Research on the Investment Value of Stocks in SSE Star Market Based on Factor Analysis and Cluster Analysis

Sini Yang(✉)

Electrical and Computer Engineering, Shanghai Jiao Tong University, Shanghai, China
yangsini020612@sjtu.edu.cn

Abstract. This paper mainly focuses on the investment value of stocks in the SSE Star Market using factor analysis and cluster analysis with Python language. The SSE Star Market is an important platform in Shanghai and even China since it gathers resources and capitals of innovation and technology to help and serve numerous startups to grow up. In this research, data of each stock in SSE Star Market such as daily opening price were collected and then categorized and processed to 12 indexes. The factor analysis is used to extract common factors by combining indexes on which one factor has similar loading. The research extracts 3 common factors: the price factor, the flow factor, and the profit factor. Cluster analysis is used in the research to categorize all stocks in SSE Star Market into 3 categories: high price stocks, high security stocks, and high potential stocks. Sample stocks in each category are picked up to analyze deeper on their investment value. Therefore, this paper explores the investment value of all stocks in SSE Star Market by doing the factor analysis and the cluster analysis and provides sample stocks for investors to refer to.

Keywords: SSE Star Market · factor analysis · cluster analysis · investment value · Python

1 Introduction

The SSE Star Market, standing for the Science and Technology Innovation Board of the Shanghai Stock Exchange, is a new science and technology innovation board to carry out the nation's strategy of innovation-driven development to promote the nation's development with high quality. This new science and technology innovation board is set up to provide a platform for the latest innovation company in the field of science and technology. Multiple high-technology industries are included in the SSE Star Market: information technology, high-end equipment, new materials, new energy, energy conservation, environmental protection and biomedicine and other high and new technology industry and strategic emerging industries, promoting Internet, big data and cloud computing, artificial intelligence, and manufacturing depth fusion industries [1, 2].

Factor analysis is a common way to descend the dimension of indexes. This method looks for the deep relationships among multiple indexes. It can be viewed as a progressing

J. Yen et al. (Eds.): ICBIS 2023, AHCS 14, pp. 238–248, 2023.
https://doi.org/10.2991/978-94-6463-198-2_27

method of Principal Component Analysis (PCA). Its principle is to categorize similar indexes with similar or even the same value while investing together in order to make the number of indexes fewer. This method can not only ensure the integrity of the information hidden in the whole data, but also include the whole complex relationships among multiple factors with only few common factors [3].

Cluster analysis is a common method to dynamically categorize objects. It is a multivariate statistical analysis method based on the idea of "birds of a feather flock together". It's come up with by James MacQueen, an IMS researcher in America, in 1967. This method is aimed at putting samples with high similarities together, as a cluster [4, 5].

Scree test is one statistical method used in this paper. It is raised by Cattell. This method uses scree plot to determine the number of factors. In the principal axis factor method, the variance contribution of a common factor is equal to the value of the characteristic root corresponding to the factor. The cumulative variance contribution rate can be replaced by the calculation of the percentage of the cumulative characteristic root, or even the number of factors can be determined by directly observing the change of the characteristic root. When the value of a feature root decreases significantly from that of the previous feature root and the feature root is small, at the same time, the subsequent feature root does not change much, adding factors corresponding to this feature root will only add little information. In other words, the first few feature roots are the number of common factors to be extracted [6].

KMO test is another statistical method used in this paper. This test tests whether the data set is suitable for conducting factor analysis. Correlation analysis is conducted to select suitable samples to do the principal component analysis so as not only to avoid the loss of the information hidden in the data, but also to help avoid the dimension reduction [7].

2 Methodology

2.1 Factor Analysis

Factor analysis is a kind of statistical analysis method of multidimensional variables which reduces multidimensional variables to a few common factors according to the correlation between variables, and then analyzes and deals with them. The basic idea of this method is to decompose the original variables into two parts. One part is a linear combination of common factors, which condenses and represents most of the information in the original variables. The other part is a special factor unrelated to the common factor, reflecting the gap between the linear combination of the common factor and the original variable [9]. The basic model of this method is described below:

$$x = Af + \varepsilon \tag{1}$$

In this way, the model can be rewritten as:

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \dots & a_{pm} \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix} \tag{2}$$

Here, f is the vector for common factors. It stands for non-directly observable objective common influencing factors in the original variables. A = (a_{ik}) is the factor loading matrix. Every element a_{ik} is the correlation coefficient of the variable xi and the common factor f_k. It reflects the loading x_i has on f_k. While the abstract value of a_{ik} is larger, the correlation between the variable xi and the common factor f_k would be greater.

The most important part of the factor analysis is to solve the equation to find the factor loading matrix A and the vector for common factors f. Firstly, standardization will be done to avoid the influence of the various dimensions. The formula is listed as below:

$$x_{ij} = \frac{x_{ij} - \frac{1}{n}\sum_{j=1}^{n} x_{ij}}{\sqrt{\frac{1}{n-1}\sum_{j=1}^{n}\left(x_{ij} - \frac{1}{n}\sum_{j=1}^{n} x_{ij}\right)^2}} \tag{3}$$

Secondly, the covariance matrix of the data will be calculated. Every element of the covariance matrix is derived as below:

$$s_{ij} = \frac{1}{n-1}\sum_{k=1}^{n} x_{ik}x_{jk} \tag{4}$$

Thirdly, do the eigenvalue decomposition to the covariance matrix of the data. P eigenvalues λ would be derived and p corresponding eigenvector γ would be derived. Factor loading matrix is made up by eigenvectors with the greatest m eigenvalues. Every eigenvector should be divided by the corresponding standard deviation to ensure the variance of every component of the common factor vector is 1. The corresponding eigenvector γ_j of the factor loading matrix should be multiplied by γ_j. Finally, the factor loading matrix can be derived as below:

$$\hat{A} = \left[\sqrt{\lambda_1}\gamma_1, \sqrt{\lambda_2}\gamma_2, \cdots, \sqrt{\lambda_m}\gamma_m\right] \tag{5}$$

The parameter m is derived by the cumulative variance contribution rate of the common factor, which can be calculated as below: [8]

$$m = arg\min_{m}\left(\frac{\sum_{i=1}^{m}\lambda_i}{\sum_{i=1}^{p}\lambda_i} \geq r\right) \tag{6}$$

Generally, the linear combination is considered as being able to still contain the information in the original variables while the cumulative variance contribution rate of m common factors is more than 90%. With the method of regression, the scores of the original variables on the common factors, the common factor vector f, can be calculated, which is shown as below:

$$\hat{f}_j = \hat{A}^T S^{-1} x_j \tag{7}$$

In this way, the special factor vectors of the original variables can be derived as below:

$$\hat{\varepsilon}_j = x_j - \hat{A}\hat{f}_j \tag{8}$$

2.2 Cluster Analysis

The cluster analysis is a methodology to analyze big data by grouping. This paper chooses one of methods of conducting the cluster analysis: the partitional clustering. This method helps determine how many groups the whole data set should be divided into by calculating out the value of k. SSE (sums of the squared errors) is calculated to draw the elbow figure, which is shown as below:

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \tag{9}$$

The value of k is determined by the place where the abstract value of the slope changes the most noticeably [10, 11].

3 Results and Discussion

3.1 Results

The following steps in this section are all processed with the help of Python language. As is mentioned in Sect. 2, the raw data set is downloaded from the CSMAR database, which has the time duration from 2020.8.13 to 2022.8.12. The raw data set is processed into the following 12 indexes to do the further analysis, which is demonstrated in Table 1.

Since the dimensions of the index are not the same, data standardization is conducted. Correlation coefficient between each index is shown in the Fig. 1.

Table 1. Economic indexes used in the following research [12]

Index Name	Explanation	Range
PE	Price Earning Ratio	[10.116, 1947.832]
PB	Price/Book Value Ratio	[1.275, 2334.046]
PS	Price/Sales Ratio	[0.630, 166914]
Turnover	Volume/Total number of shares issued x 100%	[0.009, 0.597]
Liquidity	Index of Liquidility	[0.001%, 0.021%]
Opnprc	Daily Opening Price	[3.120, 842.658]
Hiprc	Daily Highest Price	[3.151, 867.434]
Loprc	Daily Lowest Price	[3.089, 817.280]
Clsprc	Daily Closing Price	[3.114, 843.995]
ChangeRatio	How one stock changes in one day.	[3.089, 817.280]
ap	Amplitude	[1.513, 38.056]
av	Average Price	[3.118, 842.645]

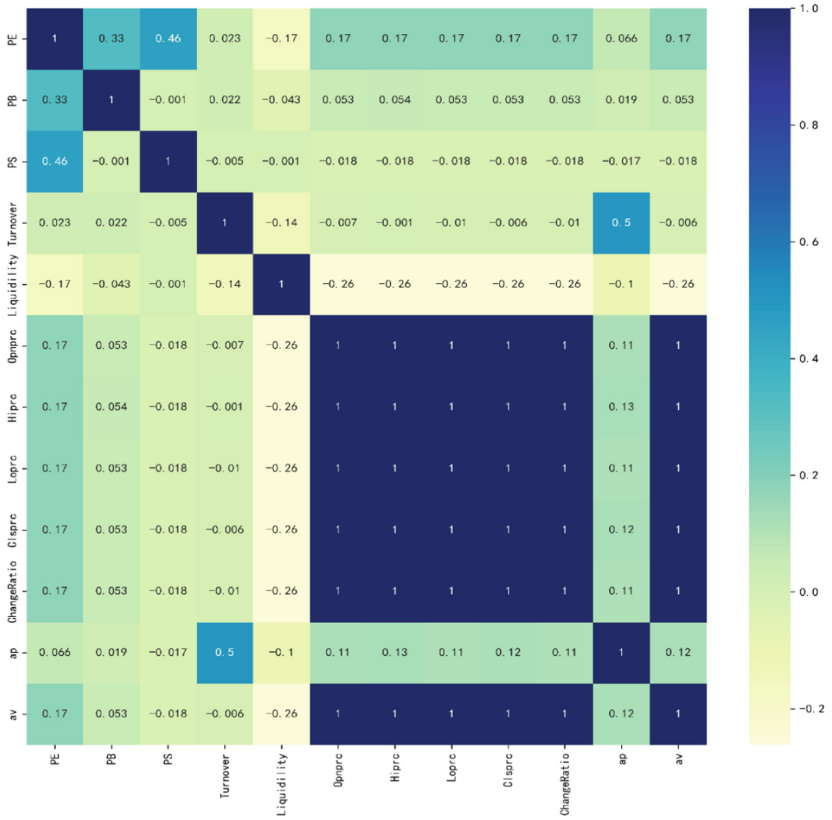

Fig. 1. Correlation efficient of each index

Since there's close correlation between some of the indexes, the Bartlett test is conducted to test whether the data set is available to do the factor analysis. The data set used in this paper has a p value of 0.001, which implies that the factor analysis can be conducted on it. Besides, the KMO test is conducted and the KMO value of this data set is 0.7405, leading to the belief that there does exist some correlation among indexes. Consequently, the common factors of this data set can be figured out.

By using the screen test mentioned in Sect. 2, according to the eigenvalues, the number of the common factors is 3. The screen figure is shown as below. For these three factors, the accumulative eigenvalue proportion is 84.89%. This implies that the selection of the 3 factors is reasonable. The result is demonstrated in Table 2.

Based on the three common factors, a heatmap of the factor loading matrix can be concluded, which is shown in the Fig. 2.

By analyzing the correlation, the 3 common factors are decided to be named as the price factor, the flow factor, and the profit factor. The price factor can reflect a stock's performance on price. It is picked out because it has a high load on indexes about price. The flow factor can reflect a company's ability to pay a debt in a short time. It is picked out because it has a high load on these two indexes: turnover and ap. The profit factor

Table 2. Eigenvalues and the accumulative eigenvalue proportion

	Factor 1	Factor 2	Factor 3
Eigenvalue	6.65	1.63	1.06
Proportion	0.60	0.15	0.10
Acc. Proportion	0.60	0.75	0.85

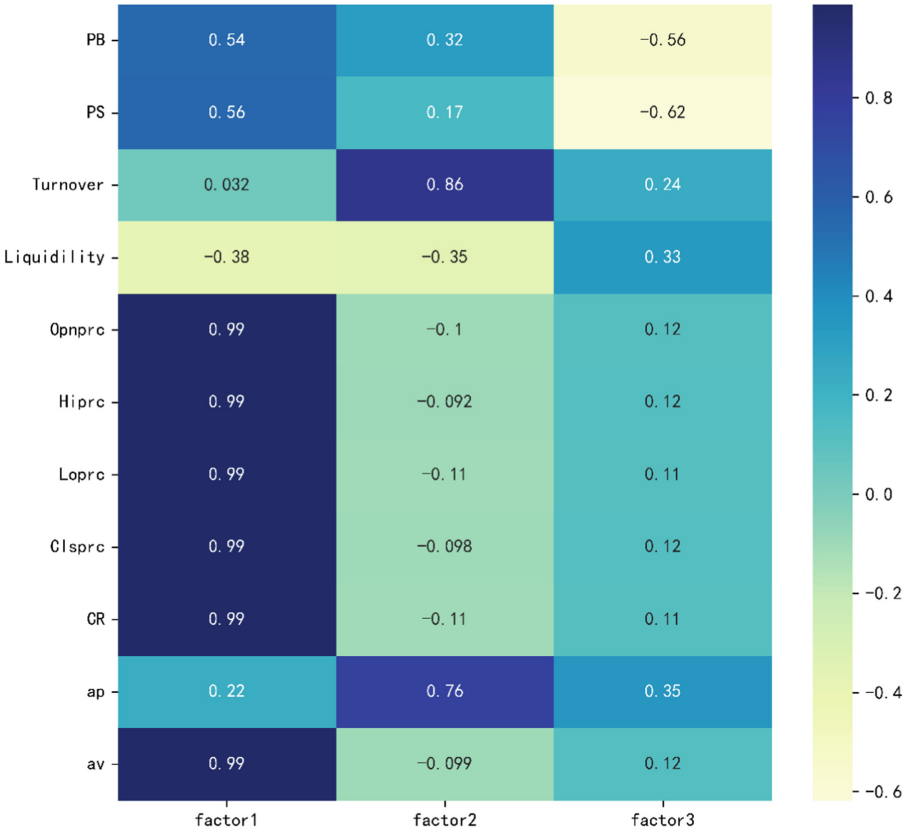

Fig. 2. Heatmap of the factor loading matrix

can reflect a company's ability to profit. It is picked out because it has a high load on these two indexes: PB and PS.

Cluster analysis is utilized after figuring out the common factors. In this paper, partitioning based clustering method with k-means algorithm is used to find out the best number of clusters.

As it is mentioned in Sect. 2, the inflection point on the figure shows the best k to choose. In this paper, k is figured out as 3. Consequently, stocks can be divided into 3 clusters. The k-means elbow figure is demonstrated here as Fig. 3.

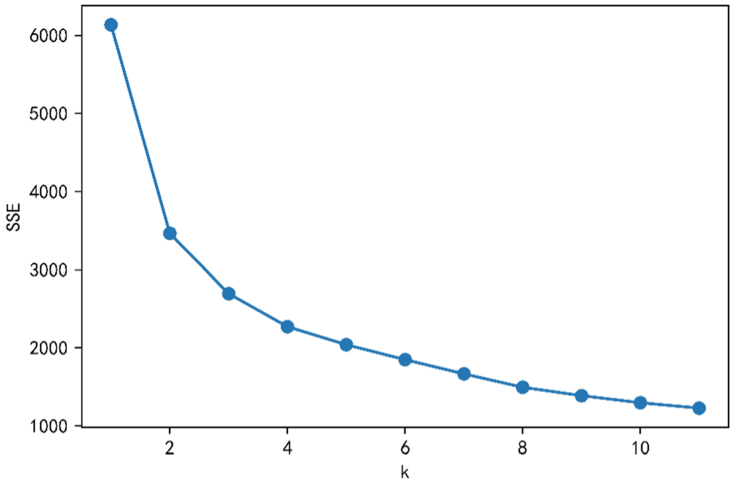

Fig. 3. The k-means elbow figure

3.2 Discussion

The 3 clusters are named after their characteristics. The cluster with stocks ranking high on the profit factor is named as the high price stock. The cluster with stocks ranking high on the flow factor high is named as the high security stock because a company would have great ability to pay debts while it has a high liquidity. The cluster with stocks ranking high on the profit factor is named as the high potential stock because the great ability to profit implies a bright future of the development of the company.

Typical stocks from the 3 clusters are listed in Table 3.

From the result, it is fair to conclude that industries of stocks in SSE Star Market are concentrate. From Fig. 4, it can be easily observed that top industries of the SSE Star Market are the next-generation IT companies, biology, high-end equipment manufacturing and new materials. The Fig. 5 demonstrates that the main part of the companies in SSE Star Market are located at the Eastern Coastal Developed Regions of China.

Table 3. Typical companies of 3 clusters

Cluster Name	Typical Company
High Price Stock	Roborock, Hoymiles, 3PEAK INCOPORATED, GIMI Technology, Beijing Huafeng Test & Control Technology Co., Ltd., Medicilon
High Security Stock	Medlander Medical Technology Inc., Wuxi Taclink Optoelectronic Technology Co., Ltd., CHINA MICRO SEMICON CO., LIMITED, Hangzhou SDIC Microelectronics Inc., Segway-Ninebot, Favored Tech
High Potential Stock	Suzhou Nanomicro Technology Co., Ltd., KINGSOFT OFFICE, QuantumCTek Co., Ltd., APsystems, HUAQIN Technology, Shanghai Awinic Technology Co., Ltd.

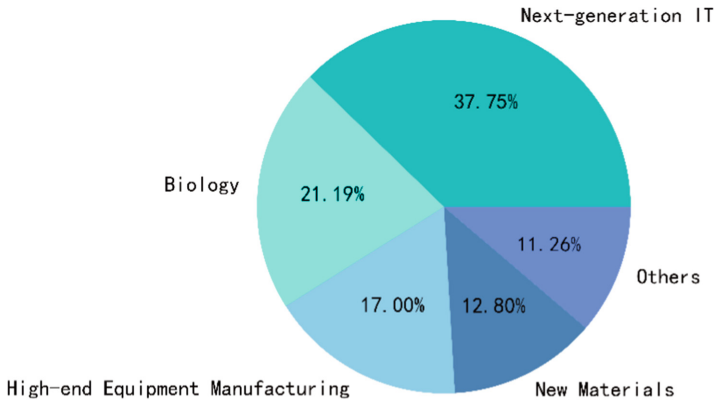

Fig. 4. Industry distribution of stocks in SSE Star Market

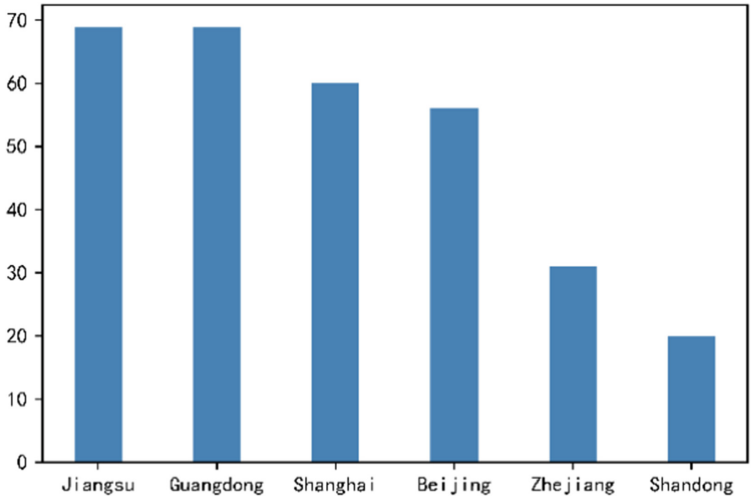

Fig. 5. Location distribution of companies in SSE Star Market

Roborock is a typical company in the cluster of high price stock. It is an intelligent hardware manufacturer focusing on technological innovation. Roborock has launched its own brand products "Stone intelligent sweeping Robot", "Stone self-cleaning sweeping and dragging robot", "Small tile intelligent sweeping robot", "Stone wireless handheld vacuum cleaner" and "Stone intelligent double scrub machine". Additionally, it has developed and produced Mijia sweeping robot, Mijia Sweeping Robot 1S and Mijia handheld wireless vacuum cleaner products for Xiaomi. Its core technical personnel are technical experts from Microsoft, Huawei, and Intel. These characteristics make Roborock a company popular with a high price in the stock market.

Wuxi Taclink Optoelectronic Technology Co., Ltd. is a typical company in the cluster of high security stock. It is a great designer and manufacturer of optoelectronic products

such as the optical amplifier products, the transceiver, and the optical components. This company features its quality control policy and environmental policy. Additionally, this is a consumer-focus company with high reputation among customers. Since their products are welcome and reliable, the sales of the company are good, implying that the flow of asset of this company is steady too. Consequently, Wuxi Taclink Optoelectronic Technology Co., Ltd. is a safe choice to choose while investing.

APsystems is a typical company in the cluster of high potential stock. It is a company focuses on the R&D and industrialization of MLPE component level power electronics. Products of this company includes the micro inverters, the intelligent shut-off, the power optimizer, the EMA intelligent monitoring and operation platform and the photovoltaic system solutions. High developing potential in the future can be observed in the company since this company had obtained 109 authorized patents (including 59 invention patents) until December 31, 2020. Additionally, this company emphasizes on the environmental protection, which just fits the heated trend in the future. As a result, APsystems is a high-tech company with focuses on sustainable development, which makes it a company to be considered with great potential to choose while investing.

4 Conclusion

To conclude, this paper mainly focuses on doing research on the investment value of stocks in SSE Star Market based on the two following statistical methods: the factor analysis and the cluster analysis. After analyzing the data set downloaded from CSMAR database with Python language, it is easy to observe that industries of stocks in SSE Star Market are limited and concentrate. More than one-third are the next-generation IT companies. Biology, high-end equipment manufacturing and new materials are also the field where companies in SSE Star Market are in. These industries are all benefited from the cutting-edge technology. Besides, most of companies in SSE Star Market are located at the Eastern Coastal Developed Regions of China. Additionally, fields of stocks in SSE Star Market are greatly applicable, with highly practical values. From the analysis, a blueprint of the SSE Star Market is formed, which can be concluded as the following three characteristics: high technology involved, geographically concentrated, and highly applicable. Therefore, the investment value of stocks in SSE Star Market is high and beneficial to the whole society.

The 12 indexes of the raw data are calculated and then standardized because of the different dimensions. From the Bartlett test, since the p value is 0.001, the factor analysis is available on this data set. Besides, the KMO value of this data set is 0.7405, implying that correlation does exist among indexes. The number of common factors is figured out to be 3 while the accumulative eigenvalue proportion is 84.89%. Consequently, the selection of the 3 factors is reasonable. According to their different load on various indexes, the 3 common factors are each named as the price factor, the flow factor, and the profit factor. The price factor reflects a stock's performance on price in the market. The flow factor reflects a company's security while investing. The profit factor reflects a company's potential to develop in the long run. From the partitioning-based clustering method with k-means algorithm, the best number of clusters is 3. As a result, all stocks are divided to 3 clusters: the high price stock, the high security stock, and the high potential

stock. Typical stocks of each cluster are listed in Sect. 3 for investors with different investing attitude to refer to. However, because of the limitation of the data tested in this paper on time, investors should pay attention to the timing difference hidden and only take the result of this paper as a reference. Stock investments are subject to market risks. Consequently, in the future research, analysis between data sets with different dates should be conducted to reach a more general conclusion.

References

1. Guo, C. and Zhang, C., "Discussion on the system innovation of the SSE Star Market," Theoretical Exploration, 6: 92-98 (2021).
2. Huang, D. and Wang, H., "SSE Star Market: a new kind of system supply," Theoretical Exploration, 5: 117-122 (2019).
3. Ye, C., Guan, X., Yang, L. and Chen, Y., "Research on Comprehensive performance Evaluation of listed pharmaceutical Manufacturing companies—based on the factor analysis and the cluster analysis," Communication of Finance and Accounting, 127–130 (2021).
4. Gu, B., Zhu, J. and Huang, T., "Research on the Investment Value of Science and Innovation Board Stock Based on Factor Analysis and Cluster Analysis," Journal of Jilin Agricultural Science and Technology University, 29(3), (2020).
5. Wu, J., "Analysis of stock investment value of listed banks based on factor analysis and cluster analysis," General Investment Guide, 19: 1-2 (2019).
6. Dai, Z., "The stock yield research based on the functional data and cluster analysis," Shanghai University of Finance and Economics Master Dissertation, (2020).
7. Yang, L., Wang, T. and Zhao, G., "The value analysis of the financial industry stocks," China Market, 10, (2014).
8. Lv, Y., Zhu, Y., Geng, Q., An, Y. and Chen, Y., "The influential factors and developing trend of circulation industry under the background of consumption driving," Journal of Commercial Economics, 6: 17-20 (2020).
9. Zhong, W., Jiao, Z. and Cai, L., "Student achievement clustering analysis based on K-means algorithm," Educational Information Technology, 5: 56-58 (2021).
10. Xie, G. and Xu, J., "Application of factor analysis and cluster analysis in financial stock investment," Coastal Enterprises and Science and Technology, 4: 11-16 (2016).
11. Guo, H. and Tang, L., "Application of clustering, Bayesian Discriminant, and factor Analysis in quantitative investment," Finance Theory and Teaching, 5: 6-12 (2017).
12. "SSE Star Market Daily Data," China Stock Market & Accounting Research Database (2022).