Research on Quantitative Investment of the CSI 300 Stocks Based on Monte Carlo Algorithm

Ziwei Wang, Weirui Liu, Yiqi Zheng, and Yanke Wu(✉)

Faculty of Mathematics and Computer Science, Guangdong Ocean University, Zhanjiang 524088, Guangdong, China
yanke.wu@163.com

Abstract. With the development of science and technology, quantitative investment was involved as the three main methods of stock investment together with fundamental analysis and technical analysis. In this paper, a quantitative investment model is proposed to automatically and accurately capture the real-time updated stock data of CSI 300 stocks. Its specific logic is that the K-means clustering model based on factor analysis is used to screen out several good stocks and then automatically calculate the optimal portfolio mode for investors. In addition, using this model, we conducted an empirical study on the Shanghai and Shenzhen 300 stock data from 2018 to 2020. It is concluded that the top ten stocks during this period are ZTE, Gree Electric Appliance, BOE A, Wuliangye, Dongfang Wealth, CITIC Securities, Sany Heavy Industry, Zhejiang Long sheng, Guizhou Mao-tai and China Ping An. Then, for these ten stocks, we allocate them according to the weights of [0.0097, 0.0189, 0.0147, 0.2403, 0.1576, 0.0863, 0.2129, 0.0403, 0.1655, 0.0535], and the expected return is 0.5019 and the volatility is 0.2979. It can be seen that using this model to make quantitative investment can obtain higher returns with lower risks.

Keywords: quantitative investment · factor analysis · k-means clustering · Monte Carlo simulation · Markowitz portfolio theory

1 Introduction

1.1 The Demand for Quantitative Investment in China's Capital Stock Market is Increasing Day by Day, Which is Mainly Reflected in the Following Aspects

1.1.1 A Large Proportion of Individual Investors

The number of investors has reached 200.087 million on February 25, 2022 when it exceeded 150 million since March 2019, and the private investors account for more than 90% of all A- and B-share accounts. Due to their lack of perfect and long-term investment concept, it is easy to cause stock market fluctuations, so quantitative investment is suitable for the Chinese market, according to China Securities Depository and Clearing Corporation Limited (CSDC).

Z. Wang, W. Liu and Y. Zheng—These authors contributed to the work equally and should be regarded as co-first authors.

G. Guan et al. (Eds.): ICBDSS 2022, AHCS 8, pp. 359–371, 2023.
https://doi.org/10.2991/978-94-6463-064-0_39

1.1.2 A Continuous Advance in the Science and Technology

Thanks to the rapid advance of network electronic equipment, a great advance has been made in the information transmission speed, which makes it more difficult to obtain high returns only by relying on the basic research of the stock market. In this case, the introduction of quantitative investment strategies in the A-share market can avoid the adverse effects caused by irrational investment strategies to the maximum extent. Therefore, quantitative investment in China is bound to usher in an unprecedented period of development.

1.1.3 Quantitative Investment Has a Lot of Room for Development in China's A-share Market

On the one hand, overseas financial markets develop relatively rapidly, while China's A-share market and quantitative investment develop relatively slowly, which means that there are fewer competitors in the Chinese market, so China embraces a broad prospect for the quantitative investment. On the other hand, quantitative funds account for only 1% to 2% of China's securities funds with a scale of more than RMB16 trillion, which is still in its infancy when compared with the overseas market where the quantitative and programmed investments account for more than 30% of total assets, therefore, there is huge room for quantitative investment in China.

This paper mainly studies the CSI 300 stocks in the stock market for the following reasons: According to the latest CSDC data, there are 3,688 companies listed on the Shanghai and Shenzhen Stock Exchanges so far, including 1,514 companies in Shanghai Stock Exchange and 2,164 companies in Shenzhen Stock Exchange. This rapid growth will pose huge challenges for the traditional investment methods. Also, the A-share market fund total scale continues to expand, with a growth trend over the number of stocks. Although the number of stocks in the index accounted for only about 8% of the total A-shares as of December 31, 2019, its total market capitalization exceeded $38.98 trillion, accounting for about 60% of the total A-shares market capitalization. As a result, CSI 300 stocks will be selected as the main research object.

1.2 Literature Review

In 1952, Professor Harry Markowitz proposed the theory of portfolio selection, which is the beginning of portfolio theory [1]. In 1963, William F. Sharpe developed an efficient computational method to analyze available capital and construct a portfolio with desired properties [2]. In 1993, Fama and French used market risk premium factor, company market value factor and book-to-market ratio factor to regressed stock return rate and explain stock return rate [3]. In 2009, By removing the limitations of the original model and incorporating the return distribution, Rom and Ferguson extend the model by using the variance of returns as a measure of investment risk [4]. Later, Woodside-Oriakhi et al. introduced a comprehensive model for portfolio optimization and implemented it by quadratic programming of complex integers [5]. Mittal and Mehlawat developed a model involving transaction costs and used genetic algorithms to optimize [6]. With the development of quantitative investment in the international investment market, the research on quantitative investment has gradually arisen in China. In 2014, Li Huilan

utilized historical data to establish a reasonable mathematical model, and then introduced computer programming language to write trading programs, using the program to analyze the risk and return of securities, and forecast the market, so as to obtain excess returns [7]. Based on the Markowitz portfolio theory, this paper adopts K-means clustering model based on factor analysis to quantitatively analyze CSI 300 Index stock selection, and designs a quantitative investment framework combined with portfolio of Monte Carlo (MC) denoising.

To study the influence of the fluctuation of Shanghai and Shenzhen Stock markets, this paper selects CSI 300 Stock from 2018 to 2020 as the research object.

2 Dataset Preparation

2.1 Sources

National Bureau of Statistics; Commerce Department; China Securities Depository and Clearing Corporation Limited; fx112 Financial Network; Sohu Finance Network; Oriental Fortune Network; TongHuaShun Stock; Morningstar Network; tushare official website.

2.2 Data Pre-processing

1. Check the continuity of dates in the data obtained for each stock. If there is not a closed period, the date is intermittent and the duration of the intermittent period is more than 3 days, the defect data in the stock will be deleted.
2. The missing values were completed by cubic spline interpolation.
3. Calculate the daily ma5 average value for each stock from its closing price.
4. All the data processed according to the three steps were sorted into a CSI 300-share dataset.

3 A k-means Clustering Model Based on Factor Analysis

Considering the data preprocessing the data set besides ma5 average value of each stock in the rest of the individual characteristics, the correlation of stock parameters should be first to factor analysis. In the stock preprocessing data, the time length obtained by downloading the data is used to calculate the average value of each parameter of each stock. The statistical factor analysis model was established after the maximum minimization standardization and dimensionality elimination processing. In our analysis of the variance table, you can figure out what each factor means together. According to the factor analysis model, we calculate the corresponding factor score of each stock and organize it into a M_score matrix (each column is the code of each stock, and the number of columns is the best number of factors). After the matrix results of factor analysis and the meaning represented by each factor are obtained, the matrix is processed with z-score standardized data to eliminate dimensionality. According to the processed data, the K-means clustering model is established. Finally, in the filter step, you can get the screening results (N is a number, you can manually decide what n is) of the top N stocks and the meaning of each cluster determined by the model Carlo algorithm.

4 Stock Screening

We use the k-means clustering model based on factor analysis to screen the stocks in the quantitative investment framework. Here's an example that we calculated using the CSI 300 Stock example. This paper downloaded and processed the data of CSI 300 Stock from 2018 to 2020 for empirical solution.

First of all, we need to conduct KMO and Bartlett test the processed data to judge whether factor analysis can be carried out.

The KMO test result shows that the KMO value is 0.752. Passing the test (>0.6) means that there is correlation between each parameter characteristic of the loaded stock, which meets the requirement of factor analysis. The result of Bartlett sphericity test shows that the significance P value is 0.000, which is significant at the level, the null hypothesis is rejected, the variables are correlated, the factor analysis is effective, and the degree is general.

Table 1. A sample table of data for one of the stocks after the data is preprocessed

code	Trade date	open	high	low	close	…	pct chg	vol	amount	ma5
300059.SZ	20201224	26.56	27.04	26.41	26.53	…	-0.4129	1360582	3631873	28.35
300059.SZ	20201223	26.33	27.1	26.32	26.64	…	1.2158	2070593	5511200	27.46
…	…	….	…	…	…	…	…	…	…	

Table 2. Variance interpretation table

Total variance explanation						
Number of components	eigenvalue			Explained rate of variance after rotation		
	eigenvalue	Percentage of variance	cumulation	eigenvalue	Percentage of variance	cumulation
1	6.209	68.992%	68.992%	5.872	65.24%	65.24%
2	1.361	15.118%	84.11%	1.048	11.645%	76.885%
3	1.000	11.107%	95.217%	1.030	11.443%	88.328%
4	0.379	4.207%	99.424%	0.999	11.096%	99.424%
5	0.052	0.576%	100.0%			
6	0.000	0.0%	100.0%			
7	0.000	0.0%	100.0%			
8	0.000	0.0%	100.0%			
9	-0.000	-0.0%	100.0%			

Fig. 1. Load matrix heat map

Table 3. Factor analysis table

Factor weight result			
Value name	Explained rate of variance after rotation	Cumulative variance explained rate after rotation	weight
Factor 1	0.652	0.652	65.618%
Factor 2	0.116	0.769	11.712%
Factor 3	0.114	0.883	11.509%
Factor 4	0.111	0.994	11.16%

According to Table 2, when the principal component is 4 in the variance interpretation table, the eigenvalue of total variance explanation is lower than 1.0, and the contribution rate of variable explanation in the eigenvalue reaches 99.424%. It can also be seen intuitively in the gravel figure that it tends to be flat when the number of factors is 4. In the selection of the number of factors in this paper, we choose 4 as the four main factors of CSI 300 Stock from 2018 to 2020.

The post-rotation factor (Factor i (i = 1, 2, …,4)) loading coefficient determined by the four factors obtained above is intuitively shown in the thermal diagram as Fig. 1, from which the importance of hidden variables in each principal component can be analyzed. The hidden variable analysis of each factor can be carried out in combination with specific business.

Table 3 shows the weight calculation results of the root factor analysis of factor analysis, which shows that the weight of Factor 1 is 65.618%, Factor 2 is 11.712%, Factor 3 is 11.509%, Factor 4 is 11.16%. The maximum value of the index weight is Factor 1 (65.618%). The minimum value is Factor 4 (11.16%).

Table 4 is the component matrix table, which explains the factor score coefficient (principal component load) contained in each component, which is used to calculate the component score and find the principal component formula. Let: x_{open} is the daily

Table 4. Composition matrix

Composition matrix table				
Value name	Factor 1	Factor 2	Factor 3	Factor 4
open	0.159	0.047	-0.05	0.396
high	0.159	0.048	-0.051	0.397
low	0.159	0.046	-0.05	0.395
close	0.159	0.047	-0.05	0.396
pre_close	0.159	0.047	-0.05	0.396
change	0.154	0.135	0.001	0.248
pct_chg	0.02	0.722	-0.062	0.322
vol	-0.016	-0.048	0.972	0.544
amount	0.047	0.115	0.265	2.391

opening stock parameter, x_{high} is the daily high price stock parameter, x_{low} is the daily lowest stock parameter, x_{close} is the daily closing stock parameter, x_{pre_close} is the daily stock parameter of yesterday's closing price, x_{change} is the daily rise and fall point stock parameters, x_{pct_chg} is the daily rise and fall of stock parameters, x_{vol} is the daily volume stock parameter, x_{amount} is the daily turnover of stock parameters. The linear relationship between component matrix and variables is:

$$\begin{aligned} F1 = {} & 0.159x_{open} + 0.159x_{high} + 0.159x_{low} + 0.159x_{close} + 0.159x_{pre\_close} \\ & + 0.154x_{change} + 0.02x_{pct\_chg} - 0.016x_{vol} + 0.047x_{amount} \end{aligned}$$

$$\begin{aligned} F2 = {} & 0.047x_{open} + 0.048x_{high} + 0.046x_{low} + 0.047x_{close} + 0.047x_{pre\_close} \\ & + 0.135x_{change} + 0.722x_{pct\_chg} - 048x_{vol} + 0.115x_{amount} \end{aligned}$$

$$\begin{aligned} F3 = {} & -0.05x_{open} - 0.051x_{high} - 0.05x_{low} - 0.05x_{close} - 0.05x_{pre\_close} \\ & + 0.001x_{change} - 0.062x_{pct\_chg} + 0.972x_{vol} + 0.265x_{amount} \end{aligned}$$

$$\begin{aligned} F4 = {} & 0.396x_{open} + 0.397x_{high} + 0.395x_{low} + 0.396x_{close} + 0.396x_{pre\_close} \\ & + 0.248x_{change} + 0.322x_{pct\_chg} + 0.544x_{vol} + 2.391x_{amount} \end{aligned}.$$

Let the component matrix be M_p and the variable factor matrix M_f, then their principal component formula is:

$$M_{score} = M_p^T * M_f$$

where M_p is composed of coefficients in the composition matrix, and the M_f is $[x_{open}, x_{high}, x_{low}, x_{close}, x_{pre\_close}, x_{change}, x_{pct\_chg}, x_{vol}, x_{amount}]^T$.

The M_{score} score matrix obtained by factor analysis was output, and the factor score data set was constructed. After correlation analysis between the data in the component matrix and factor loading coefficient table and the stock parameters after preprocessing

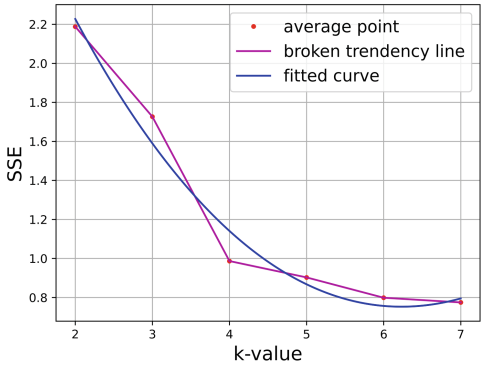

Fig. 2. K-means SSE Elbow chart

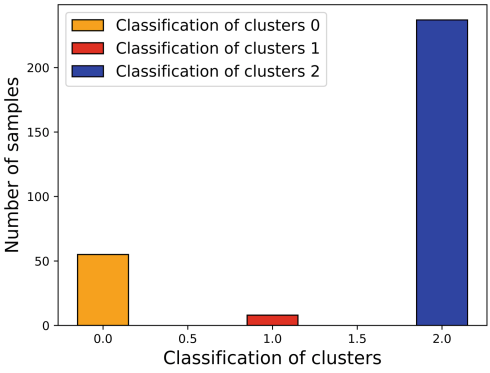

Fig. 3. Cluster sample number plot

data, the meanings represented by the four factors were obtained and named as follows: Factor1: In terms of the parameter characteristics of the stocks in open, high, low, close, pre_close, change, the factor loads are relatively high, and the loads are all close to 1, which can reflect the performance of each stock in price. In this paper, it is determined as the price factor. Factor 2: In terms of pct_chg stock parameter characteristics, the factor load is high and the performance is good, which can reflect the situation of the rise and fall of each stock, so this paper named it as the rise and fall factor. Factor 3: In terms of the vol stock parameter characteristics, the factor loading is high, which generally shows the change of the trading volume of each stock, so this paper named it as the trading volume factor. Factor 4: As the *amount* stock parameter characteristics, the factor load is high, with sound performance, showing the overall turnover situation, so this paper named it as turnover factor.

After substituting in factor analysis and calculation, M_{score} conducted cluster analysis on the characteristic factors of each stock in the data.

According to the SSE measurement quantity in Fig. 2, we can see that the stock classification effect is the best when we select $K = 3$. After the model calculation, the stocks of each classification cluster are counted and the Monte Carlo algorithm

simulation analysis is carried out to know the stock property score represented by each cluster.

According to the classification cluster attributes determined by the stock property score, we can analyze that the comprehensive property of stocks in classification cluster 1 is the strongest compared with the other two clusters, followed by classification cluster 0, and classification cluster 2 is the last. Therefore, we choose the stocks in categorical cluster 1 as the stock data for the next stage of quantitative investment. Since there are less than ten stocks in classification cluster 1, we select the first two stocks from classification cluster 0, which is slightly weaker than classification cluster 1, and add them to our candidate set data. As a result, we obtained the Stock data after CSI 300 Stock screening, namely:

[600030,300059,000063,000858,000651,600519,601318,000725, 600031,600352].

5 MArkowitz's Portfolio Model

We choose Markowitz portfolio theory as the theoretical support when constructing the quantitative investment framework, and take the model designed based on this theory as the backtest part of the quantitative investment framework.

For this model, we take the following assumptions as the premise:

- Investors can reconsider their portfolio based on the probability distribution of asset returns over the trading holding period.
- Investors estimate the risk of a portfolio based on the expected return on assets.
- Investors make decisions based solely on the risk and return of an asset.
- Investors expect maximum return given a certain level of risk. Instead, investors expect minimal risk for a given level of income.

Based on the above hypothesis, we have the following theoretical support:

Let x_i represents the investment proportion of the i asset and r_i represents the expected rate of return of the i asset, this model will use Eq. (1) to calculate the expected rate of return of the portfolio.

$$E(r_p) = \sum_{i=1}^{n} x_i E(r_i) \tag{1}$$

Let σ_i^2 be the variance of the first asset; i and j are different assets. $cov(r_i, r_j)$ is the covariance between asset i and asset j, which can measure the co-activity of the return rate of two assets. ρ_{ij} is the correlation coefficient between asset i and asset j, which can be used to compare the correlation between the two assets. σ_i And σ_j are the standard deviations of assets i and j, respectively.

Equation (2) will be used to calculate the variance of portfolio in this model:

$$\sigma_p^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j cov(r_i, r_j) = \sum_{i=1}^{n} x_i^2 \sigma_i^2 + \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} x_i x_j \rho_{ij} \sigma_i \sigma_j \tag{2}$$

According to Eq. (2), it can be concluded that the risk of portfolio mainly depends on the investment proportion of each asset, the correlation coefficient between different

stocks, and the standard deviation of each asset. This indicates that the model can estimate future returns and risks with the sample average and sample variance of existing stock value data.

Also, this model involves the efficient frontier of asset portfolio [8].

The effective margin of asset portfolio refers to the requirement that the portfolio needs to achieve the minimum risk at a given rate of return or the maximum return at the same risk level.

6 Backtesting

Based on Markowitz's Portfolio Model, we backtest the selected ten stocks.

Firstly, we get the weight of the portfolio based on the minimum volatility is through empirical study, and the expected yield is 0.189, and the global minimum volatility is [0.0404 0.0859 0.1594 0.0196 0.0096 0.0221 0.1164 0.1184 0.2160 0.2122].

Then, we get a portfolio weight of [0.0097 0.0189 0.0147 0.2403 0.1576 0.0864 0.21294522 0.0403 0.1656 0.0535] based on the maximum Sharpe ratio, an expected return of 0.5019, a volatility of 0.2979, and a Sharpe ratio of 1.6848.

And the cumulative return rate comparison chart and effective frontier chart of the two portfolio methods are obtained as Fig. 4–5.

According to Fig. 5, the trend of this effective front line shifts downward to the right, which indicates that the return rate is lower when the risk is higher, it is contrary to the theoretical economics that the expected return of the portfolio is higher and the corresponding risk is higher. The capital market line is consistent with theoretical economics. According to Fig. 4, the cumulative return of the portfolio based on the minimum volatility is worse than the cumulative return of the portfolio based on the maximum Sharpe ratio. Comparatively speaking, the expected return rate of the portfolio based on the maximum Sharpe ratio is 0.5019, which is larger than that of the portfolio based on the minimum volatility and the difference is nearly 30%. But its volatility is 0.2979, compared with the portfolio based on the minimum volatility, the volatility is higher,

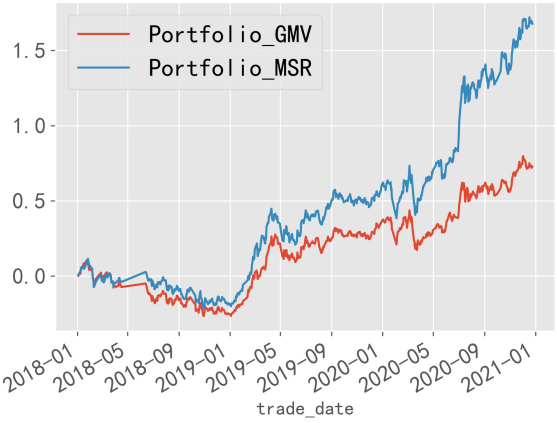

Fig. 4. Comparison chart of cumulative returns of two portfolios

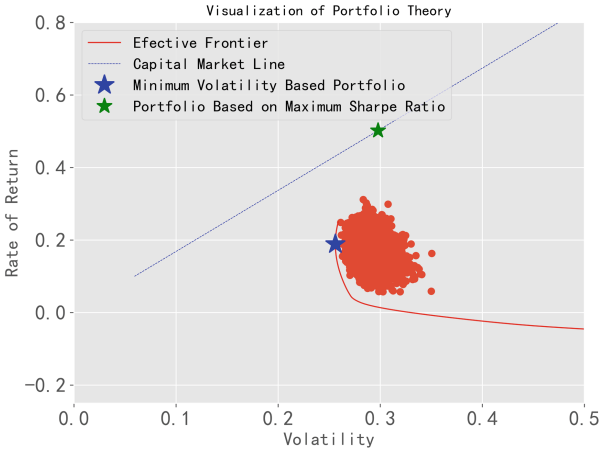

Fig. 5. Global optimal point supplemental plot based on maximum Sharpe ratio

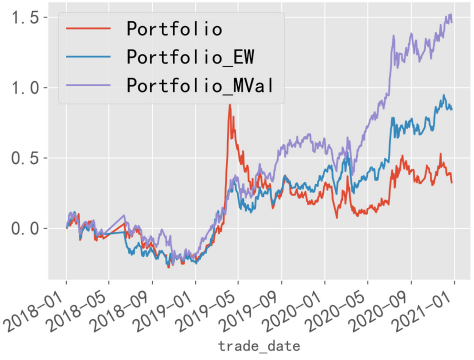

Fig. 6. Graph of the cumulative return of the portfolio

indicating greater risk. According to the Sharpe ratio: each additional unit of risk in the portfolio generates an excess return. According to Markowitz theory, the portfolio based on the maximum Sharpe ratio can be determined as the optimal portfolio, and the weight derived from the Sharpe ratio is the theoretical optimal weight.

We conduct an empirical study on three traditional portfolio methods, including the given weight portfolio, equal weight portfolio, and market weighted portfolio to determine the traditional optimal weight.

We use empirical research to find that the volatility of the portfolio with a given weight is 0.3638, the expected return is 0.1849 and the volatility of the portfolio with equal weight is 0.2865, and the expected rate of return is 0.3056, the volatility of the market-weighted portfolio is 0.3118, and the expected rate of return is 0.4626, and the comparison chart of the cumulative yield of the three portfolio methods is obtained, as shown in Fig. 6.

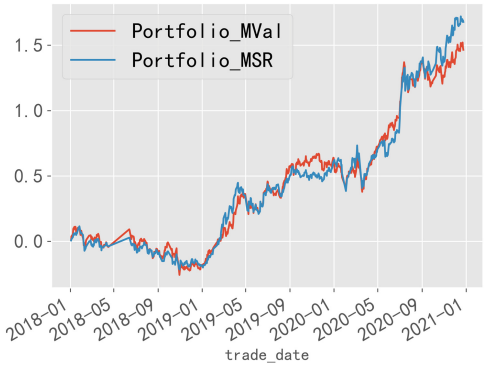

Fig. 7. Comparison chart of the combined returns of the optimal traditional and optimal theoretical portfolios

According to Fig. 6, the market-weighted portfolio has the highest cumulative rate of return and the best effect. After analyzing the expected rate of return of the three portfolios, it is found that the market-weighted portfolio has the highest return. According to Markowitz theory, the market-weighted portfolio is the optimal traditional portfolio, and its weight [0.0222 0.0376 0.0030 0.0872 0.0120 0.0161 0.0104 0.0094 0.7483 0.0537] is the optimal weight of the traditional portfolio.

Finally, we take the weight of the theoretical optimal portfolio as the standard weight and compare it with the traditional optimal portfolio weight to determine the optimal portfolio of the model, it can be intuitively concluded that the optimal theoretical portfolio has the best effect by according to Fig. 7. Compared with the expected return rate and volatility of the optimal traditional portfolio of 0.4626 and 0.3118, the expected return rate and volatility of the optimal theoretical portfolio are 0.5019 and 0.2979. In the case of less risk, therefore, the yield of theoretical optimal portfolio can be higher than that of traditional optimal portfolio.

Therefore, the optimal weight of this model is [0.0097 0.0189 0.0147 0.2403 0.1576 0.0864 0.2129 0.0403 0.1656 0.0535], indicating that in the selected period of time to invest in the selected ten stocks according to this weight, the maximum return can be obtained by investing in the same number of CSI 300 stocks.

7 Conclusions

This paper provides the reader with a complete quantitative investment framework. Firstly, this paper uses Python to obtain CSI 300 Stock data from the official website of Tushare to ensure the availability and authenticity of the data. According to the data, it is substituted into the K-means clustering model based on the factor analysis model for Stock screening, and the stocks most suitable for portfolio in the selected time interval are selected. This model not only overcomes the overelaborate parameters of stock data, realizing the data dimensionality reduction, but also solves the problem of large amount of stock data and slow convergence speed, improving the computational efficiency. According to the selected stocks, the Monte Carlo simulation method is used

to simulate the portfolio by random sampling, and the financial noise in stock selection is eliminated, so as to determine the effective frontier of the portfolio and the optimal portfolio weight of each asset, and overcome the defects of random selection of portfolio weight. Combined with Markowitz theory, the optimal portfolio weight is determined as well. According to the selected stocks, three traditional portfolio methods are used to generate the optimal portfolio weight of each asset, the traditional optimal portfolio is determined by comparing the returns. Finally, the empirical results show that the optimal portfolio based on theory is significantly better than the traditional portfolio, which also means that the reasonable weight allocation of portfolio can diversify the investment risk and increase the expected rate of return of portfolio. In the process of portfolio selection, blindly avoiding risks and selecting the portfolio with the least risk is not the optimal solution. According to the least volatility in the process of empirical portfolios, although the risk is the least of all portfolio, it is also based on the theory of the portfolio, but the yield is not the highest, even below the market weighted portfolio (traditional portfolio). The balance between expected return rate and risk should be further considered, and the Sharp index and other factors should be combined to bear less risk to obtain higher expected return rate. For example, portfolios based on the maximum Sharpe ratio in this paper have significantly better returns and volatilities than portfolios in traditional markets.

References

1. Markowitzharry M. Portfolio Selection[J]. Journal of Finance, 2007, 7(1):77-91.
2. Sharpe W F. A Simplified Model for Portfolio Analysis[J]. Management Science, 1963, 9(2):277-293.
3. Fama, E., French, K. Common Risk Factors in the Returns on Stocks and Bonds [J]. Journal of Financial Economics, 1993, 33(3): 3-56
4. Rom B M, Ferguson K W. PostModern Portfolio Theory Comes of Age[J]. Journal of Investing, 2009, 3(3):11-17.
5. Woodside-Oriakhi M, Lucas C, Beasley J E. Portfolio rebalancing with an investment horizon and transaction costs [J]. Omega, 2013, 41(2):406-420.
6. Mukesh, Kumar, Mehlawat, et al. A multiobjective portfolio rebalancing model incorporating transaction costs based on incremental discounts [J]. Optimization A Journal of Mathematical Programming & Operations Research, 2014.
7. Li Huilan. An Empirical Study on Quantitative Investment Strategies Based on Data Mining [D]. Zhejiang University, 2014.
8. Sun Libo. An Empirical Study of Markowitz's Portfolio Theory Based on Python [J]. Times Finance, 2020(25):3.