Quantitative Research on Stock Multi-Factor Models in Empirical Investment

Chengzhao Zhang[1], Xun Huang[2*], Hairong Li[2], Jingxin Zhao[2]

[1]Chengdu Polytechnic, Innovation and Practice Base for Postdoctors, No. 83 Tianyi Street, North Section of Yizhou Avenue, High tech Zone, Chengdu, China
[2]Chengdu University, Business School, 2025 Chengluo Avenue, Chengdu, China

[a]zhangchengzhao@cdp.edu.cn,
[*]Corresponding Author: huangxun1118@126.com, [b]210843296@qq.com,
[c]LZ190404@163.com

Abstract. This study aims to empirically examine the efficacy of a multi-factor model in quantitative stock investment. The model integrates various factors such as value, size, momentum, and volatility, which are well-established in academic research and have been shown to significantly influence stock returns. Utilizing a comprehensive dataset of stock market information, we establish factor portfolios based on these factors and evaluate their performance over a specific time frame. The empirical findings indicate that the multi-factor model surpasses conventional single-factor models in terms of risk-adjusted returns. Specifically, the value and momentum factors demonstrate notable positive alphas, signifying their capacity to generate excess returns beyond market risk considerations. Additionally, the size and volatility factors also exhibit substantial impacts on stock returns, further underscoring the efficacy of the multi-factor model in capturing stock price movements. During a specific time period, this study focused on the constituents of the CSI300 Index (China Securities Index 300) and selected 12 factors, including industry, technical, and financial factors, to establish an initial factor pool. Subsequently, a multi-factor stock selection model based on neural networks was developed and backtested. The empirical results demonstrated that this investment strategy yielded higher excess returns compared to the benchmark index.

Keywords: multi-factor; quantitative stock investment; volatility; portfolio; CSI300

1 Introduction

Quantitative stock investment, also known as quantitative finance or algorithmic trading, is a method of investing that relies on mathematical and statistical models to make trading decisions. This approach involves the use of computer-based algorithms to analyze large sets of financial data, execute trades and identify patterns. Quantitative in-

ICPDI 2024, September 06-08, Jinan, People's Republic of China
Copyright © 2024 EAI
DOI 10.4108/eai.6-9-2024.2353667

vestment strategies aim to systematically exploit market inefficiencies and generate alpha, or excess returns, by leveraging data-driven insights and quantitative techniques. These strategies often encompass a wide range of approaches, including factor investing, statistical arbitrage, high-frequency trading, and machine learning-based models.

Factors such as value, quality, momentum, size, and volatility are commonly used in quantitative stock investment to identify stocks with potential for outperformance. By analyzing these factors and their historical relationships with stock returns, quantitative investors seek to construct optimized portfolios and systematically manage risk. The use of quantitative techniques allows investors to process vast amounts of data efficiently and objectively, enabling them to uncover insights that may not be readily apparent through traditional fundamental analysis or qualitative methods. Moreover, quantitative models can be designed to execute trades with precision and speed, taking advantage of short-term market inefficiencies or exploiting opportunities across various asset classes.

In recent years, advancements in technology and data analytics have further propelled the growth of quantitative stock investment, leading to the development of sophisticated algorithmic trading platforms and quantitative investment funds. These tools and strategies have attracted interest from institutional investors, hedge funds, and wealth managers seeking to enhance portfolio performance and risk management through systematic and data-driven approaches.

Multi-factor stock model investment is a quantitative approach that integrates multiple factors to systematically select stocks with the potential for superior returns. This investment strategy goes beyond traditional single-factor models by incorporating various fundamental, technical, and macroeconomic factors to identify stocks that exhibit desirable characteristics and performance. Key factors commonly considered in multi-factor stock models include value (e.g., price-to-earnings ratio), size (market capitalization), momentum (price trends), quality (profitability and stability), and volatility (risk measures). By combining these factors, investors aim to construct diversified portfolios that capture different dimensions of stock returns and manage risk effectively. The multi-factor stock model leverages empirical evidence and academic research to identify factors that have historically influenced stock performance. This data-driven approach allows investors to systematically evaluate and select stocks based on a comprehensive set of criteria, aiming to enhance portfolio returns while mitigating exposure to specific risks.

2 Literature Review

Li et al.[1] employ XGBoost (eXtreme Gradient Boosting) and random forest for rolling tests on multiple factors. They utilize the multi-factor selection model for a rolling backtest on CSI300 Constituent Stocks. Pan et al.[2] introduce the Intelligent Portfolio Theory for stock market investment. Test results confirm the Intelligent Portfolio Theory with the suggested multi-factor models. Xing et al.[3] introduce a novel GARCH (Generalized autoregressive conditional heteroskedasticity) model by incorporating a nonlinear potential function. The empirical results show that the proposed GARCH

model outperforms the benchmark GARCH in predicting prices during financial price crashes period. Guidolin et al.[4] investigate the empirical performance for estimating multi-factor models. Predictive log-likelihood scores indicate the need in both risk exposures and variances. Ballarin et al.[5] present a new framework, MFESEN (the multi-frequency echo state network), which is built on a relatively novel machine learning paradigm known as reservoir computing. Skočir et al.[6] explore the impact of multi-factor asset pricing models. Their analysis suggests that the multi-factor model results in decreased estimation errors. Guerard Jr. et al.[7] provide a proposed model by utilizing the data sources from 1997 to 2011. Their research also supports the use of APT (Arbitrage Pricing Theory) and multi-factor models for portfolio construction.

Grilli et al.[8] investigate the effectiveness for improving forecasting accuracy. The results indicate that Boltzmann's entropy is a valuable indicator suitable for application to individual stocks. Liu et al.[9] integrate environmental data with financial indicators for stock price forecasting by using a novel PCA-GRU-LSTM (principal component analysis-gated recurrent units-long short term memory) model. Yilmaz et al.[10] assess the capabilities of advanced deep learning models in multi-horizon forecasting for stock returns. The empirical findings affirm that the TCN (temporal convolution networks) model exhibits the strongest out-of-sample predictive performance across all forecasting horizons when compared to other models, based on analysis of three financial datasets. Aydinhan et al.[11] extend the jump model and the empirical tests demonstrate that the approach surpasses the traditional models. Salles et al.[12] introduce MSED (Multi-Scale Event Detect), a method for identifying events in financial time series. The findings revealed a correlation between the uncertainty fluctuations in the EPU (Economic Policy Uncertainty) and the events detected in a financial time series. Zhang et al.[13] apply the modified iterative cumulative sum of squares (ICSS) algorithm and the EEMD (Ensemble empirical mode decomposition) method to decompose the index returns. Pokou et al.[14] seek to effectively underscore the relevance of hybrid models for economic or financial agents. The empirical results demonstrate the effectiveness of the hybrid models under consideration. Hafiz et al.[15] have created predictive models based on neural networks to tackle this challenge; however, the selection of an appropriate neural architecture is rarely discussed. The study's findings compellingly illustrate that the proposed approach can produce relatively efficient and concise networks.

Other chapters can be summarized as follows. Section 3 provides summary of multi-factor model. Moreover, Section 4 presents the construction and testing of the multi-factor model. Section 5 conducts empirical analysis. Finally, Section 6 draws the conclusions and prospects for future research.

3 Summary of Multi-factor Model

In 1952, Markowitz[16] introduced a basic model for determining optimal investment portfolios by using mean and variance to quantitatively describe the return and risk of assets. Subsequently, Sharpe[17] and others developed the CAPM (Capital Asset Pricing Model) based on this theory, which suggests that the expected excess return on an asset is proportional to the market's excess return. The CAPM contends that a stock's price

is solely tied to market risk and not connected to the fundamentals of listed companies, with high stock prices requiring high beta values to support them. However, the assumptions of the CAPM model were considered too restrictive, leading subsequent scholars to break away from these original assumptions and derive the APT model. This model features looser assumptions than the CAPM model but does not determine which factors play a decisive role.

After scholars found that many phenomena could no longer be explained by the CAPM model, such as the significant outperformance of small-cap and value stocks in the market, Fama and French introduced new factors to explain asset returns based on the CAPM model[18]. The three-factor model indicates that returns are related not only to market risk but also to the returns of simulated portfolios based on the book-to-market ratio and size factors[19]. Subsequently, researchers found that the momentum phenomenon in the market could not be explained by the three-factor model. Therefore, they added the momentum effect on top of the existing three factors, creating the four-factor model, which takes into account the tendency for stocks that have been continuously rising or falling to continue in the same direction[20]. The previous models did not include factors characterizing a company's asset quality to explain asset returns. Therefore, factors representing profitability and investment patterns were added, leading to the proposal of the five-factor model[21].

The multi-factor model is a statistical model used to explain asset returns by considering the impact of multiple factors on asset price movements to more accurately estimate the expected returns of assets. These factors typically include market factors, risk factors, industry factors, and style factors. The fundamental idea behind the multi-factor model is that asset returns are influenced not only by the overall performance of the market but also by other factors. By introducing multiple factors, a more comprehensive analysis of asset return volatility and characteristics can be achieved, thereby enhancing risk control and return forecasting for investment portfolios. When constructing the multi-factor model, statistical tools such as regression analysis can be used to identify and quantify the impact of each factor on asset returns, enabling asset pricing and risk assessment based on this information. The multi-factor model is widely used in asset allocation, portfolio construction, and risk management, providing investors with more effective decision support.

4 Construction and Testing of the Multi-factor Model

Outliers can significantly impact SVM performance due to its sensitivity to noise and outliers. Employ robust estimators to identify and manage outliers, or use robust optimization models to mitigate their effects. Missing values can be addressed by various methods, such as imputing with mean, median, or using more complex approaches like KNN (K-Nearest Neighbors) to estimate missing values. Choosing the right imputation method is crucial for enhancing the model's robustness.

4.1 Data Preprocessing before Model Construction

The selection of raw data forms the basis of the multi-factor model. Only high-quality data can ensure the accuracy of subsequent research. The quality of data preprocessing largely determines the reliability of the model analysis results. Therefore, it is necessary to preprocess the raw data before constructing the model.

In practical use, the appropriate method should be selected according to the specific situation. In statistics, due to the extremely low probability of values occurring outside of plus or minus three standard deviations under normal distribution, they can be considered as rare events. Rare events are almost impossible to occur, so these values can be regarded as outliers. This method is practical and can identify significant outliers. However, the drawbacks are also evident. Firstly, calculating the mean and variance based on outlier data can lead to inaccurate mean values and may potentially inflate the variance, resulting in excessively wide upper and lower bounds, making it difficult to filter out certain outliers. Another drawback is that this method is not applicable to non-normally distributed datasets.

4.2 Missing Value Treatment

A factor dataset with a high number of missing values is generally considered to be of poor quality, mainly because missing values lead to the loss of useful information and increase the system's uncertainty. Therefore, adopting a reasonable approach to handling missing values can improve the quality of factor data and lay a solid foundation for subsequent steps.

The exclusion method involves removing elements with missing data to obtain a complete information table, which is the simplest method of handling missing data. This approach is suitable for factors with a large number of missing values or stock objects with a large number of missing values, provided that these factors or stock objects constitute a very small proportion of the total data. Otherwise, discarding this information could affect the accuracy of the results.

The imputation method involves filling in missing values with appropriate values, and finding suitable values is the key to determining the quality of the imputation. There are various methods for filling in missing values, such as manual entry, special value imputation, mean imputation, etc., with mean imputation being more common. Different types of companies and different ratios have different factor averages, hence the imputation methods also vary.

4.3 Model Factor Selection

Based on macroeconomic, industry, company fundamentals, and market characteristics, combined with various specific factors to construct an investment portfolio, the factors that affect stock price returns can be roughly divided into several categories. These include market-wide factor, valuation factor, growth factor, profitability factor, momentum and reversal factor and size factor.

Generally, the sorting method is commonly used to test the effectiveness of candidate factors. For example, monthly testing involves calculating the size of the factor for each asset in the market at the beginning of the first month, sorting the assets in ascending order according to the factor, and dividing all assets into N portfolios. These portfolios are held until the end of the month. At the beginning of each month, the same method is used to reconstruct the N portfolios and hold them until the end of the month, repeating this process until the end of the period.

If there is correlation between factors, using only one factor to construct a long-short portfolio cannot eliminate the influence of other factors, so the results obtained are not entirely the final performance of that factor. Additionally, when constructing a long-short portfolio, only the stocks at the top and bottom ends of the ranking are generally used, and the information of stocks in the middle is not fully utilized, leading to a waste of information.

Constructing a long-short investment portfolio based on factor values serves as the factor's return rate. Regression is performed on all portfolios within each time period, and the returns are regressed against the risk premium levels of the risk factors. The average return rate ER_i of a single portfolio over the entire time period is calculated, and ER_i is used to regress the estimated β value obtained from different stocks in different time periods in the previous stage. After the single-factor test, all factors that are indeed effective are included in the model to build a regression equation, forming a multi-factor model.

Based on the summary of previous research, this paper selects 12 indicators from three major categories：growth, valuation, profitability as candidate factors for the model, drawing on the research conclusions of scholars at home and abroad on multi-factor quantitative stock selection models, as shown in Table 1.

Table 1. Candidate factors.

category	factor name	factor explanation
valuation factor	Price-Earnings Ratio	The current total market value of the stock / the net profit of the company in the last four quarters
	Price to Book Ratio	The current total market value of the stock / the current net asset value
	Price to Cash Flow Ratio	The current total market value of the stock / the net operating cash flow for the last four quarters
	Price to Sales Ratio	The current total market value of the stock / the sales revenue for the last four quarters
	Net Profit Growth Rate	The net profit of the company for the current quarter / the net profit for the same quarter last year -1

growth factor	Earnings Per Share Growth Rate	(The growth rate of earnings per share for the current period - The growth rate of earnings per share for the same period last year)/(The growth rate of earnings per share for the same period last year)
	Quarterly Change in ROA	(current period ROA - previous period ROA)/ The absolute value of the ROA for the previous period
	Quarterly Change in ROE	(current period ROE - previous period ROE)/ The absolute value of the ROE for the previous period
profitability factor	Return on Equity	The net profit attributable to the parent company for the last four quarters / the average net assets
	Return on Assets	The net profit attributable to the parent company for the last four quarters / the average total assets.
	Return on Invested Capital	ROIC= Earnings Before Interest and Tax(EBIT)× (1- tax rate)/ invested capital
	Total asset turnover ratio	Net operating revenue / the average total assets

5 Empirical Analysis

This paper extracts feature indicators from research reports for machine learning to study the multi-factor quantitative model, which helps to retain the broad investment perspective, strong investment discipline, and high utilization rate of historical data of traditional multi-factor models. At the same time, it transforms the stock selection prediction capabilities of securities research institutions into actual investment value, broadens the analysis methods of multi-factor stock selection strategies, and enriches the theoretical research related to multi-factor models in China. Based on the selected effective factor combinations, a multi-factor stock selection model is established using support vector machine method. Stocks with top 10% returns in the A-share sample are selected as the investment portfolio, with quarterly rebalancing. The model is tested and evaluated using the CSI 300 Index as benchmarks to test the stock selection ability of the model.

5.1 Model Construction

The paper uses the selected effective factor combination as the input feature set for the support vector machine multi-factor stock selection model, and divides stocks into three

categories based on their returns in descending order: the first category is the high-quality stock portfolio A (top 10%), the second category is the average stock portfolio B (middle 80%), and the third category is the low-quality stock portfolio C (bottom 10%). These portfolios are used as the output feature set of the model. The support vector machine model is used to classify all stocks, and the high-quality stock portfolio judged as Class A is selected. An equal-weight method is used to construct an investment portfolio strategy, and its performance is analyzed during the out-of-sample test period.

The paper utilizes the Libsvm toolbox of Matlab software to construct the multi-factor stock selection model. Previous studies have shown that the setting of the time length for the training and testing sets is a critical issue affecting the construction of the support vector machine classification model. Through analysis, the paper adopts a continuously extended dynamic training set and a fixed-length testing set, with the training set length increasing by 3 months each time, while the testing set remains fixed at 3 months. The specific steps are as follows:

(1) The paper selects the sample screening period from January 2019 to December 2023, totaling 60 months of data as the initial training set for the model, gradually increasing to 108 months. The forecast set is set as a fixed value of 3 months, and the out-of-sample testing period data are used as the model's testing set.

(2) After standardizing the effective factors, they are used as the feature input set for the model. Stocks are then sorted into three categories A, B, and C, based on their portfolio returns, to serve as the corresponding target variables.

(3) The model parameters are optimized using a method of network parameter optimization, with the cross-validation error serving as the objective function. The model is continuously trained to minimize the error, thus obtaining the optimal parameterization for the model.

(4) To test the stock selection ability of the model, we use the effective factors in the testing set to predict the categories of stocks from January 2020 to November 2021, and calculate the monthly return accuracy. We select the stocks classified as category A to form an investment portfolio, establish an equal-weight investment strategy, and then track its comprehensive performance relative to the CSI 300 index during the out-of-sample testing period.

5.2 Testing of Multi-factor Stock Selection Model

From the perspective of stability, during the period from January 2016 to November 2017, the returns of the investment portfolio strategy constructed by the support vector machine multi-factor stock selection model exceeded those of the CSI 300 index. Table 2 presents the statistical results of the support vector machine multi-factor stock selection model. The multi-factor model's quarterly average return rate was 1.21%, surpassing the return rate of the CSI 300 index. In terms of risk, the Sharpe ratio[22] comprehensively reflects an important indicator of model risk. Larger values indicate higher excess returns per unit of risk, leading to higher overall model returns[22]. As shown in Table 2, the Sharpe ratios[22] of the multi-factor stock selection model and the CSI 300 index are 0.7254 and 0.7028 respectively, while the model's Beta values are 0.4278 and

0.3782[22]. This suggests that the multi-factor stock selection model can achieve decent returns while also demonstrating good stability.

Table 2. Statistical Results of the Support Vector Machine Multi-Factor Stock Selection Model.

Indicators	multi-factor stock selection model	CSI 300 index
quarterly average return	1.21%	0.83%
Beta	0.4278	0.3782
Sharpe ratio	0.7254	0.7028
maximum drawdown	18.78%	28.56%

6 Conclusion and Outlook for Future Research

Utilizing the Fama-Macbeth regression, it has been observed that the turnover rate and abnormal turnover rate factors exhibit significant returns. From the commonly used indicator in the quantitative finance industry, ICIR (Information Coefficient), it can be seen that both turnover rate factors, adjusted for the impact of small market capitalization, can achieve robust excess returns. This indicates that in emerging markets like A-shares, factors constructed based on behavioral finance theory have a low correlation with fundamental factors and can effectively complement the traditional multi-factor model. Researching empirical asset pricing models from the perspective of behavioral finance still has great potential for development.

Studying multi-factor models not only supplements existing asset pricing models and provides quantitative investment advice to investors but also assists regulators in analyzing the relationship between capital market fluctuations and investors' irrational behaviors. This is of significant importance for building a multi-level capital market, improving investor protection policies, and achieving the sustainable development of the capital market.

This study demonstrates the efficacy of multi-factor models in quantitative stock investment, revealing significant predictive power for stock returns and identifying key factors that contribute to excess returns. The limitations, such as sample selection bias, are acknowledged, suggesting avenues for model refinement. Looking forward, the integration of emerging technologies like machine learning and further exploration into behavioral finance are poised to enhance model robustness and investment strategies, offering promising prospects for sustainable capital market development.

Acknowledgments

We express our gratitude for the assistance provided by the National Natural Science Foundation of China under grant No.72201042, the Youth Program of Sichuan Natural Science Foundation under grant No. 2023NSFSC1027, and Chengdu Polytechnic under grant No. 23KYTD11.

References

1. Zimo Li, Weijia Xu, Aihua Li. Research on multi factor stock selection model based on LightGBM and Bayesian Optimization [C]. 9th International Conference on Information Technology and Quantitative Management, 2022,214: 1234-1240.
2. Heping Pan, Manxiao Long. Intelligent Portfolio Theory and Application in Stock Investment with Multi-Factor Models and Trend Following Trading Strategies [C]. International Conference on Identification, Information and Knowledge in the internet of Things, 2021,187: 414-419.
3. Dun-Zhong Xing, Hai-Feng Li, Jiang-Cheng Li, Chao Long. Forecasting price of financial market crash via a new nonlinear potential GARCH model [J]. Physica A: Statistical Mechanics and its Applications, 2021,566: 1-16.
4. Massimo Guidolin, Francesco Ravazzolo, Andrea Donato Tortora. Alternative econometric implementations of multi-factor models of the U.S. financial markets [J]. The Quarterly Review of Economics and Finance, 2013,53: 87-111.
5. Giovanni Ballarin, Petros Dellaportas, Lyudmila Grigoryeva, Marcel Hirt, Sophie van Huellen, Juan-Pablo Ortega. Reservoir computing for macroeconomic forecasting with mixed-frequency data [J]. International Journal of Forecasting, 2024,40: 1206-1237.
6. Matevž Skočir, Igor Lončarski. On the importance of asset pricing factors in the relative valuation [J]. Research in International Business and Finance, 2024,70: 1-18.
7. John B. Guerard Jr., Harry Markowitz, GanLin Xu. Earnings forecasting in a global stock selection model and efficient portfolio construction and management, 2015,31: 550-560.
8. Luca Grilli, Domenico Santoro. Forecasting financial time series with Boltzmann entropy through neural networks [J]. Computational Management Science, 2022,19: 665-681.
9. Bingchun Liu, Mingzhao Lai. Advanced Machine Learning for Financial Markets: A PCA-GRU-LSTM Approach [J]. Journal of the Knowledge Economy, 2024,5: 1-35.
10. Firat Melih Yilmaz, Engin Yildiztepe. Statistical Evaluation of Deep Learning Models for Stock Return Forecasting [J]. Computational Economics, 2024,63: 221-244.
11. Afşar Onat Aydinhan, Petter N. Kolm, John M. Mulvey, Yizhan Shu. Identifying patterns in financial markets: extending the statistical jump model for regime identification [J]. Annals of Operations Research, 2024,4: 1-37.
12. Diego Silva de Salles, Cristiane Gea, Carlos E. Mello, Laura Assis, Rafaelli Coutinho, Eduardo Bezerra, Eduardo Ogasawara. Multi-Scale Event Detection in Financial Time Series [J]. Computational Economics, 2024,2: 1-29.
13. Yue-Jun Zhang, Han Zhang, Rangan Gupta. A new hybrid method with data-characteristic-driven analysis for artificial intelligence and robotics index return forecasting [J]. Financial Innovation, 2023,9: 1-23.
14. Frédy Pokou, Jules Sadefo Kamdem, François Benhmad. Hybridization of ARIMA with Learning Models for Forecasting of Stock Market Time Series [J]. Computational Economics, 2024,63: 1349-1399.
15. Faizal Hafiz, Jan Broekaert, Davide La Torre, Akshya Swain. A multi-criteria approach to evolve sparse neural architectures for stock market forecasting [J]. Annals of Operations Research, 2024,336: 1219-1263.
16. Markowitz, H.M. Portfolio selection. [J].Journal of Finance, 1952,7: 77-91.
17. William F. Sharpe. Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk [J].Journal of Finance, 1964,19: 425-442.
18. Fama, E.F., French, K.R. The CAPM is wanted, dead or alive [J].Journal of Finance, 1996, 51: 1947–1958.

19. Fama, E. F., French, K. R. Common risk factors in the returns on stocks and bonds [J].Journal of Financial Economics, 1993,33(1): 3-56.
20. Carhart, M. On persistence in mutual fund performance [J].The Journal of Finance, 1997,52(1): 57-82.
21. Fama, E.F., French, K.R. A five-factor asset pricing model [J].Journal of Financial Economics, 2015,116(1): 1-22.
22. William F. Sharpe. Mutual Fund Performance [J].Journal of Business, 1966,39: 119-138.