Analysis and Data Mining 4*(2), 229–246.

Giordani, P. and H. Kiers (2004). Principal component analysis of symmetric fuzzy data. *Computational statistics & data analysis 45*(3), 519–548.

Huang, Z. and S. Sun (2008). Literature review and prospect of research on stock market bubbles. *Finance & Economics 9*, 50–57.

Lauro, C. and F. Palumbo (2000). Principal component analysis of interval data: a symbolic data analysis approach. *Computational statistics 15*(1), 73–87.

Le-Rademacher, J. and L. Billard (2012). Symbolic covariance principal component analysis and visualization for interval-valued data. *Journal of Computational and Graphical Statistics 21*(2), 413–432.

Long, W., H. Mok, Y. Hu, and H. Wang (2009). The style and innate structure of the stock markets in china. *Pacific-Basin Finance Journal 17*(2), 224–242.

Shi, Q. and S. Xu (2003). Techno-economic analysis on stock market behavior. *Journal of Beijing University of Technology 3*, 54–57.

Wang, L., Y. Hu, and H. Wang (2005). The canonical correlation analysis of interval data and its application in the stock market. *System Engineering Theory and Practice 1*, 128–133.