A Study on Machine Learning-Enhanced Efficient Frontiers in Multifactor Models

Zhongting Wang
School of Economics and Management, Harbin University of Science and Technology, Harbin, Heilongjiang Province, 150080, China
18945454188@139.com

ABSTRACT

Traditional multi-factor models and Markowitz's efficient frontier theory struggle to adapt to complex financial markets, making machine learning a crucial tool for expanding the efficient frontier of investment portfolios. This paper systematically reviews the research landscape in this field, delineates its developmental stages, and summarizes the current applications and pros and cons of various machine learning algorithms across four dimensions: factor discovery, risk measurement, portfolio optimization, and constraint refinement. The study finds that the current field suffers from issues such as a lack of underlying theory, a fragmented research framework, the presence of pseudo-factors, and overfitting. Based on these findings, future research directions are projected, and it is proposed that the theoretical framework should be strengthened, causal and robust techniques should be integrated, and localized research should be conducted, with the aim of providing guidance for both theoretical exploration and practical applications in this field.

Keywords: Multifactor Models; Efficient Frontier; Machine Learning

1. INTRODUCTION

In 1952, Markowitz proposed the mean-variance portfolio theory, defined the efficient frontier of investment portfolios, and laid the foundation for modern portfolio optimization theory [1]. Since then, Capital Asset Pricing Model (CAPM) and Arbitrage Pricing Theory (APT) have emerged. In addition, Fama-French's series of multi-factor models modified the asset price mechanism in more detail [2]. The combination of multi-factor model and mean variance optimization has become the mainstream choice of global institutional investors.

With the advent of the era of big data, the working methods of financial markets have undergone great changes. The market state often changes, and the effect of factors has the characteristics of obvious change with time. Alternative data sources such as public opinion and industrial chain are constantly spreading, and the amount of data is growing. At the same time, the strategies of market participants are similar, the factors are crowded, and the benefits of traditional factors have been declining. The traditional linear static model is analyzed according to fixed rules, which is not only difficult to capture the complex nonlinear relationship between factors, but also difficult to adapt to market changes dynamically. In the end, it is difficult for a portfolio to stay at the forefront of optimal efficiency. However, the development of machine learning technology has given us a new way to solve these problems. Academic circles have recognized the practical value of machine learning algorithm in the fields of income prediction, risk management, portfolio adjustment and so on, and there are a lot of empirical evidence [3]. However, after investigating the research papers, it is found that the current research is still very fragmented. Dynamics, causality, robustness and other important constraints are often studied separately, and the basic theoretical research lags far behind the practical application. Therefore, the basic logic and applicable conditions of using machine learning to improve the efficiency frontier have not been solved so far.

Based on this research situation, the paper makes an overall survey of the development path in this field, classifies and summarizes the use scenarios, advantages and disadvantages of different algorithm, finds out the core problems existing at present, and points out the direction for future research. The purpose of this paper is to clarify the research methods and help the cooperative development of theory and application in this field.

2. EVOLUTION OF FUNDAMENTAL THEORIES AND APPLICATION BOTTLENECKS OF TRADITIONAL MODELS

Efficient frontier theory and multi-factor model are two important basic theories in this field. Two people helped each other and worked out the framework of the previous portfolio optimization. In this part, we will track how these theories change, and analyze in detail the real reason why the old framework can't continue to expand the efficient frontier according to the current market situation.

2.1 Evolution and Refinement of Core Theories

2.1.1 The Development of Efficient Frontier Theory

The mean-variance model made by Markowitz (1952) aims at increasing the return of the portfolio and reducing the variance of the portfolio, so as to find the best asset ratio. The form of this formula laid the foundation for the theory of efficient frontier. However, this model has two shortcomings. First, when there are many types of asset, the workload of calculating covariance matrix will increase exponentially, thus reducing the calculation efficiency. Second, the model is very sensitive to input parameters such as return and volatility. Even if there is a slight estimation error, the portfolio's asset ratio will be seriously deviated, resulting in poor out-of-sample stability. In order to correct these shortcomings, the following research focuses on three main improvement points. The first is covariance matrix optimization, which uses sample shrinkage estimation, Bayesian estimation, principal component analysis and other methods to reduce data noise. The second is to add short selling restrictions, trading times restrictions, sector neutrality rules and other trading restrictions to meet the actual trading conditions. Thirdly, by extending robust optimization, the robust mean-variance model and risk parity model are created to reduce the negative impact caused by unclear parameters [5].

All traditional improvement methods are based on two important premises: static and linear models. They believe that the market structure, asset correlations and return patterns have been stable for a long time. Therefore, the traditional model cannot adapt to the fast-moving modern financial markets. 2.1.2 Iterative Improvement of the Multi-Factor Model Framework

Multi-factor models serve as the key bridge connecting asset pricing and portfolio optimization. Their core expression is:

$$R_{i,t} = \alpha_i + \sum_{j=1}^{K} \beta_{i,j} F_{j,t} + \varepsilon_{i,t}, \tag{1}$$

where Ri,t is the excess return of asset i at time t, αiis the asset-specific return, βi,jis the factor exposure coefficient, Fj,t is the common factor return, and εi,t is the random disturbance term.

The development of multi-factor models has evolved from single-factor to multi-factor approaches and from single-style to multi-style frameworks: starting with the CAPM's single-market-factor model, progressing to the Fama-French three-factor and five-factor models, and further expanding with the addition of specialized factors such as sector, volatility, liquidity, and sentiment factors, thereby forming a vast traditional manual factor library. Multi-factor models effectively simplify the computational logic of portfolio optimization and, during periods of stable market styles and low volatility, can support the maintenance of the efficient frontier at an optimal level [6].

2.2 Phases of Research in the Field

Combined with the progress of technology and the change of research hotspots, this paper divides the field of multi-factor models and efficient frontier into three development stages, which clearly shows the logic of research:

1. Traditional Statistical Optimization Phase (1952-2010): This phase is centered on the classical mean-variance model and linear multi-factor models, and the focus of traditional statistical research is to reduce estimation error and better formulate trading rules, which is a pure theoretical and traditional empirical research stage.

2. Machine Learning Empirical Application Stage (2010–2020): Machine Learning began to enter the field of quantitative finance. The research focuses on comparing the revenue forecasting ability of various algorithms to confirm whether machine learning can improve the efficient frontier [7]. The biggest feature of this stage is that it pays more attention to empirical research than theory.

3. Constraint Deepening and Theoretical Exploration Phase (2020–present): Students are beginning to realize the shortcomings of the model that works only by data. Causal inference, adversarial robust learning and explainable AI have become the main research topics. At the same time, some scholars began to study how machine learning affects the basic theory of efficient production line, and this field gradually changed into an experimental and theoretical mode.

2.3 Inherent Limitations of Traditional Model Frameworks

In today's rapidly changing and complex market, the shortcomings of the static linear research framework are gradually becoming clear. This is the fundamental reason why the efficient frontier of the extended portfolio meets the developmental bottleneck. First of all, it is difficult for traditional models centered on linear regression to grasp the complex interaction between many factor and the nonlinear relationship between asset returns. This will lose a lot of new market information, and the scope of expanding the efficient frontier will become smaller. Secondly, the determined factor weighting and risk parameters can't keep up with the frequently changing market conditions, such as ups and downs, sharp or stable price fluctuations. Therefore, the investment portfolio will always be separated from the best efficient frontier. Third, the traditional factors only relies on statistical correlation for selection, and it is easy to have similar factors, and everyone uses the same factor crowding, resulting in a decline in revenue and becoming an efficient front-end. This will produce many pseudo-factors without economic reasons, and the performance when using the actual data of the model will drop significantly [8]. Finally, for the increasing number of factor framework and alternative data, the linear model is more prone to multicollinearity problems. This will greatly reduce the efficiency of data processing and calculation, and it is difficult to meet the requirements of quantitative research in the era of big data. Generally speaking, the traditional statistical methods can't overcome the obstacles of today's development, which is the central reason for integrating machine learning into this research field and improving efficient frontier.

3. APPLICATION FRAMEWORK AND RESEARCH PROGRESS OF MACHINE LEARNING IN EMPOWERING THE EFFICIENT FRONTIER

According to the usual workflow of Quantitative investment, that is to say, according to all the processes including factor discovery, risk measurement, portfolio construction and risk constraints, this paper categorizes the application of machine learning into four major modules. It systematically reviews the current state of research, algorithmic characteristics, academic perspectives, and application outcomes for each module, while comparing the advantages and disadvantages of various technical methods.

3.1 Factor Discovery and Return Forecasting: Return-Side Optimization

This module represents the most researched and widely used field in industry. The primary goal of this model is to find nonlinear factor by using machine learning to improve the accuracy of return prediction. In this way, an efficient boundary can be extruded outward from the side of return. According to the nature of the algorithm, it is divided into three categories, and there are obvious differences in the use methods between different algorithms.

3.1.1 Regularized Linear Models

These models represented by LASSO, Ridge and Elastic Net are slight extensions of linear models. Their core function is to automatically select features in high-dimensional feature space, thus weakening the multicollinearity problem. These models retain the advantages of "easy explanation" and "fast calculation" of linear models, so they are suitable for scenarios with high-frequency trading and low delay strategy. It is widely believed at home and abroad that the regularization model can only make the factor structure slightly better. Because of its linear nature, it is impossible to capture the relationship between factor, and its ability to improve the efficiency frontier is relatively limited.

3.1.2 Tree-Based Ensemble Learning Models

Tree-based models such as Random Forest, XGBoost and LightGBM, are the mainstream configurations in the field of quantitative factor. As a nonlinear and nonparametric model, the tree-based perceptual learning model can automatically capture the interaction of factors, and has a good adaptability to price, turnover and basic data. Previous studies at home and abroad show that the gradient tree model is better than the traditional linear factor model in both the accuracy of prediction and the performance of portfolio. However, in academic circles, tree model is widely considered to have the risk of overfitting, and its explanatory power is not enough. When the market structure changes, the generalization ability of the model will drop sharply.

3.1.3 Deep Learning Models

Deep learning algorithms such as LSTM and CNN are very good at dealing with alternative data with incomplete structure such as time series data and text, and their ability to find information is very strong. This can promote the efficient frontier of portfolio [9]. But these models cost you money to learn, and it is difficult to know what you are doing, and it is difficult to stabilize when the market is very chaotic. Therefore, people in the industry often use it in combination with traditional models. Generally speaking, in the factor discovery stage, the tree model is the most suitable one in the A-share market, which is supplemented by regularized models, while Deep learning focuses on the special exploration of alternative data, and these three methods are put together to form a ladder-like application framework.

3.2 Covariance Matrix Optimization: Risk-Side Optimization

Efficient frontier is determined by interests and risks. Moreover, covariance matrix is the center to measure the risk of portfolio. The traditional covariance matrix is easily influenced by noise, delay effect and dimension problems, which may distort efficiency. There are two main methods of machine learning to measure risk well. The first is to use algorithm (such as principal component analysis or clustering) to reduce noise or dimension to improve the stability of the portfolio. The second is to use hidden Markov models and LSTMs to create dynamic covariance matrix to adapt to the changing market situation. This optimization does not change the framework of the original factor, but only needs to correct the error of risk calculation, which is an important auxiliary method to polish the efficient frontier.

3.3 Dynamic Portfolio Weight Optimization: Direct Optimization at the Portfolio Level

This method is different from the combination of factor forecasting and traditional mean-variance optimization. Use machine learning to determine the weight in one breath and directly make the efficient frontier. Commonly used algorithms are reinforcement learning and intelligent optimization algorithms. Deep reinforcement learning transforms portfolio construction into Markov decision process, and automatically adjusts shareholding according to profit, decline, Sharpe ratio and other indicators. This works very well in the market with changeable styles. Genetic algorithms and particle swarm optimization solve limited optimization problems to improve the accuracy of the answers [10]. This practice has a good place in theory, but it is difficult to train and know the reason why it is moving. In addition, limited by rules and risk management, there are actually not many scenarios that can be used.

3.4 Causal and Robust Constrained Optimization: Advanced Improvements in Risk Management

With people gradually discovering that the model only runs on data, in recent three years, causal inference, adversarial robust learning and explainable AI have become the latest research hotspots. These important functions are to make rules for machine learning model and keep the effective boundary line stable all the time, and they are also important bridges to the original research introduced in this paper.

3.4.1 Applications of Causal Inference

A major challenge for financial markets is that correlation does not mean causation. Purely data-driven machine learning is prone to identifying pseudo-factors that are statistically correlated but lack economic logic, thereby exacerbating factor crowding and leading to strategy failure. Causal inference uses methods like causal graphs, causal forests and confounding path blocking. To distinguish the real causal relationship between factors and returns, so as to eliminate pseudo-factors at the source. At present, some researches at home and abroad have added causal constraints to the process of factor selection, confirming that causal factors demonstrate significantly greater persistence than conventional data-driven factors. However, the existing research has clear limitations: most of the research only stays in the stage of factor selection, and fail to integrate causal theory with efficient boundary conditions and portfolio optimization objectives.

3.4.2 Applications of Adversarial Robust Learning

The signal-to-noise ratio of Financial data is very low. Random noise, extreme market conditions, and changes in sample distribution will all cause large fluctuations in model output, thus causing large changes in efficiency. “Adversarial robust learning” uses “adversarial training” and “regularization” constraints to represent the allowable range of data disturbance in digital form, thus enhancing the anti-interference ability of the model. The research so far has confirmed the advantages of robust model in the case of black swan events. However, most of the research is carried out separately, and there is no standard framework for making robust efficient frontier.

3.4.3 Application of Explainability Techniques

Explainability tools such as SHAP and LIME can explain the logic behind the relationship between elements in the nonlinear model and establish the connection between machine learning elements and traditional economic elements. “At present, these technologies can only explain part of it, and it is still a difficult problem to figure out the whole process of the “black box” problem.

4. CORE GAPS AND ISSUES IN CURRENT RESEARCH

Based on the previous analysis of application framework, this section will focus on specific areas of multifactor models, machine learning and efficient frontiers. Beyond the general topic, this paper summarizes the five key points in the research so far in order, and finds out the important research gaps that have not been solved.

4.1 Lack of Underlying Theoretical Research

Most of the research is only to verify the phenomenon that machine learning can expand efficiency, but it does not answer the core questions of asset pricing and portfolio optimization theory. What kind of factor interaction intensity, market integration and model complexity can machine learning stably expand the efficient frontier, and how machine learning can carry out this idle part in the theoretical framework is the most concerned point of this study.

4.2 Model Overfitting Causes Out-of-Sample Contraction of the Efficient Frontier

For data with various elements with high dimensions, a complex machine learning model can easily become overfitting. Although these models can match the best efficient frontier of all the examples collected so far, when the new market or market structure changes, the portfolio will immediately fall out of the best range, thus improving the efficient frontier is particularly obvious in deep learning models, and it has always been a problem in practical use.

4.3 Fragmented Research Framework

Now the research on various modules is very loose. Factor discovery, risk optimization, Weight Adjustment and causal constraints are in different fields. If these modules are optimized separately, errors will continue to increase. So far, there is no research that unifies three important factors, namely, dynamism, causality and robustness, in a theoretical framework. Therefore, it is impossible to maximize the expansion of the efficient frontier and keep it stable for a long time [12]. This is the most important new opportunity in this field.

4.4 The Issues of Spurious Factors and Factor Crowding Remain Unresolved

Even in the emergence of machine learning today, if we only rely on statistical relations to select factors, the problem that any factor is similar is also very serious. Although causal inference is an effective solution, the work of causal modeling is very complicated and requires a lot of energy to calculate. At present, it is still in the academic experimental stage and has not been widely used in industry. Therefore, the risk of factor return reduction or strategic failure still exists.

4.5 Insufficient Depth in Localized Research

With the support of mature capital market, foreign research has established a relatively perfect empirical framework. Most of the domestic research directly uses foreign algorithms and models, and there are few original theoretical studies on the characteristics of A-share market, such as large fluctuation, fast action and many individual investors. In order to cater to the domestic market and improve machine learning-enhanced efficient frontiers, there is almost no research on specific rules and parameter thresholds.

5. FUTURE RESEARCH TRENDS AND DEVELOPMENT DIRECTIONS

Considering the insufficient research mentioned above, this part puts forward a targeted forecast for the future development direction. Each trend corresponds to the questions raised above, which is completely consistent with the original research written on this paper.

5.1 Filling Theoretical Gaps to Establish a Theoretical Framework for Machine Learning-Enhanced Efficient Frontiers

In order to solve the shortcomings of this theory, the core direction in the future is to put matrix analysis, asset pricing theory and portfolio optimization theory together. Then, the available conditions of machine learning-enhanced efficient frontiers, factor interaction thresholds, and model complexity matching rules are given, and the theoretical frontiers under different market conditions is given. According to this, the field will change from the state of only experiment to the state of both theory and experiment. This is also the central research goal of this paper.

5.2 Developing Dynamic Adaptive Hybrid Models to Balance Fitting Capability and Generalization Ability

Only one algorithm can hardly adapt to all market environments. The hybrid adaptive model combining linear model, tree-based model and powerful function of deep learning will become the mainstream. These models use market status identification to automatically adjust their structure, factor weight and risk parameter according to market conditions. Always cooperate with the dynamic efficient frontier working in various situations, and technically reduce the problem of overfitting.

5.3 Promoting the Deep Integration of Causal, Robust, and Portfolio Optimization Techniques

On the other hand, we simplify the causal discovery algorithm and reduce the cost of calculation. Causal constraints are added to the whole process of factor discovery and portfolio optimization, which fundamentally inhibits pseudo-factors and factor crowding. On the other hand, we have perfected the theory of adversarial-robust efficient frontiers, quantify the tolerance range for disturbances under extreme market conditions by numbers, formulated standardized robust constraint rules, and ensure the long-term stability of the efficient frontier. The “dynamic-causal-robust” unified framework constructed in this paper aligns precisely with this development trend.

5.4 Deepening Research on Interpretability to Facilitate Model Compliance and Implementation

We will also make the extended explainability tools better in the future. Create a whole mapping system connecting machine learning elements and traditional economic elements. Then the model logic is easy to understand from beginning to end. Solve the “black box” problem of nonlinear model, and eliminate the obstacle that the asset management institutions of China driver’s license follows the rules when using complex machine learning model.

5.5 Conducting Localized Original Research Rooted in the A-Share Market

We will cooperate with the structure of domestic capital market, and make different investigations on the selection method, the model parameters and the extent to which they can be used of the factor that conform to A-share market. Our goal is not only to use foreign research results, but to establish a quantitative theory and mechanism that can use China market with obvious characteristics.

5.6 Leveraging Large Language Models and Multimodal Data to Uncover Incremental Insights

Language model and multimodal learning are the driving forces for the development of new technologies. By putting unstructured data such as text, research reports, public sentiment and images together, we have created new incremental elements and further expanded the scope of the efficiency frontier.

6. CONCLUSION

Looking back on the whole development process, multi-factor models and efficient frontier theory have been fully developed. From static linear modeling to dynamic nonlinear modeling, from manual feature extraction to intelligent mining based on data, from single target income to comprehensive goal of comprehensive income, risk and constraint. The traditional theoretical framework has gradually shown its drawbacks in today’s complex financial market. At the same time, machine learning has powerful information mining and dynamic matching capabilities, and has become the core technology to solve these problems and promote efficient frontier.

This field has now entered a new stage of investigating theories and deepening constraints. Although there are a lot of empirical findings, the problems such as no basic theory, scattered research ecosystem, pseudo-factors and overfitting have not been solved. Cutting-edge technologies such as causal inference, adversarial robust learning and explainable AI provide us with good ways to solve these challenges. From an academic point of view, the most urgent task is to fill the gap in the underlying theory, integrate scattered research modules together, and establish an overall framework combining dynamics, causality, and robustness. From the perspective of industrial practice, we should not only look at the profit, but also strike a balance between the predictability, stability and interpretability of the model, so that these technologies can be used well. This paper system reviews the research situation, summarizes the technical application, finds important research gaps, and makes it into a “dynamic-causal-robust” framework. Thus, the research value and innovation scope of the effective boundary theme of machine learning-enhanced multifactor models are clarified. Under the background of the close combination of machine learning and finance, this research direction has great potential for exploration and practical application of theory.

REFERENCES

[1] Markowitz H. Portfolio Selection [J]. The Journal of Finance, 1952, 7 (1):77-91.

[2] 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.
[3] Gu S, Kelly B, Xiu D. Empirical Asset Pricing via Machine Learning [J]. The Review of Financial Studies, 2020, 33 (5):2223-2273.
[4] Tan J. Improving portfolio efficiency with principal component analysis [J]. Journal of Portfolio Management, 2012, 38 (3): 87-96.
[5] López de Prado M, Simonian J, Fabozzi F J. Enhancing Markowitz’s portfolio selection paradigm with machine learning [J]. Annals of Operations Research, 2024, 346 (1):319-340.
[6] Zhao Zhenquan, Ding Zhiguo. Research Progress and Review of Multi-Factor Asset Pricing Models [J]. Economic Perspectives, 2019 (08): 116-132.
[7] Zhang Chengsi, Li Jintao. Advances and Challenges in the Application of Machine Learning in Quantitative Factor Investing [J]. China Finance Review International, 2023 (2):89-106.
[8] Wang Haoyu. A Study on the Application of Causal Inference in Financial Factor Screening [J]. Statistics & Decision, 2024 (8):172-176.
[9] Chen Xingyu. Application of Deep Learning in Multi-Factor Stock Selection for A-Shares and Research on Overfitting Prevention [J]. FinTech Era, 2024 (05): 45-52.
[10] Liu Chang. Dynamic Factor Models and Portfolio Efficient Frontier Optimization [J]. Journal of Applied Statistics and Management, 2022 (3): 521-533.
[11] Zhou Mingyuan. A Review of the Application of Explainable Machine Learning in Factor Investing [J]. Times Finance, 2024 (12): 68-72.
[12] Li Yang, Chen Shoudong. A Review of Methods for Optimizing the Portfolio Efficient Frontier [J]. Contemporary Economic Research, 2021 (05): 91-98.