Investment Value Evaluation of Listed Companies: Based on Principal Component Analysis, Factor Analysis and Cluster Analysis

Yang Zhoufan

Shandong University, Weihai, Shandong 264209

Keywords: Principal Component Analysis, Factor Analysis, Cluster Analysis, Investment Value.

Abstract: Since stock price reflects the operating conditions of listed companies, the epidemic in early 2020 made the fluctuation of stock prices more uncertain. In this paper, we take the data of listed companies on the Shenzhen Main Board in the first quarter of 2020. For the seven financial indicators, we adopt principal component analysis and factor analysis to extract components to calculate the score. Comparing these two methods, we find that factor analysis performances better in this case. On this basis, companies are classified by their performance through cluster analysis. It is found that in the first quarter of 2020, despite the influence of the epidemic, large-scale enterprises still behave well, especially those in the pharmaceutical and consumer goods industries. However, small enterprises, especially in the tourism industry and catering industry, are severely affected by the epidemic. Based on the findings, we give suggestions to investors.

1. Introduction

Stock investment has high returns and high risks. Recently, affected by the epidemic, there are more uncertainties in the fluctuation of stock prices. Investors should be more cautious when choosing stocks for investment and should make choices after scientifically and reasonably measuring the investment value of listed companies. The development and investment value of listed companies are reflected in their financial data, but companies' financial data are numerous and relevant, which should not be directly used for evaluation.

Principal component analysis, factor analysis, and cluster analysis are widely used in evaluating the investment value of listed companies. In this paper, principal component analysis and factor analysis are used to reduce the dimension of financial data of listed companies. Then cluster analysis is carried out on listed companies according to their financial situation to analyze the financial condition of each type of listed company. R language is used for data processing. Based on the results, this paper evaluates the investment value of listed companies and provides investors with stock selection suggestions.

2. Variable Selection and Data Preprocessing

The data comes from the financial data of 506 listed companies in the first quarter of 2020. To evaluate the investment value of listed companies in a comprehensive way, we select seven variables from four sections: capital stock expansion ability, profitability, operation ability, and solvency, as shown in Table 1.

Firstly，we remove the line with "ST" before the company name, and ST means "special treatment," which means that the listed company has financial status or other abnormal conditions. There may be significant accounting errors or false records in the business and accounting reports of such companies. This paper does not analyze the data of such listed companies.

A-shares are RMB paid shares issued by listed companies, while B-shares are foreign currency paid shares issued by listed companies. A-shares and B-shares belong to the same company and use the same financial data, so one of them can be selected when analyzing the financial data of listed companies. Delete the line with "b" after the company name.

Some listed companies miss the values of some variables, so delete the missing lines of these data to ensure that each data has the same number of variables.

After data preprocessing, there are 402 pieces of data and seven variables.

Table 1 Classification of investment value evaluation indicators

Classification	Variable
Capital stock expansion ability	Earnings per share (yuan) X1
	Net assets per share (yuan) X2
Profitability	Return on net assets (%) X3
	Gross profit margin (%) X6
	Operating profit rate (%) X5
Operational capability	Total assets turnover rate (%)X4
Debt paying ability	Current ratio (%) X7

3. Data Analysis

3.1 Calculating Correlation Matrix

Calculating the correlation matrix and expressing it as Table 2. The color from light to deep indicates the correlation coefficient from small to large, and the correlation from weak to strong. It can be seen that there is a strong correlation among some variables, and the information overlaps. For example, the correlation coefficient between earnings per share and net assets per share is 0.655015, the correlation coefficient between return on net assets and earnings per share is 0.636657, and the correlation coefficient between return on net assets and gross profit margin is 0.378967. Therefore, the data is suitable for principal component analysis.

Table 2 Correlation coefficient matrix

	Earnings per share	Net assets per share	Return on net assets	Total asset turnover	Operating profit rate	Gross margin	Flow ratio
Earnings per share	1	0.652502	0.636258	0.113157	0.103418	0.339321	0.055578
Net assets per share	0.652502	1	0.313818	0.045093	0.098951	0.183614	0.015568
Return on net assets	0.636258	0.313818	1	0.155561	0.210004	0.37756	0.033179
Total asset turnover	0.113157	0.045093	0.155561	1	0.070407	-0.11956	-0.09276
Operating profit rate	0.103418	0.098951	0.210004	0.070407	1	0.077569	0.015314
Gross margin	0.339321	0.183614	0.37756	-0.11956	0.077569	1	0.197953
Current Ratio	0.055578	0.015568	0.033179	-0.09276	0.015314	0.197953	1

3.2 KMO and Bartlett's Sphere Test

The value of KMO statistics is 0.6>0.5, and the p-value of Bartlett's sphere test is less than 0.05, so the data is suitable for factor analysis.

3.3. Descriptive Analysis

The calculated maximum value, minimum value, mean value, and variance of 7 variables are recorded in Table 3.

Among them, the variance of operating profit margin and gross profit margin is significant, while the variance of earnings per share and total asset turnover rate is small. Drawing the distribution histogram or scatter diagram of each variable for analysis.

From Figures 1 and 2, we can see that earnings per share are mainly distributed in the interval (-0.066, 0.154), and net assets per share are primarily distributed in the range (2.209, 6.209). The data's left deviation indicates that earnings per share and net assets per share of some listed companies are much larger than those of most listed companies.

Table 3 Characteristic number of each variable

Statistic	Earnings per share	Net assets per share	Return on net assets	Total asset turnover	Operating profit rate	Gross margin	Current Ratio
Maximum value	2.680748	42.631	9.396	1.347	1990.048	99.9	37.02
Minimum value	-0.61592	0.2087	-18.864	0	-4866.28	-219.85	0.205
Mean value	0.052329	5.077884	0.264047	0.118353	-18.4205	22.59398	1.987331
variance	0.056291	17.75945	8.312292	0.017717	75327.05	754.4888	6.436492

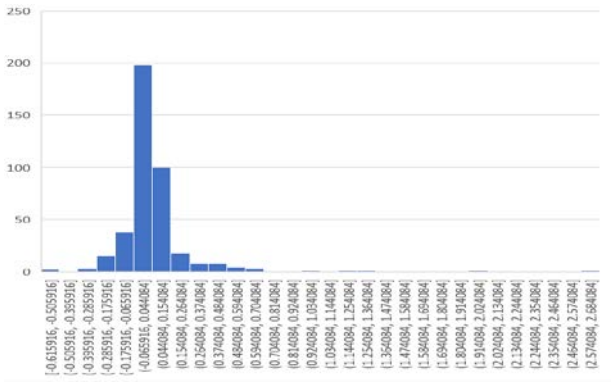

Figure 1 Distribution of earnings per share

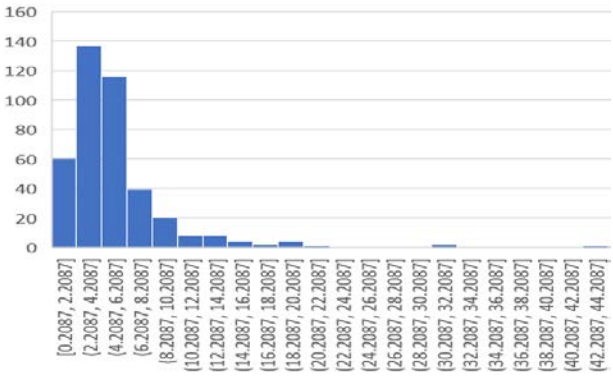

Figure 2 Distribution of net assets per share

Figures 3 and 4 illustrate that net asset turnover rate is mainly distributed in (0,0.126), and return on Net Assets is primarily distributed in (-0.664,2.136). The net assets per share of some listed companies are much larger than that of most listed companies.

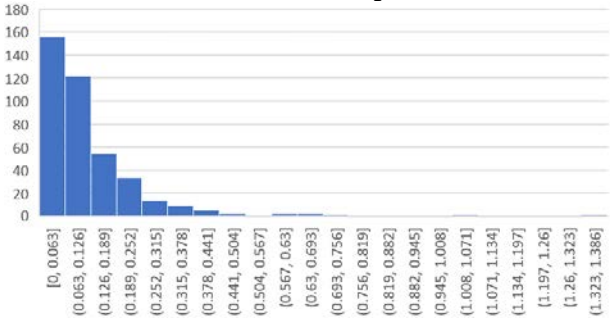

Figure 3 Distribution of net asset turnover rate

The variance of the operating profit rate is considerable. However, from the scatter diagram, it can be seen that the distribution of the operating profit rate of most listed companies is relatively concentrated. The difference between the operating profit rate of a few listed companies and other companies is vast, with the extreme value reaching 6856, resulting in significant variance. It can be seen from Figure 6 that the amount of current ratio is relatively concentrated, while Figure 7 shows that the distribution of gross profit margin is relatively scattered compared with the current rate.

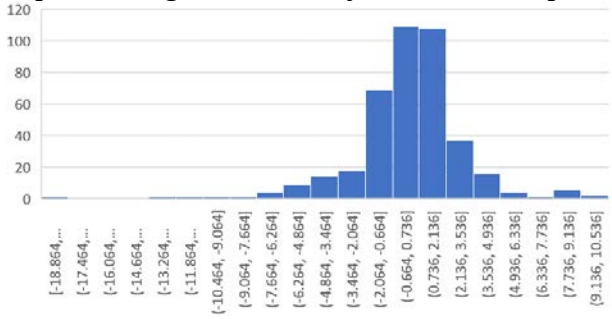

Figure 4 Distribution of return on Net Assets

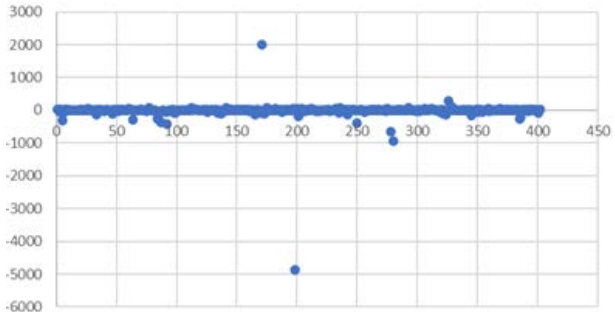

Figure 5 Scatter chart of operating profit margin

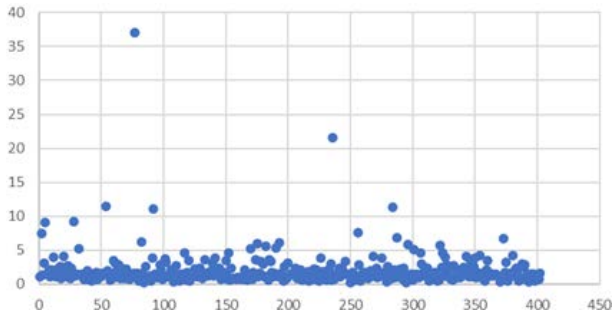

Figure 6 Scatter chart of flow ratio

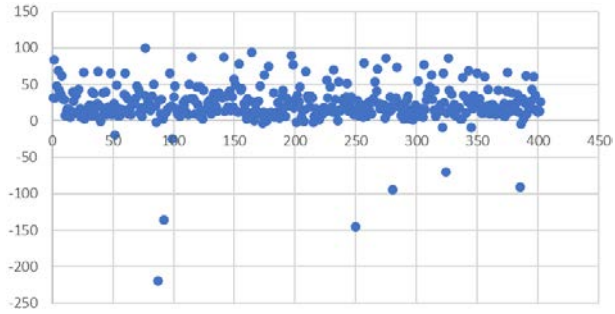

Figure 7 Scatter chart of gross profit margin

4. Principal Component Analysis

4.1 Eigenvalues and Eigenvectors of Correlation Matrix

Eigenvalues and eigenvectors are shown in Tables 4 and 5.

Table 4 Eigenvalues of Correlation Matrix

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7
2.3647100	1.2321974	0.9911745	0.8788349	0.7833237	0.5232539	0.2265056

Table 5 eigenvectors of correlation matrix

	L1	L2	L3	L4	L5	L6	L7
X_1^*	-0.57845	0.076027	0.222873	0.015576	0.106011	0.162332	0.756377
X_2^*	-0.46217	0.105308	0.347776	-0.12802	0.54349	-0.35773	-0.46327
X_3^*	-0.51795	0.079116	-0.11498	0.039028	-0.35413	0.616096	-0.45358
X_4^*	-0.09369	0.627791	-0.2075	0.665054	-0.12019	-0.31191	-0.00352
X_5^*	-0.18362	0.149043	-0.82772	-0.42361	0.252182	-0.09554	0.082373
X_6^*	-0.36459	-0.4564	-0.09127	-0.02346	-0.55534	-0.58437	-0.00232
X_7^*	-0.09	-0.59349	-0.28223	0.599629	0.42606	0.134899	-0.02703

4.2 The Number and Explanation of Principal Components

The principal components were selected according to the principle that the cumulative contribution rate reached more than 70%. The results are as follows:

Table 6 Variance contribution of principal components

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
Standard deviation	1.5377614	1.1100439	0.9955775	0.9374619	0.8850558	0.72336289	0.47592609
Proportion of Variance	0.3378157	0.1760282	0.1415964	0.1255478	0.1119034	0.07475055	0.03235795
Cumulative Proportion	0.3378157	0.5138439	0.6554403	0.7809881	0.8928915	0.96764205	1.00000000

The cumulative contribution rate of the fourth principal component reaches 0.7809881 >0.7, so according to the rule that the cumulative contribution rate reaches more than 70%, we choose the first four principal components. Get Z_1,Z_2,Z_3,Z_4:

$$Z_1 = \mathbf{0.578}X_1^* + \mathbf{0.462}X_2^* + \mathbf{0.518}X_3^* + 0.094X_4^* + 0.184X_5^* + 0.365X_6^* + 0.09X_7^*$$

$$Z_2 = 0.076X_1^* + 0.105X_2^* + 0.079X_3^* + \mathbf{0.628}X_4^* + 0.149X_5^* - \mathbf{0.456}X_6^* - \mathbf{0.593}X_7^*$$

$$Z_3 = 0.223X_1^* + 0.348X_2^* - 0.115X_3^* - 0.207X_4^* - \mathbf{0.828}X_5^* - 0.091X_6^* - 0.282X_7^*$$

$$Z_4 = 0.016X_1^* - 0.128X_2^* + 0.039X_3^* + \mathbf{0.665}X_4^* - 0.424X_5^* - 0.023X_6^* + \mathbf{0.6}X_7^*$$

Each component represents a different meaning. For example, Z_1 reflects the level of return given by the company to shareholders.

4.3 Principal Component Score

Substitute the standardized data into the above formula to get the scores of principal components of each company. Taking the variance contribution rate of each principal component as the weight, we construct the full evaluation function and then calculate each company's overall score.

$$Z = \sum_{i=1}^{4} \alpha_i Z_i / \sum_{i=1}^{4} \alpha_i$$

$$Z = (0.3378157Z_1 + 0.1760282Z_2 + 0.1415964Z_3 + 0.1255478Z_4)/0.7809881$$

The principal component scores of 402 companies are obtained. The top ten companies and the bottom ten companies are listed below.

The top ten companies with principal component scores are plentiful, and most of them belong to medicine and alcohol. Changchun Hi-Tech is located in the biopharmaceutical and health industries.

Jinsai Pharmaceutical and Baike Bio, which belongs to Changchun Hi-Tech, continues to contribute to its stable income, and in March 2020, the company obtained the drug registration approval for the nasal influenza vaccine. It can be seen that large companies have strong resilience in the face of the epidemic, and pharmaceutical listed companies have performed well during the pandemic.

After the principal component score, the top ten are mostly small companies, including "Xi 'an Diet" in the catering service category and "Shanxi Road and Bridge" in the transportation category. In the first quarter of 2020, the service industry and transportation industry were significantly affected by the epidemic situation.

Table 7 Top Ten Principal Component Scores

	Z1	Z2	Z3	Z4	Z	Z.rank
Changchun Hi-Tech	12.57324	0.755843	4.847384	-0.84952	6.351191	1
Wuliang liquid	9.028871	0.157267	1.984898	0.575546	4.393269	2
Gujinggong liquor	6.482523	0.442752	1.517021	0.238965	3.217259	3
Yunnan Baiyao	5.895294	0.508574	2.413436	0.04396	3.109265	4
National medicine yi zhi	4.397375	2.317392	2.181621	0.361067	2.877982	5
Luzhou laojiao	6.126096	-0.63472	0.829888	0.158937	2.682789	6
Huadong medicine	3.683105	1.52142	-0.08491	1.247683	2.121217	7
Shuanghui Development	2.66657	2.774456	-0.67823	2.3336	2.030933	8
Midea group	3.516332	0.999439	1.103125	0.009606	1.947797	9
Tianyin	0.555337	4.912708	-1.61472	4.445117	1.769313	10

Table 8 Top 10 Principal Component Scores

	Z1	Z2	Z3	Z4	Z	Z.rank
Huajin stock	-2.73817	0.578628	-0.03186	0.052625	-1.05129	393
Yihua healthy	-2.36485	-0.09653	0.109468	-0.56808	-1.11614	394
Zhonghe technology	-2.23856	-0.49012	0.246384	-0.62201	-1.13408	395
Xi 'an Diet	-2.69461	0.002656	0.352054	-0.52396	-1.18535	396
Guoxin healthy	-4.01452	-0.51935	0.686902	2.402941	-1.34271	397
Jiakai city	-3.31302	-0.28851	0.726989	-0.57327	-1.45842	398
Xinhualian	-3.06408	-0.33161	0.044892	-0.67406	-1.50033	399
Brand new	-2.44702	-3.57058	0.143877	1.573038	-1.58428	400
Shanxi Road and Bridge	-5.76472	1.979549	2.000247	-0.35667	-1.74203	401
Ziguang Xueda	-4.4233	-0.10693	0.253547	-0.29733	-1.93922	402

5. Factor Analysis

5.1 Determination of the Number of Factors

The number of factors determined by the gravel map is 2. If the eigenvalue of the correlation coefficient matrix is greater than one, the number of factors identified is two. If the cumulative variance contribution rate of four elements is more than 70% and the cumulative variance contribution rate of four factors is 0.78, four factors should be selected. At last, we choose four factors.

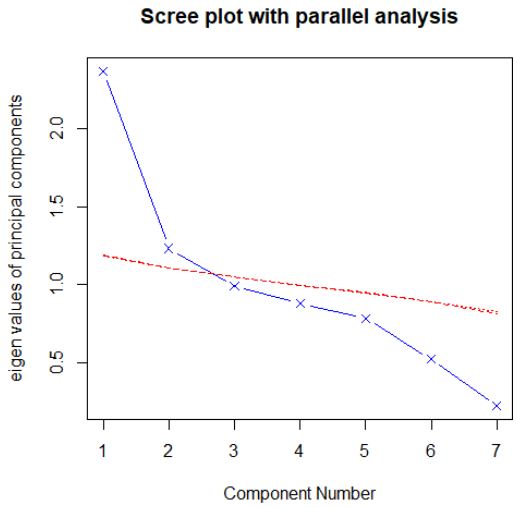

Figure 8 Gravel diagram

5.2 Construct Factor Variables and Calculate Factor Load Matrix

Common factors are extracted based on the principal component method. To highlight the typical representative variables of each common factor, orthogonal transformation with maximum variance is performed. Use the principal function, in which the rotate parameter is set to' varimax,' and the number of factors is 4, and the variance contribution and factor load matrix of the rotated factors are shown in Table 9 and Table 10.

Table 9 Variance contribution of factors after rotation

Factor	total variance	Variance contribution rate	Cumulative variance contribution rate
1	2.225	0.318	0.318
2	1.142	0.163	0.481
3	1.051	0.150	0.631
4	1.049	0.150	0.781

Table 10 Factor Load Matrix after Rotation

Variable	F1	F2	F3	F4
Earnings per share	**0.916**			
Net assets per share	**0.796**	-0.114		
Return on net assets	**0.725**	0.168	0.155	0.280
Total asset turnover			**0.962**	
Operating profit rate				**0.969**
Gross margin	0.460	**0.502**	-0.307	0.149
Current Ratio		**0.918**		

F1 has a significant load on earnings per share, net assets per share, and return on net assets, reflecting the level of return given by listed companies to shareholders, and is called return shareholder ability factor. F2 is called the short solvency factor. Since F3 has a significant load on the turnover rate of total assets, reflecting the operational capability of listed companies, we could call it the functional capability factor. F4 could be called the profitability factor. Compared with the principal component, the four factors have more explicit meanings and are easier to explain.

5.3 Factor Score

Take the variance contribution rate of each factor α_i as the weight, construct the comprehensive evaluation function, and then calculate the overall score of each company according to the following formula.

$$F = \sum_{i=1}^{4} \alpha_i F_i / \sum_{i=1}^{4} \alpha_i$$

$$F = (0.318F_1 + 0.163F_2 + 0.150F_3 + 0.150F_4)/0.781$$

Table 11 Top Ten Factor Scores

	F1	F2	F3	F4	F	F.rank
Changchun Hi-Tech	9.317575	-0.85496	-0.77246	-1.86284	3.140432	1
Wuliang liquid	6.12079	0.856042	0.42234	-0.61139	2.650335	2
Minsheng	-1.61141	12.95148	0.697908	-0.37531	2.057662	3
Gujinggong liquor	4.462057	0.264649	0.354064	-0.40119	1.875811	4
Luzhou laojiao	3.954552	1.089704	-0.22498	0.048569	1.811983	5
Yunnan Baiyao	4.390539	-0.27689	0.031057	-1.17457	1.524542	6
Huadong medicine	2.270141	0.324082	2.011786	0.253899	1.433527	7
Shuanghui Development	1.486798	0.233006	3.678161	0.285113	1.419933	8
Tianyin	-0.07803	0.157501	6.684193	0.100372	1.305019	9
National medicine	3.522508	-1.38404	1.365638	-1.11969	1.208524	10

List the top ten listed companies and the bottom ten listed companies with comprehensive factor scores. The top ten companies with comprehensive factor scores are the same as the top ten companies with principal component scores, both of which are large-scale pharmaceutical and liquor listed companies. In the factor overall score, "Guilin Tourism" and "Zhangjiajie" in tourism category, "Xi 'an Diet" in catering category and "Shanxi Road and Bridge" in transportation category are ranked in the bottom ten, and the listed companies in the bottom ten are all small in scale.

Among them, "Changchun Hi-Tech" and "Wuliang liquid," which ranked first and second, scored higher in factor F1, reflecting the high returns given to shareholders by the two listed companies.

Table 12 Top Ten Factor Scores

	F1	F2	F3	F4	F	F.rank
Xi 'an catering	-1.51213	-0.81625	-0.56521	-0.47644	-0.98811	393
Guoxin healthy	-2.48456	1.01299	1.179292	-2.33571	-1.0339	394
Xinhualian	-1.84774	-0.64189	-0.82728	-0.238	-1.09458	395
Tieling new city	-0.80999	-1.18012	0.428308	-3.588	-1.18201	396
Jiakai city	-1.79963	-0.86949	-0.88411	-0.90887	-1.26147	397
Guilin tourism	-1.48386	-2.10704	0.064181	-1.32171	-1.28281	398
Ziguang Xueda	-2.63095	-0.81651	-0.50472	-0.76443	-1.49092	399
Zhangjiajie	-1.89081	-3.57579	1.397916	-1.96272	-1.61783	400
Shanxi Luqiao	-2.75662	-2.96792	0.291147	-2.20666	-2.10809	401
Tianxiazhihui	0.931269	1.255847	-0.90474	-16.7624	-2.75786	402

6. Cluster Analysis

6.1 Determining the Number of Clusters

Selecting four factors, using factor scores as variables for cluster analysis, and using NbClust function in the NbClust package to determine the number of clusters. When using the sum of squares of deviation method (Ward method) in the systematic clustering method, the results are shown in Figure 9, which indicates that it is better to divide into eight categories. When k-means is used for classification, the function results shown in Figure 10 are better divided into three groups.

However, if clustering is carried out according to the number of clusters mentioned above, there is only one sample in some categories, and the classification effect is not good. Therefore, based on

NbClust function and the actual situation, the number of clusters using the systematic clustering method and the k-means method is determined to be four.

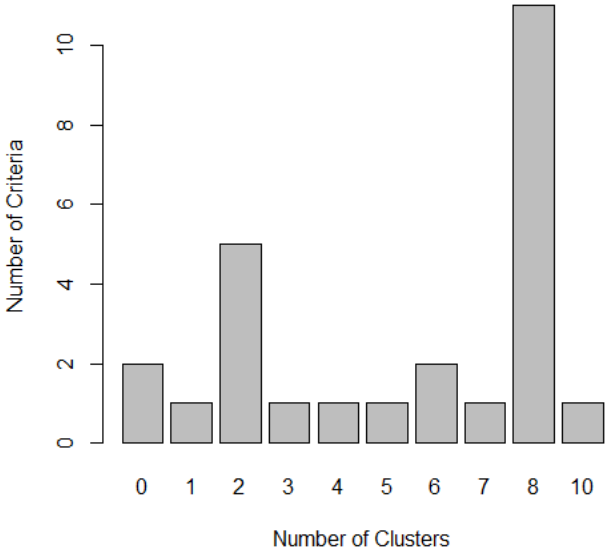

Figure 9 Cluster number discrimination diagram of system clustering method

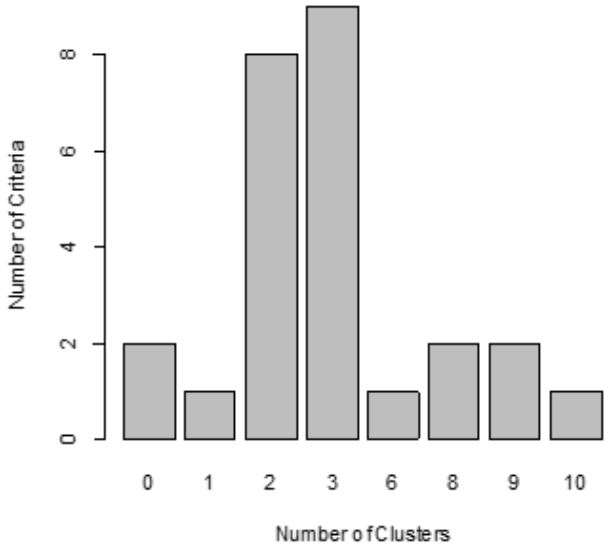

Figure 10 discriminant diagram of clustering number by k-means clustering method

6.2 Systematic Clustering Method

See Figure 11 for the pedigree diagram obtained by using the sum of squares of deviation method (Ward method) in the systematic clustering method, and see table 13 for the clustering results.

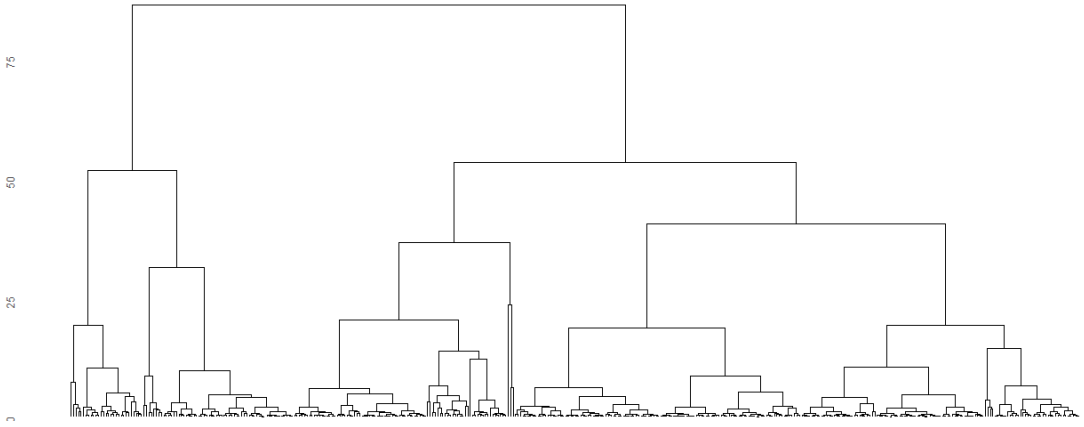

Figure 11 System cluster pedigree diagram

Table 13 Clustering results

Classification serial number	Quantity	Company name (list the first category)
1	29	Wanke A, Sinopharm, Desai Battery, Shentianma A, Tianjian Group, Midea Group, Livzon Group, Suihengyun A, guangyu development, Yunnan Baiyao, Luzhou Laojiao, Weifu Hi-Tech, Gujing Distillery, Jilin Jidong, Gree Electric Appliances, Changchun Gaoxin, Shangfeng Cement, CITIC Special Steel, Zhenhua Technology, Wannianqing, faw liberation, Wuliangye, Shunxin Agriculture
2	88	
3	225	
4	60	

Listed companies are classified into four categories by systematic clustering, and the ranking of overall scores of various factors is shown in Figure 12, among which the first category ranks high. It can be seen from Table 13 that the first type of company is large-scale liquor and pharmaceutical companies.

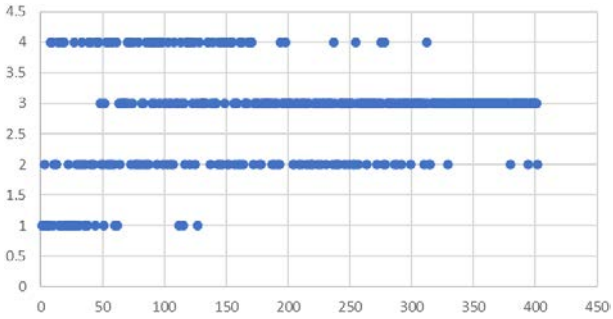

Figure 12 Systematic clustering of various comprehensive scores

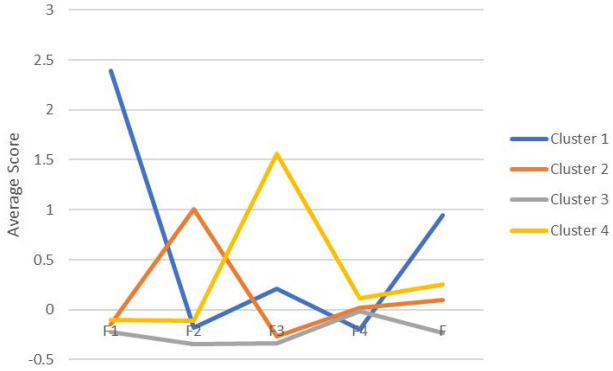

Figure 13 Systematic clustering of various scores

Figure 13 shows the average scores of various factors, among which the first type of companies have the highest average ratings of return shareholders' ability factors, substantial return shareholders' ability, and the most significant average comprehensive scores.

6.3 K-Means Clustering Method

Using k-means rapid clustering method, listed companies are divided into four categories, the clustering results are shown in Table 14, the overall scores of various factors are shown in Figure 14, and the scores of different types are shown in Figure 15.

We can see from Figures 14 and 15 that the third category ranks at the top, and one sample ranks at the bottom. The average score of return shareholder's ability factor is much higher than the other three categories, but the profitability factor's average rating is low. The analysis found that "Tian Xia Zhihui" ranked last in the overall score of factors, and its lower F4 score resulted in the lower average rating of the third category F4. Comparing the performance of "Tian Xia Zhihui" with other

companies, we find that Tian Xia Zhihui was not suitable. This situation may be related to the sensitivity of the k-means method to outliers, and the variance of operating profit margin is significant, which may lead to poor classification results using the k-means method.

Table 14 Clustering results

Classification serial number	Quantity	Company name (list the third category)
1	29	
2	48	
3	7	National Medicine Consistency, Yunnan Baiyao, Luzhou Laojiao, Gujing Distillery, Changchun Gaoxin, Tian Xia Zhihui, Wuliangye
4	318	

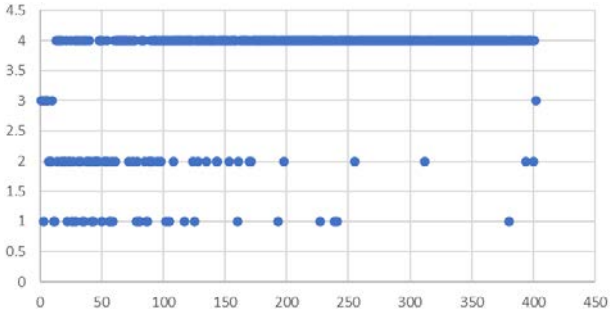

Figure 14 ranking of all kinds of comprehensive scores in k-means clustering

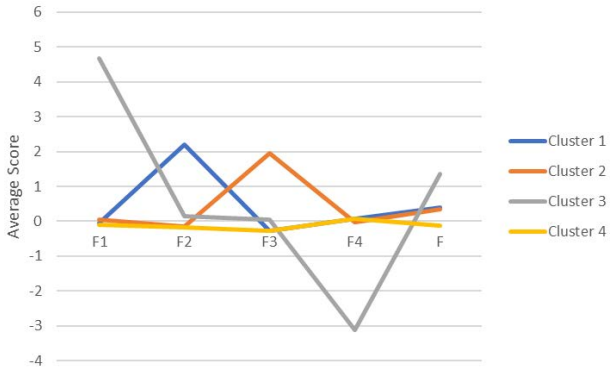

Figure 15 k-means clustering scores

7. Conclusion

Pharmaceutical and liquor companies performed well in the first quarter of 2020. Affected by the epidemic, the demand for medical and epidemic prevention materials such as masks, protective clothing, and respirators increased, so the financial data of pharmaceutical companies performed well in the first quarter. Consumer goods such as alcohol are not affected by the epidemic, so their financial data also show well.

In the epidemic situation, large companies' anti-risk ability is prominent, and the top ten of principal component scores and factor scores are all large companies. Among them, Midea Group is an electrical appliance company. Although electrical appliance companies are greatly affected by the epidemic, Midea Group, as a leading company, has strong resilience, and its principal component score is still relatively high.

From the analysis of the last ten principal component scores, it can be concluded that small companies are greatly affected by the epidemic situation, especially the tourism industry, the transportation industry, and the catering industry.

Affected by the epidemic in 2020, most listed companies are disturbed to vary degrees, and their

performance fluctuates obviously in the short term. Large-scale companies should be selected to avoid risks, and massive consumption industries such as medicine and alcohol should be chosen for investment, while small companies, such as transportation, tourism, and catering, which are greatly affected by the epidemic, should be avoided.

References

[1] Wang Zhaoxi. Application of factor analysis and cluster analysis in diversified financial stock investment [J]. Guangxi Quality Supervision Herald, 2019(09):79-81.

[2] Hu Shuwen, Xu Jianwu. Application of Principal Component Analysis and Factor Analysis in Chinese Stock Evaluation System [J]. Journal of Chongqing University of Technology (Natural Science), 2017, 31(05):192-202.

[3] Zhang Zongqiang, Ren Jingxi. Factor Analysis of Investment Value of Listed Automobile Companies in 2002 [J]. Value Engineering, 2004(05):109-112.

[4] Hua Han, Fei Tang. Investment Value Analysis for Listed Companies of China Communications Industry[C]. Intelligent Information Technology Application Association. Proceedings of the 2011 International Conference on Computing, Information and Control(ICCIC 2011 Part1).Intelligent Information Technology Application Association, 2011:490-495.

[5] Ruohan Sun. Analysis of Investment Value of Listed Companies in New Energy Lithium Battery Industry Based on Factor Model[C]. Hainan University, Sanya University, Xiamen University Tan Kah Kee College. Proceedings of 4th International Conference on Social Science and Higher Education(ICSSHE 2018)(Advances in Social Science, Education and Humanities Research, VOL. 181).Hainan University, Sanya University, Xiamen University Tan Kah Kee College, 2018:875-878.