

Stock Evaluation Based on Factor Analysis and Clustering

Wei Wang

School of Economics
Shanghai University
Shanghai, China
1754886359@qq.com

Abstract—The purpose of this paper is to assess the value of listed companies. We used R software for factor analysis and K-means clustering to analyze the eight financial indicators of these companies. Our innovation was to use the most simplified and the most condensed indexes for decision making. The comprehensive results of the two methods show that the sample stock is divided into three categories. Combining CAPM model, we put forward two investment strategies suitable for long-term and short-term investors respectively.

Keywords—*factor analysis; clustering; Stock evaluation*

I. INTRODUCTION

At present, China's stock has become an important investment target for many investors. More and more stock picking theory began to be treated seriously. Individual investors prefer to regard MACD, trading volume and other technical indicators as decision-making tools of stock investment, while the cost of stock picking of institutional investors is high. Even in the weak market condition of the effective market hypothesis, technical analysis has been invalid, but the fundamental analysis can also get excess returns. So it is essential to find effective and simple indicators of fundamental analysis.

In order to examine the rapid development of the listed companies in financial and information industry, this paper will analyze 150 stocks in CSRC(China Securities Regulatory Commission) industry in Wind database as the sample. We change eight indicators such as operating profit, total profit, and earnings per share and so on into two factors through factor analysis to make the indicators as simple as possible. Based on these two indicators, we construct a comprehensive evaluation pattern considering a number of indicators in value analysis to evaluate the 150 stocks of listed companies. We do further analysis through clustering for investors with different investment preferences.

II. LITERATURE REVIEW AND THEORETICAL OVERVIEW

A. Literature Review

Yingying Zhang, Fuyun Ren and Xuemin Wang have explored factor analysis for stock evaluation. Yingying Zhang (2015) used the object-oriented robust factor analysis to analyze the ten major financial indicators of 957 listed

companies in the third quarter of 2013[1]. It is found that the results of the classical estimator and the robust Sde estimator are different. By combining the principal factor method and the robust Sde estimator, the ten financial indicators are reduced to three factors: the yield factor, the scale factor and the asset main factor. Stock is evaluated comprehensively according to the three factor scores of each stock. Fuyun Ren (2005) used factor analysis to analyze the financial data of China's securities market in mid-2005 and the corresponding financial data of the listed companies, and rank the investment value of each industry, which represents the relative investment value of these industries[2]. Through the ranking analysis, we can see that the index system is reasonable. Finally, the influence of the sample data selection method on the result of factor analysis is analyzed, and the whole analysis system is improved. Xuemin Wang (2004) used SAS software to analyze the ten major financial indicators in the financial reports of 604 listed companies in Shanghai, and summed up the ten indicators into three factors, each with a specific meaning[3]. Finally, a comprehensive evaluation of the stock is made on the basis of the three factor scores for each stock.

Based on the factor analysis method, this paper combines the clustering algorithm to make a relatively comprehensive evaluation of the stock of different listed companies.

B. Theoretical Overview

1) Factor Analysis

According to Richard A. Johnson, the essence of factor analysis is that, as far as possible, several potential but unobservable random variable (that is factor) are used to describe covariance relationships among many variables[4]. It was first proposed by C.Spearman, British psychologist.

The steps to establish a factor analysis model are as follows.

- a) *Preprocess data .*
- b) *Determine the number of factors.*
- c) *Select the factor and rotate the factor.*
- d) *Explain the results.*
- e) *calculate the factor score.*

2) Clustering

Clustering analysis is a data reduction technique designed to expose a subset of observations in a data set. It can make a

large number of observations reduced to several categories. The similarity of observations in each category is higher than that between categories.

The general steps of clustering analysis are:

- a) *Select appropriate variable .*
- b) *Data standardization.*
- c) *Select appropriate clustering algorithm.*
- d) *Determine the number of classes.*
- e) *Obtain final clustering result and interpret the class.*

Because of the large amount of data in this paper, we use K-means algorithm to cluster.

3) CAPM model

Capital Asset Pricing Model (CAPM) was developed by William Sharpe, American scholar, in 1964 on the theory of portfolio theory and capital market theory, which focuses on the relationship between the expected return rate on assets and risk assets in the securities market. The CAPM model is as in

$$E(R_i) - R_f = \alpha + \beta_i (E(R_M) - R_f) \quad (1)$$

where: $E(R_i)$ is expected return of the risky asset i . R_f is risk-free rate (assumed constant). $E(R_M)$ is expected return on the market portfolio. β_i is measure of systematic risk of asset i . (referred to as *beta*). α is excess return.

III. EMPIRICAL ANALYSIS

A. Data Collection

The sample of this study is 150 stocks in China Securities Regulatory Commission industry, including banking, securities, insurance and information industry. The data come from the financial reports of the corresponding listed companies in Wind database in fiscal year 2016. The original data are omitted.

TABLE I. INDICATORS LIST

Variable	Indicator	Unit	Directionality
X1	operating income	hundred million yuan	positive
X2	operating profit	hundred million yuan	positive
X3	total profit	hundred million yuan	positive
X4	earnings per share	yuan	positive
X5	net assets per share	yuan	positive
X6	ROE ^a	%	positive
X7	total assets	hundred million yuan	positive
X8	capitalization	hundred million shares	positive

^a Rate of Return on Common Stockholders' Equity

Taking into account the financial indicators contained in the existing literature, this paper selects eight financial indicators of listed companies as shown in Table 1.

Because past financial indicators of listed companies considered in data analysis are more complex, not easy to analyze directly, this paper only selects the eight financial

indicators shared by the existing literature[6]. The selected indicators are more concise than others. The final results of data analysis make investors not confused by unimportant indicators, but pay attention to the most important factor of stock selection and investment.

B. Factor Analysis

1) Data preprocessing

First of all, according to some data screening principle mentioned by Man Guo, excluding the company whose financial data and transaction data is incomplete in 2016, excluding the stocks in ST or PT, and eliminating the companies listed in 2016. Finally 150 listed companies' financial data are obtained.

Secondly, the numerical value of indicators selected in this paper is different and data units are not the same. According to the method provided by Xuemin Wang, we need to standardize the indicators in order to make factor analysis treat each indicator equally[3]. The standardized data correlation coefficient matrix shown in Table 2, shows that there is a high correlation between the majority of variables in eight indicator variables, so it is suitable for factor analysis.

TABLE II. CORRELATION COEFFICIENT MATRIX

Value	X1	X2	X3	X4	X5	X6	X7	X8
X1	1.000	0.976	0.975	0.240	0.290	0.055	0.968	0.925
X2	0.976	1.000	1.000	0.300	0.310	0.091	0.967	0.934
X3	0.975	1.000	1.000	0.300	0.310	0.092	0.965	0.934
X4	0.240	0.300	0.300	1.000	0.760	0.720	0.262	0.139
X5	0.290	0.310	0.310	0.760	1.000	0.228	0.259	0.150
X6	0.055	0.091	0.092	0.720	0.228	1.000	0.074	0.039
X7	0.968	0.967	0.965	0.262	0.259	0.074	1.000	0.895
X8	0.925	0.934	0.934	0.139	0.150	0.039	0.895	1.000

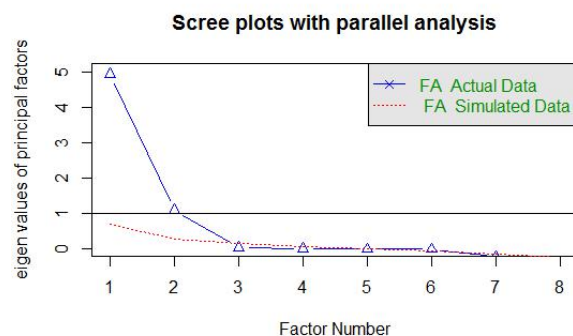


Fig. 1. Scree plot

2) Determining the number of common factors and making factor rotation

From the sample correlation coefficient matrix, factor analysis is carried out, and eigenvalues and eigenvectors are computed. There are two methods in R software to determine the number of common factors, one method is counting the number of common factors whose eigenvalues is larger than 1, the other method is to determine the number of common factors before the inflection point through the scree plot with

parallel analysis. As illustrated in Fig. 1, two common factors are needed.

Factor extraction is carried out by using the number of common factors. After calculating the total variance contribution rate, the cumulative total variance contribution rate of the two common factors can be as high as 85%. It implies that the factor analysis model can achieve the goal of simplifying indicators while maintaining the total variance of 85%.

In order to further explain the significance of the factor loading matrix, the factor variances after varimax orthogonal rotation are continuously analyzed, and the output is shown in Table 3.

The second and the third columns in the Table 3 are eigenvectors corresponding to the two eigenvalues. You can see the first three indicators and the last two indicators in Table 3 (total operating income, operating profit, total profit, total assets and capitalization), all have large positive loading in the first common factor F_1 . Table 2 also shows their correlation coefficients are close to one, so the first factor can be represented the company's size factor. The bigger the first factor is, the larger the size of the listed company is. The fourth, fifth and sixth indicators (earnings per share, net assets per share and ROE (Rate of Return on Common Stockholders' Equity)) all have large positive loading in the second common factor F_2 , and can be interpreted as the company's profitability factor. The bigger the second factor is, the better the profitability of the listed company is. These will be verified below.

3) Calculating factor scores

Based on the analysis above, we can obtain the weight matrix of factor score after the rotation.

Based on the weight matrix of factor score, the score of each factor of each stock can be further calculated. Then, we compute the weighted mean whose weights are the proportion of the contribution ratio of each factor variance (58% and 27%) in the contribution ratio of total variance (85%), and the comprehensive score of each stock is obtained, as in

$$F = (0.58 F_1 + 0.27 F_2) / 0.85 \quad (2)$$

TABLE III. FACTOR LOADING MATRIX

Value	X1	X2	X3	X4	X5	X6	X7	X8
MR1	0.95	0.98	0.98	0.47	0.44	0.23	0.95	0.89
MR2	-0.24	-0.19	-0.18	0.88	0.63	0.69	-0.21	-0.32

TABLE IV. THE RANK OF OVERALL SCORE

Rank	Name	Score	Rank	Name	Score
1	SPDB	6.516	141	Zhaori Technology	-0.302
2	CITIC Bank	4.780	142	Gao Weida	-0.303
3	Bank of Shanghai	2.102	143	Rongji Software	-0.307
4	Guotai Junan Securities	1.303	144	Jianqiao Communication	-0.309
5	Bank of	1.269	145	Hai Lianxun	-0.320

Rank	Name	Score	Rank	Name	Score
	Ningbo				
6	Jiangsu Bank	1.223	146	Jiachuang Shixun	-0.322
7	CITIC Securities	1.195	147	Yi Lianzhong	-0.324
8	Haitong Securities	0.859	148	Anshuo Information	-0.330
9	Bank of Guiyang	0.773	149	Huahong Jitong	-0.334
10	Flush	0.670	150	Mei Ansen	-0.335

The rank of overall score of 150 listed companies is listed in Table 4, which shows top and post ten stocks.

C. Clustering

On the basis of factor analysis, we use clustering method. Taking the two important factors above as variables, we use K means algorithm to cluster the 150 stocks, which is to divide similar stocks into the same category. As shown in Fig. 2, we get the highest score of the three categories, which is the optimal number of categories, according to 26 kinds of standards in R software. The number of stocks in various categories and various centers of various categories are shown in the Table 5.

D. Result Analysis

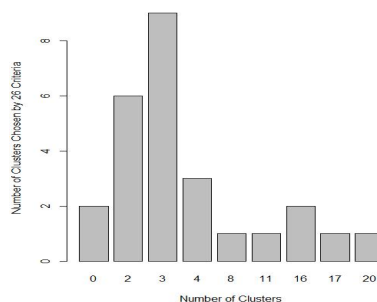


Fig. 2. Determine the number of classes through 26 kinds of standards

According to the results of clustering analysis and Table 5, the first category of clustering results is composed of Shanghai Pudong Development Bank and CITIC Bank. The common characteristics of these two banks are large scale and good profitability, and these two banks are ranked on the top two both in overall score (Table 4) and company's size factor (Table 6) in the result of factor analysis, so the results of two methods are very consistent. From the perspective of CAPM model, the beta coefficient of such stocks is small and the risk of the stock is small, so the short-term expected rate of return is low. If investors have less free time, or prefer value investment, pay more attention to the depth of investment, and are more willing to explore the intrinsic value and growth of listed companies, it is more appropriate to invest in such stocks. They have strong ability to resist risks and have the development prospects in long term.

TABLE V. THE RESULT OF CLUSTERING

Class	Amount of Stocks	Size Factor	Profitability Factor
1	2	8.054	0.48
2	12	-0.394	2.57
3	136	-0.084	-0.23

The second category is based on the Bank of Shanghai, Bank of Ningbo, Flush and Glacier Network as the representative of the 12 listed companies which have strong profitability and small size. The 12 listed companies is also ranked in the top 20 in profitability factor (Table 6) in the result of factor analysis, so the results of the two methods are similar. This problem is interpreted from the perspective of the CAPM model, that the beta coefficient is large and the risk of the stock is high, so the risk premium leads to a high rate of short-term expected return. If investors usually have more leisure or work on full-time stock investment, or prefer high-frequency transactions, pay more attention to the breadth of investment, and are more willing to transact frequently through the law of large numbers to obtain profit, it is more appropriate to invest in such stocks. There will be risks in the long term, but at least recent data show good short-term profitability.

The third category is the remaining 136 listed companies. From the perspective of company's mean of two factors in Table 5, they have small overall size (-0.084) and unremarkable profitability (-0.23). However, Jiangsu Bank, CITIC Securities, Haitong Securities, Guotai Junan Securities, Shenwan Hongyuan and Guosen Securities classified as the third category in clustering analysis, these six listed companies rank the top nine in the size factor ranking of factor analysis, and have also entered the top 15 in the overall ranking. The results of the two methods can be considered comprehensively that part stocks of the third category with comprehensively front ranking can be considered by investors to invest appropriately.

TABLE VI. THE RESPECTIVE RANK OF TWO SCORE

Rank	Name	MR1 Score	Rank	Name	MR2 Score
1	SPDB	8.60	1	Bank of Shanghai	4.37
2	CITIC Bank	7.51	2	Flush	4.09
3	CITIC Securities	1.55	3	Glacier Network	3.22
4	Jiangsu Bank	1.46	4	Huijin Technology	3.16
5	Guotai Junan Securities	1.32	5	Bank of Guiyang	3.09
6	Haitong Securities	1.16	6	Bank of Ningbo	3.08
7	Bank of Shanghai	1.05	7	Wangsu Technology	2.62
8	Shenwan Hongyuan	0.76	148	SPDB	2.05
9	Guosen Securities	0.54	149	Sheng Xunda	1.75

IV. SUMMARY AND RECOMMENDATIONS

In the circumstances that market is not effective in CAPM model, in order to obtain excess returns, it is essential for investors who take stocks as risky assets to select the appropriate stock and the corresponding investment strategy. In order to simplify the investors' stock selection process, a number of financial indicators should be condensed into two major indicators to facilitate the comprehensive stock sorting and cluster analysis. Combining the results of factor analysis and clustering analysis, this paper provide appropriate investment strategy and stock picking guidance for the investors prefer value investment or high-frequency investment respectively. However, there are some tiny shortcomings in this paper. The results of factor analysis and clustering are affected by different data, which may result in great difference, so we need to consider the results of the two methods comprehensively and does more robustness test.

First of all, as a rational investor, they should know themselves and identify their own investment pattern. Secondly, China's stock market is not an effective market, so both fundamental analysis and technical analysis are effective for investors. We can first consider the fundamentals for stock selection and classification. Value investors can invest in business with long-term growth, while high-frequency traders can use technology analysis tools to invest in the stocks with short-term big gains. Finally, it is essential to follow only one strategy in the investment process. If being optimistic about the long-term gains of the listed companies at the beginning, investors should not give up because of the price reduction during the holdings of the stocks so as to get the final long-term gains. On the contrary, if the initial decision is to conduct high-frequency operation, investors should not hesitate to buy and to sell because of the long-term trend of the stock; otherwise they may miss the best trading opportunity.

REFERENCES

- [1] Z.Y. Ying, "Application of Robust Factor Analysis in Stock Evaluation," *Statistics and Decision Making*, vol. 16, pp.76-79, 2015(In Chinese).
- [2] R.F. Yun, "The Application of Factor Analysis in the Evaluation of Investment Value of China 's Stock Market," Zhongnan University, 2005(In Chinese).
- [3] W.X. Ming, "The Application of Factor Analysis in Stock Evaluation," *Mathematical Statistics and Management*, vol. 3, pp. 6-10, 2004(In Chinese).
- [4] J.A. Richard, *Applied multivariate statistical analysis*, 6th ed., Beijing, vol. 9, pp.374-419, 2008(In Chinese).
- [5] Y.S. Si, "Statistical Analysis Based on 100 Stock Index of Central Enterprises," Chongqing University, 2015(In Chinese).
- [6] F.Kun and F.C. Huan, "How the index data is correctly preprocessed in factor analysis," *Accounting Monthly: Theoretical Edition*, vol. 3, pp.85-88, 2013(In Chinese).