DOI: 10.54254/2754-1169/90/20242003

Whether ARIMA-based Portfolio Can Beat the Market Index

Weiping Li[1,a,*]

[1]*School of Economics and Management, Shandong Agriculture University, Taian, China*
a. Liweiping@stu.xaau.edu.cn
**corresponding author*

Abstract: Portfolio optimization is a long-term topic in the financial field, which can maximize returns while minimizing risks. It is widely used in production and daily life, such as stock investment, production optimization, and engineering models. This article selects data from 10 stocks, namely KO, PG, PEP, CL, MDLZ, STZ, PM, KMB, GIS, and K, from January 24, 2024, to June 14, 2024. Firstly, this article uses the ARIMA model to learn the first 70% of the data and predict the last 30% of the data, and then uses portfolio strategy to evaluate the results. This article obtained the maximum Sharpe ratio model on the Monte Carlo effective boundary. By combining the maximum Sharpe ratio model with the SP 500 index within the same time frame, this article concludes that the market performs better. This article provides different portfolio allocation strategies for risk averse investors seeking stable positive returns in turbulent markets.

Keywords: Portfolio optimization, Mean-variance, ARIMA model.

1. Introduction

Since Markowitz initially put forth the classical mean-variance (MV) theory in 1952, it has served as the cornerstone of contemporary portfolio theory. Markowitz's work provides a rational basis for portfolio management decisions by quantifying the risk-return trade-off. Since then, there has been extensive research into developing model variants that are more adaptable to real-life conditions [1]. However, the limitation of MV model is also clear: since the traditional MV model only exploits past information, it can only present the optimal strategy up to the data input [2]. Moreover, the famous Efficient Market Hypothesis (EMH) stipulates that the stock prices already contain and reflect all available information, and therefore in theory there exists no techniques that can produce excess economic profits in the long run [3]. For a long time, there has been a debate about whether the daily stock price is predictable for its intrinsically chaotic, non-parametric properties [4]. The once-dominated theory of EMH underwent skeptics when more and more economists came to believe that at least some predictable patterns that could lead to excess market profit exist [4]. Furthermore, significant incidents in the financial sector have brought spotlight on the significance of diversity. For instance, the 2022 cryptocurrency market crash demonstrated the risks of heavy investment in highly volatile assets. The sharp decline in cryptocurrency values highlighted the necessity of diversification and effective risk management in portfolio construction. Therefore, it remains crucial to delve into the subject of portfolio allocation so that investors can better employ it to mitigate risks while boosting profits. The most pressing issue among all is how to accurately predict future returns, ensuring that the mean-variance model can perform effectively in real-world scenarios.

The Autoregressive Integrated Moving Average model (The ARIMA model) holds a significant position in the field of time series forecasting. Since its inception, the ARIMA model has been widely acknowledged for its solid theoretical foundation and extensive applications. It is extensively used in various fields such as economics, finance, and meteorology to analyze and forecast time series data, demonstrating its powerful utility and predictive capability. In recent years, research on the ARIMA model has been continuously deepened and expanded. Box and Jenkins systematically proposed the modeling steps and methods of the ARIMA model, providing an essential tool for time series forecasting [5]. Subsequently, Makridakis et al., in their time series forecasting competition, confirmed the excellent performance of the ARIMA model in short-term forecasting through a large number of real-world datasets [6]. To enhance the predictive accuracy of the ARIMA model, researchers have attempted to combine it with other forecasting methods. For instance, Khashei and Bijari proposed a hybrid model by integrating ARIMA with artificial neural networks, significantly improving the forecasting performance [7]. Moreover, parameter selection methods based on intelligent optimization algorithms such as genetic algorithms and particle swarm optimization have been widely studied. These methods not only improve the efficiency of parameter estimation but also enhance the predictive accuracy of the model [8]. The application domains of the ARIMA model have also been continuously expanding. For example, Shen et al. used the ARIMA model to predict stock market volatility, achieving remarkable results [9]. In the field of environmental science, Zhang et al. analyzed meteorological data changes using the ARIMA model, providing crucial insights for climate change research [10]. With the advent of the big data era, the scale and frequency of time series data have increased significantly. To address this challenge, Liu et al. made adaptive improvements to the ARIMA model by introducing distributed computing technology, enhancing its applicability in big data environments [11]. These research findings indicate that the ARIMA model has extensive application prospects in handling complex and large-scale time series data.

The purpose of this research is to leverage machine-learning techniques to generate a more informed prediction of future returns and covariance in order to facilitate more effective asset allocation strategies. To achieve this aim, the study first selects 10 distinct sector stocks from the latest SP500 constituents, adhering to certain constraints. The study uses the previous 70 days of stock price data to train a ARIMA neural network to project the next day's stock prices. The projection can be easily shifted into percent change, i.e. daily returns, and a shrinkage method is applied to calculate the covariance matrix. The mean-variance optimization method is then utilized to derive the optimal portfolio weights for each day. The study iterates the above step for the next 30 days, updating the portfolio weights each day based on the latest predictions. Ultimately, after the 30-day period has elapsed, the overall portfolio returns are calculated and benchmarked against the SP500 index. The research results indicate that portfolio construction methods based on complex models may not necessarily perform better than market performance.

2. Data and Methodology

2.1. Data source and pre-process

This study uses the Python package yfinance (https://finance.yahoo.com/) to obtain daily stock data, which is provided by Yahoo Finance. The purpose of this study is to improve the performance of the SP500 index by constructing a portfolio using the latest SP500 constituent stocks. The stocks are ranked based on their average return over a 70-day window period, and then combined with the corresponding company market capitalization to select the most suitable stocks. Finally, a diversified investment portfolio consisting of 10 stocks was constructed, which are representative stocks in the consumer goods industry. The codes of these ten stocks are KO, PG, PEP, CL, MDLZ, STZ, PM, KMB, GIS, K (See Table 1). These stocks belong to the consumer goods industry and have high market stability and risk resistance capabilities, suitable for long-term and short-term investment strategies. At the same time, these companies have good brand reputation and financial health, and their products have a wide range of market demand, making them reliable choices in the investment portfolio.

Table 1: 10 Stocks Selected for Portfolio Optimization

KO	PG	PEP	CL	MDLZ
STZ	PM	KMB	GIS	K

This study examines market data from January 24, 2024, to June 14, 2024, a total of 100 trading days. The model uses the first 70 trading days from January 24, 2024, to May 2, 2024, to predict future stock prices. Then, the remaining 30 trading days from May 3, 2024, to June 14, 2024, are used to evaluate the performance of the investment portfolio. The basic information of these stock prices in the first 70 days, as shown in the Table 2 below:

Table 2: Descriptive statistics of five stocks

	KO	PG	PEP	CL	MDLZ
Mean	60.11	159.17	169.67	86.88	71.48
Std	0.93	2.53	3.73	2.53	2.71
Min	58.06	152.12	162.04	80.08	65.87
Max	62.04	163.84	177.41	92.91	76.87
	STZ	PM	KMB	GIS	K
Mean	256.31	91.87	125.11	66.87	55.75
Std	9.03	2.37	4.98	2.74	1.61
Min	242.55	88.60	118.04	62.34	52.94
Max	272.04	99.02	137.78	71.61	61.28

2.2. Method

The research method includes these steps, as shown in Figure 1. Firstly, select 10 stocks from the SP500 index based on certain constraints. Secondly, use the stock price data from the first 70 days to predict the stock price for the next trading day using the ARIMA model. Simultaneously perform rolling window prediction, rolling forward 30 days with a window of 70 days, and updating the window data for each rolling. Thirdly, use mean variance optimization method and maximum Sharpe ratio to calculate the daily optimal portfolio weights. Fourthly, the cumulative return of the portfolio assets can be obtained by calculating the 30-day weight and collecting the 30-day real stock data. Finally, compare the cumulative yield of the portfolio assets with the SP500 yield for the same period.

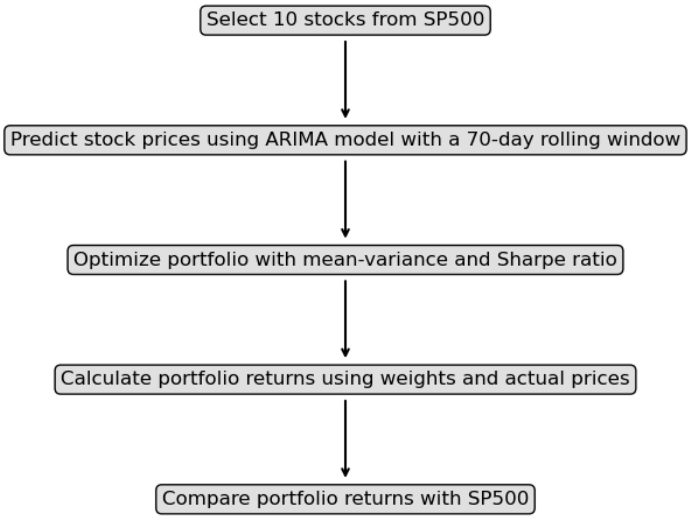
Select 10 stocks from SP500
Predict stock prices using ARIMA model with a 70-day rolling window
Optimize portfolio with mean-variance and Sharpe ratio
Calculate portfolio returns using weights and actual prices
Compare portfolio returns with SP500

Figure 1: Flowchart of the Study

2.2.1. ARIMA model

The ARIMA model offers a comprehensive statistical approach for time series analysis, capturing the linear dependencies within the data. The core principle of ARIMA (AutoRegressive Integrated Moving Average) is to model the differences between consecutive observations to achieve stationarity. This model is particularly useful for forecasting financial time series such as stock prices, where historical data is leveraged to predict future values. Yt represent the observed time series at time t. The ARIMA model is characterized by three parameters: p (the order of the autoregressive part), d (the degree of differencing), and q (the order of the moving average part). The model can be expressed as:

$$\Delta^d Y_t = \sum_{i=1}^{p} \emptyset_i \Delta^d Y_{t-i} + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} \tag{1}$$

Δ^ddenotes the differencing operator applied d times,$\emptyset_i$ are the autoregressive parameters, θ_j are the moving average parameters, and ε_t represents the error term at time t. The ARIMA modeling process involves three key steps: identification, estimation, and diagnosis. Once the model parameters are identified, the estimation phase involves fitting the ARIMA model to the data. This is typically achieved through methods such as maximum likelihood estimation (MLE) or least squares estimation (LSE), which determine the optimal values of $\emptyset_i$ and θ_j. In the context of stock price prediction, the ARIMA model's effectiveness is derived from its ability to capture the underlying temporal dependencies in the data. By differencing the series to achieve stationarity and modeling the autocorrelations, ARIMA can provide robust forecasts that are crucial for financial decision-making.

2.2.2. Mean-Variance Model

The MV model provides a mathematical framework for obtaining the optimum weights for each asset for the investor [11]. The key insight of MV is to find the best portfolio that gives the maximum returns for a given rate of risk. Let w_i be the weight of the i-th asset such that $\sum_i w_i = 1$ and μ_i be the expected return of the i-th asset. Then the expected returns of the portfolio can be expressed as follows.

$$\mu_p = \sum_i w_i \, \mu_i . (2) \tag{2}$$

Denote σ_i as the standard deviation of the i-th asset and ρ_{ij} as the correlation between the returns of the i-th and j-th asset. Then the portfolio return variance is shown as follows.

$$\sigma_p^2 = \sum_i w_i^2 \, \sigma_i^2 + \sum_i \sum_{j \neq i} w_i \, w_j \sigma_i \sigma_j \rho_{ij} . (3) \tag{3}$$

The most ideal portfolios are those that are situated on the efficient frontier. Also, the above equations can be rewritten in matrix form, which is more convenient to implement in programming and calculate efficient frontiers. The objective function for MV is as follows:

$$min \, w^{\mathrm{T}} \sum w - qR^{\mathrm{T}} w \tag{4}$$

where w is a vector of asset weights and Σ is the sample covariance matrix for asset returns. The latter is called the sample covariance because it is computed with historical data directly. In the MV model, two portfolios are interesting: the minimum volatility portfolio, and the maximum Sharpe ratio portfolio. The minimum volatility portfolio is where $q = 0$, i.e., the investor is completely averse to risk. The Sharpe ratio is a popular metric to evaluate risk-adjusted return, which is calculated as:

$$Sharpe \; ratio = \frac{R_p - R_f}{\sigma_p} \tag{5}$$

Where R_f is the current risk-free rate of the market. This study uses the maximum Sharpe ratio portfolio as target portfolio.

3. Results

This study first investigated the predictive performance of the ARIMA model. Using ARIMA for rolling window testing, after completing the stock price prediction, the investment portfolio weight for each asset on a daily basis is obtained through the expected return rate and mean covariance matrix to confirm the investment portfolio. To evaluate the performance of each model, the historical returns of the S&P 500 index during the testing period serve as the market benchmark (See Figure 2). This study conducted post hoc analysis to determine the actual return of the investment portfolio (See Table 3).

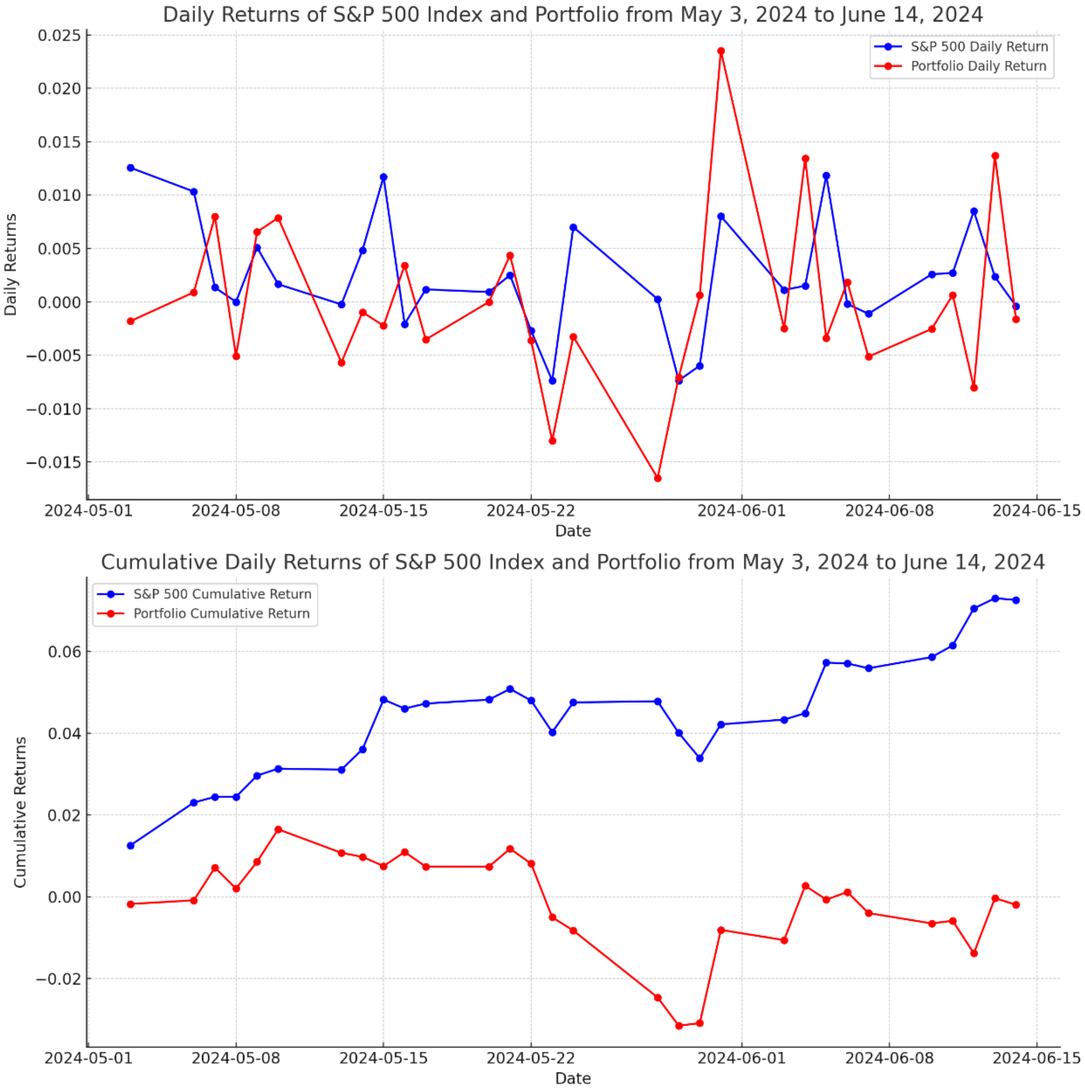

Figure 2: Comparison between SP 500 returns and Maximum Sharpe Ratio Model portfolios

Table 3: Cumulative Daily Return during the observation period

SP 500 Index	7.25%
Maximum Sharpe Ratio Model portfolios	-0.20%

The detailed analysis of the daily and cumulative returns of the S&P 500 index and investment portfolio from May 3, 2024 to June 14, 2024 is as follows. From the data, it can be seen that the daily returns of the S&P 500 index fluctuated significantly during this period, with a maximum daily increase of 1.18% and a maximum daily decline of -0.74%. The cumulative yield of the S&P 500 index has reached a cumulative increase of 7.25%, showing a strong upward trend. In contrast, the performance of the investment portfolio is slightly inferior, and its daily return fluctuates greatly, with a maximum daily increase of 0.80% and a maximum daily decrease of -0.51%. Although the investment portfolio has shown some growth potential during certain periods, overall, its cumulative return has shown negative growth throughout the analysis period, ultimately recording a cumulative decline of -0.20%.

4. Conclusion

This study first systematically evaluated the fitting performance of the ARIMA model. By optimizing model parameters and utilizing stock price data from the first 70 trading days, the ARIMA model is used to predict stock prices on subsequent trading days. In addition, the rolling window prediction method is applied, with a 70-day window as the window, rolling forward for 30 days each time, and updating the window data during each rolling to ensure the adaptability of the model in dynamic market environments. The results show that the ARIMA model can effectively predict the future trend of asset prices to a certain extent.

On this basis, the mean variance optimization method and the maximum Sharpe ratio method are used to calculate the daily optimal investment portfolio weight, and based on this, the investment portfolio is constructed to calculate the cumulative return of the portfolio assets. This optimization method aims to achieve optimal portfolio allocation by balancing risk and return.

Finally, compare and analyze the cumulative returns of the portfolio assets with the cumulative returns of the SP 500 Index (SP500) during the same period. The cumulative returns of the SP 500 index have been significantly higher than the Maximum Sharpe Ratio Model portfolio since mid-May and has maintained this advantage throughout the entire observation period. This phenomenon reflect that the overall market performance of SP 500 index components is better than that of investment portfolios, especially during periods of high market volatility. This difference may be related to various factors such as asset allocation, market risk exposure, and management strategies in the investment portfolio. The cumulative daily return of the final SP500 index is 7.25%, which is better than the maximum Sharpe ratio investment portfolio of -0.20% This analysis provides valuable reference for the formulation of investment strategies.

Overall, this approach represents a compelling solution for portfolio optimization, as it leverages advanced techniques to effectively manage risk and generate returns. However, this study does not fully consider realistic constraints such as transactional fees. It is also salubrious to investigate whether the strategy remains valid in a wider time horizon and with a higher trading frequency. In this sense, the proposed model can be further refined and expanded in future studies to achieve even better results.

References

[1] Kalayci, C. B., Ertenlice, O., & Akbay, M. A. (2019). A comprehensive review of deterministic models and applications for mean-variance portfolio optimization. Expert Systems with Applications, 125, 345–368.

[2] Chaweewanchon, A., & Chaysiri, R. (2022). Markowitz Mean-Variance Portfolio Optimization with Predictive Stock Selection Using Machine Learning. International Journal of Financial Studies, 10(3), 64.

[3] Jensen, M. C. (1978). Some anomalous evidence regarding market efficiency. Journal of Financial Economics, 6(2), 95–101.

[4] Basak, S., Kar, S., Saha, S., Khaidem, L., & Dey, S. R. (2019). Predicting the direction of stock market prices using tree-based classifiers. The North American Journal of Economics and Finance, 47, 552–567.

[5] Box, G. E. P., & Jenkins, G. M. (1976). Time Series Analysis: Forecasting and Control. Holden-Day.

[6] Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (1982). Forecasting: methods and applications. Wiley.

[7] Khashei, M., & Bijari, M. (2011). A new hybrid methodology for forecasting using autoregressive integrated moving average and artificial neural networks. Applied Soft Computing, 11(2), 2664-2675.

[8] Wang, J., Ding, J., & Zhang, Y. (2012). An improved particle swarm optimization algorithm for solving optimization problems. Applied Mathematics and Computation, 219(4), 2239-2259.

[9] Shen, Y., Li, X., & Huang, Y. (2012). Stock market forecasting using ARIMA model. Journal of Business Research, 65(10), 1425-1432.

[10] Zhang, X., Zhang, Y., & Li, H. (2017). Climate change analysis using ARIMA model. Journal of Climate, 30(5), 1745-1756.

[11] Liu, Y., Liu, Z., & Zhang, W. (2018). Distributed ARIMA models for big data analysis. Journal of Statistical Computation and Simulation, 88(12), 2401-2413.