Asset pricing models with machine-learning method

Cancan Zhang
School of Business, City College of Dongguan, Guangdong, China
Email: 36690772@qq.com

Liangliang Zhang*
School of General Education, City College of Dongguan, Guangdong, China
Email: 406666093@qq.com

Yajuan Yang
School of Digital Economics, City College of Dongguan, Guangdong, China
Email: 15201666@qq.com

Kongyan Chen
School of Digital Economics, Dongguan City College, Guangdong, China
420720639@qq.com

2022 18th International Conference on Computational Intelligence and Security (CIS) | 979-8-3503-4627-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/CIS58238.2022.00021

***Abstract*—Traditional asset pricing theories and models are facing more and more challenges in empirical study. Machine learning provides a new tool for asset pricing research. Due to the low signal-to-noise ratio and concept drift of financial data, the theoretical constraints of economics are very important for the applicability of machine learning in asset pricing. Firstly, this paper introduces seven multi-factor asset pricing models based on ad hoc sparsity constraints, summarizes the characteristics and shortcomings of traditional asset pricing models. Then, we display the challenges of machine learning facing in empirical application of asset pricing, formulate the targeted economic constraints. Finally, we further discuss the possible future trends of machine learning algorithms in asset pricing.**

***Keywords*-Asset pricing, Machine learning, Expected return, Sparsity constraint, Concept drift**

I. Introduction

The prediction of asset price has always been the core content in the field of asset pricing, which mainly estimate the trading price of financial products, especially the stock price in the capital market. Since 1970s, many scholars have begun to look for predictor variables that can explain the cross-sectional differences or time series changes of asset returns, on this basis, many classical asset pricing models have been proposed. Many predictor variables are usually calculated by stock price indices or financial data. In the overseas financial community, these predictor variables have a proper name: return predictors. The current consensus is that systemic risk and investor behavior's deviation jointly drive the difference in the cross section of asset return. Therefore, return predictors include predictors representing systemic risks and emotional factors representing investors' behavioral deviations.

Traditional asset pricing models explain cross-sectional differences of stock returns through several multi-factor models with a few predictors. In fact, these multi-factor models only select a few factors from a large number of optional company characteristics. The isolated study, only focusing on a small number of factors, shows that scholars have added a strong sparsity constraint to the models. Although the sparsity constraint can ensure the good performance of traditional statistical methods, but the sparsity constraint is ad hoc and varies with each individual.

With the development of technology, people can construct hundreds of predictor variables from public information. However, a large number of potential predictor variables will lead to inapplicability of traditional statistical methods (OLS). In addition, based on the sparsity assumption, traditional asset pricing models assume that investors have enough time to learn the functional relationship between predictor variables and predicted variables in a stable environment, while the model reflects the equilibrium state after investors learning the functional relationship. However, in the high-dimensional environment, the assumption that investors have learned the functional relationship becomes unconvincing. Therefore, there is a mismatch between the prediction difficulty faced in theoretical models and that faced in the real world by investors, which leads to unsatisfactory empirical performance of existing asset pricing theories.

Machine learning toolbox offers a new opportunity for people to analyze asset prices, without imposing extreme and ad hoc sparse assumptions. In empirical studies, machine learning tools allow econometricians to consider the interactions among a large number of predictor variables. In theoretical research, machine learning tools can provide inspiration for economic decision-making modeling in complex and uncertain high-dimensional environment, without making unrealistic simplification of the environment.

II. Multi-factor models based on "ad hoc" sparsity assumption

Asset pricing follows two basic methods: absolute pricing and relative pricing. The typical representative of absolute pricing theory should be the modern portfolio theory(MPT) proposed by Markowitz, a Nobel laureate in economics [1]. As the popularity of multi-factor pricing models, the attention of academic circles shifts from absolute pricing models to relative pricing models.

Since Fama & French (1993) [2] published and put forward the first multi-factor asset pricing model, the academic circles have studied the multi-factor models for more than 30

***For correspondence**

979-8-3503-4627-5/22/$31.00 ©2022 IEEE
DOI 10.1109/CIS58238.2022.00021

years. Many new models have been proposed successively, at present, there are seven mainstream multi-factor models.

A. Fama-french three-factor model

CAPM was the first paradigm of asset pricing before multi-factor models were proposed. However, since the 1970s, scholars have gradually found that stocks "packaged" according to a certain style can beat the market and achieve high returns, such as the earnings-to-price (EP) effect discovered by Basu (1977) [3] and the small-market value effect discovered by Banz (1981) [4]. Although the discovery of a single anomaly challenged CAPM, but they didn't form a combined force, so people did not have too much doubt on CAPM. Until Fama et al. (1993) integrated a variety of previously anomalies which completely subverting people's view of CAPM.

On the basis of CAPM, Fama et al.(1993) selected two firm characteristics: BM and size, according to 2×3 independent double sorting, formed the value factor (High-Minus-Low, HML) and the size factor (Small-Minus-Big, SMB), and proposed the classic three-factor asset pricing model (FF-3):

$$\begin{aligned} E[R_i] - R_f &= a_i + \beta_{i,MKT}(E[R_M] - R_f) \\ &+ \beta_{i,SMB}E[R_M] + \beta_{I,HML}E[R_{HML}] \end{aligned} \tag{1}$$

where $E[R_i]$ denotes the expected return of stock i and R_f is the return of risk-free asset, while$E[R_M]$ is the expected return of market portfolio, $E[R_{SMB}]$ and $E[R_{HML}]$ are the expected return of SMB and HML factors. While $\beta_{i,MKT}$, $\beta_{i,SMB}$ and $\beta_{i,HML}$ are the exposures of stock i on the corresponding factors respectively.

The FF-3 model can explain the co-movement of asset returns to a large extent, making most of the financial anomalies disappear. Along the FF-3 modeling idea, scholars have been looking for anomaly factors that cannot be explained by the FF-3 model in order to obtain new discoveries.

B. Carhart four-factor model & Novy-Marx four-factor model

With more and more test for FF-3 model, many unexplained financial anomalies have been discovered. For example, Jegadeesh & Titman(1993) [5] proposed the famous cross-sectional momentum anomaly. At the end of t month, they sorted the stock returns of a total of 11 months from $t-12$ to $t-1$—excluding the t-month to avoid the interference of short-term reversal. They selected stocks with the highest return to construct the winner portfolio and buy long the winner, meanwhile chose stocks with the lowest return to construct loser portfolio and short selling the loser. Finally, they discovered the long/short portfolios can obtain excess return.

subsequently, Carhart (1997) [6] introduced the cross-sectional momentum factor (MOM) into the FF-3 model and proposed the Carhart four-factor model:

$$\begin{aligned} E[R_i] - R_f &= a_i + \beta_{i,MKT}(E[R_M] - R_f) \\ &+ \beta_{i,SMB}E[R_M] + \beta_{i,HML}E[R_{HML}] \\ &+ \beta_{i,MOM}E[R_{MOM}] \end{aligned} \tag{2}$$

Here $E[R_{MOM}]$ is the return of momentum factor, $\beta_{i,MOM}$ is the exposure of stock i on momentum factor.

At the same time, Novy-Marx (2013) [7] found that profitability was closely related to future expected return, and proposed a four-factor model:

$$\begin{aligned} E[R_i] - R_f &= a_i + \beta_{i,MKT}(E[R_M] - R_f) \\ &+ \beta_{i,HML}E[R_{HML}] + \beta_{i,UMD}E[R_{UMD}] \\ &+ \beta_{i,PUM}E[R_{PUM}] \end{aligned} \tag{3}$$

where $E[R_{PUM}]$ and $E[R_{UMD}]$ are the return of profit factor and momentum factor, $\beta_{i,PUM}$ and $\beta_{i,UMD}$ are the exposure of stock i on these factors.

It is worth noting that Novy-Marx(2013) uses UMD to represent momentum factor is constructed by univariate sorting, while the momentum factors in Carhart's four-factor model is constructed by independent double sorting, they are completely different.

C. Fama-french five-factor model

Based on the dividend discount model (DDM) and the results of Miller & Modigliani (1961) [9], Fama & French (2015) [8] added profit and investment factors in the FF-3 model and proposed a five-factor model (FF-5):

$$\begin{aligned} E[R_i] - R_f &= a_i + \beta_{i,MKT}(E[R_M] - R_f) \\ &+ \beta_{i,SMB}E[R_{SMB}] + \beta_{i,HML}E[R_{HML}] \\ &+ \beta_{i,RMW}E[R_{RMW}] + \beta_{i,CMA}E[R_{CMA}] \end{aligned} \tag{4}$$

where $E[R_{RMW}]$ and $E[R_{CMA}]$ are the expected return of profit factor and investment factor respectively, $\beta_{i,RMW}$ and $\beta_{i,CMA}$ are the exposure of stock i on two factors respectively. It is worth noting that the positive relationship between expected investment and expected return is contradictory to the negative relationship between expected investment and expected return derived from DDM.

D. Houxue Zhang four-factor model

Hou et. al. (2015) [10] proposed a pricing model including four factors: market, scale, profit and investment based on the "economics theory of physical investment "which is also known as Q-theory. So, the model is also named Q-factor model by the academic community.

$$\begin{aligned} E[R_i] - R_f &= a_i + \beta_{i,MKT}(E[R_M] - R_f) \\ &+ \beta_{i,ME}E[R_{ME}] + \beta_{i,I/A}E[R_{I/A}] \\ &+ \beta_{i,ROE}E[R_{ROE}] \end{aligned} \tag{5}$$

where $E[R_{ME}]$, $E[R_{I/A}]$ and $E[R_{ROE}]$ are the expected return of size, investment and profit factor respectively, $\beta_{i,ME}$, $\beta_{i,I/A}$ and $\beta_{i,ROE}$ are the exposure of stock i on the corresponding factor.

The four-factor model follows the NPV principle in corporate finance: A firm should continue to invest until the marginal return on investment equals the marginal cost of investment. This model uses ROE and the change rate of total assets as proxy variables representing earnings and investment. The results confirm that investment is negatively correlated with future stock returns, which is completely opposite to the empirical results of FF-5 model.

E. Multi-factor models based on behavioral finance

A large number of studies have shown that cognitive constraints and behavioral biases of investors cause various mispricing, while mispricing can explain many anomalies in the financial market.

Stambaugh & Yuan (2017) [11] proposed the first multi-factor model of behavioral finance. On the basis of market factor and scale factor, this model introduces management and performance factors to construct a four-factor model:

$$\begin{aligned} E[R_i] - R_f &= a_i + \beta_{i,MKT}(E[R_M] - R_f) \\ &+ \beta_{i,SMB}E[R_{SMB}] + \beta_{i,MGMT}E[R_{MGMT}] \\ &+ \beta_{i,PERF}E[R_{PERF}] \end{aligned} \tag{6}$$

where $E[R_{MGMT}]$ and $E[R_{PERF}]$ are the expected return of management factor and performance factor respectively, $\beta_{i,MGMT}$, $\beta_{i,PERF}$ and $\beta_{i,ROE}$ are the exposure of stock i on the corresponding factors.

The management and performance factors derive from the study of mispricing. Mispricing means that price deviates from intrinsic value. When the price is higher than the intrinsic value, the asset is overvalued, while when price is lower than intrinsic value, the asset is undervalued. The overvalued asset will have a low return rate in the future due to the price correction, whereas meanwhile the undervalued asset will have a high return rate in the future. In order to find indicators to measure that stock is overvalued or undervalued, this paper constructs mispricing indicators based on 11 anomalies that can't be explained by FF-3 model. The value of anomaly variables can be used to describe the direction and size of mispricing.

The three-factor model proposed by Daniel Hirshleifer & Sun (2020) [12] is another attempt to apply behavioral finance to asset pricing. In this paper, two behavioral factors are put forward from the long time and short time scales. The two factors together with the market factor form a three-factor model:

$$\begin{aligned} E[R_i] - R_f &= a_i + \beta_{i,MKT}(E[R_M] - R_f) \\ &+ \beta_{i,FIN}E[R_{FIN}] + \beta_{i,PEAD}E[R_{PEAD}] \end{aligned} \tag{7}$$

Here $E[R_{FIN}]$ and $E[R_{PEAD}]$ respectively represent the expected return of two behavior factors with long period and short period, While $\beta_{i,FIN}$, $\beta_{i,PEAD}$ are the exposure of stock i on the corresponding factors. Two behavioral factors aim to capture mispricing due to overconfidence and limited attention. According to the research results, the two models based on behavioral finance can explain most anomalies of the stock market. The explanatory power of the two models is no less than that of traditional multi-factor models.

At present, the mainstream multi-factor models all contain a very limited number factors. There are two main reasons for this. Firstly, due to the constraints of data acquisition and computing, the early multi-factor model could only contain a few number factors. Secondly, following the law of parsimony, so, the seven mainstream multi-factor models try to contain 3 up to 5 factors.

The above seven mainstream multi-factor pricing models are all based on their own special sparsity assumption, and only contain a few predictors, while the influence of other pricing factors on the return is regarded as zero. However, this sparsity assumption of theoretical models is imposed by the researchers for their own purposes which is not general and does not fully reflect the investment environment for investors. At the same time, considering a large number of potential predictor variables will raise a statistical problem. If the number of predictor variables is larger than the number of stock returns, the OLS solutions are not unique, and OLS will overfit the noise of the data, resulting in a good fit in the sample, but bad performance out of sample.

III. Asset pricing via machine learning

In recent years, the application of machine learning methods in empirical asset pricing generally includes four aspects: (1) factor premium estimation; (2) Aggregate company characteristic information or latent factor model; (3) Forecast stock return; (4) Estimate the random discount factor. In the field of asset pricing, the two most crucial functions of machine learning are "price prediction" and "feature selection".

A. Challenges in empirical of asset pricing

Although machine learning methods have achieved impressive success in the application of artificial intelligence, medicine and other fields, compared with the data environment of traditional machine learning, the noise of financial data is very high, so, machine learning algorithms will inevitably face some unique challenges of the empirical application in asset pricing.

Firstly, the signal-to-noise ratio of financial data set is very low which is prone to overfitting. In the application of return prediction, since the expected return is unobserved, the realized return is regarded as a noisy signal of the expected return. Because the volatility of the expected return is only a small part of the volatility of the realized return, once the realized return of assets is used as the training

dataset of the algorithm, in both the cross section and the time series, the SNR of the training dataset is very low.

Secondly, the data-generating process in financial markets may be undergoing continuous structural change. In other machine learning application background, the availability of data and whether it is used in the decision-making process or not does not affect the subsequent data generation process, that is, the data generation process is stable. However, in asset pricing field, investors learn from data. When the historical data is analyzed and used, it may change the investment behavior of investors, affect the subsequent data generation process, and make the previous prediction relationship that is no longer valid, that is, the internal data generation process of asset return is non-stationary. In addition, the overall economy is constantly undergoing structural change. Production technologies, regulatory policies and the institutional environment all have changed dramatically over the past few decades. Therefore, under this background, there is no stable relationship between firm characteristics and future returns. So, data structural change that is also known as concept drift challenges the traditional machine learning approach in model validation and hyperparameter tuning using set-aside and cross-validation methods.

Finally, the sample number is limited. Since there is only one real asset pricing path, it is easy to overfit by repeatedly training based on this path. Although relevant studies often use walk-forward back-testing, but this method can only avoid future data problems and can't completely eliminate overfitting. In fact, this problem may be particularly serious in return-forecasting research. In addition, return prediction is usually based on the monthly stock returns, but even for the long-established U.S. stock market, the relatively complete historical data only goes back to 1962, with a sample number of about 700 months, which is too small for machine learning algorithms.

To sum up, machine learning algorithms that is only driven by data are not applicable in asset pricing field under the unconstrained and flexible framework. Therefore, the theoretical constraints of economics are very important for the applicability of machine learning.

B. Adding economically motivated constraints

To more closely integrate machine learning methods with asset pricing theory and existing empirical methods, it is necessary to add constraints with economic motivation and formulate targeted machine learning methods.

Firstly, the Bayesian prior assumption is introduced which is crucial to obtain a good out-of-sample R^2. The nature of data in asset pricing applications is very different from the nature of data in technology, medicine, and other scientific fields. Therefore, if we want to successfully use machine learning methods in asset pricing field, we need to make some adjustments and introduce the prior knowledge of economics about the data generation environment. In addition, according to the "no free lunch theorem " (Wolpert, 1996) [13], unless we have some prior knowledge of prediction problems, there is no reason to believe that one algorithm will be superior to another. Even if several algorithms perform equally well in the training dataset, we still cannot determine without prior knowledge which algorithm will produce better predictions performance in the test dataset. Therefore, only with specific prior knowledge to the prediction problem that you are trying to solve, it is possible to find an algorithm that still produces good prediction performance in the test data. Parameter estimation and Bayesian statistics of regularization provide a framework for incorporating prior knowledge into statistical estimation, bridging the gap between economic theory of asset pricing and machine learning.

Secondly, if there are structural changes in the data, it is necessary to find out whether it is appropriate for the verification data to precede all or part of the training data in time, because the directionality of time is very important. To address the problem, many methods allow parameters to vary over time in order to accommodate structural changes of the data. The simplest example is the rolling window estimation method, in which data beyond a certain period is discarded directly and not used in the estimation process.

Finally, the predictor variables are standardized. In the prediction of stock return, we are not necessarily interested in the accurateness of individual stock return prediction, We are more concerned about building a portfolio with good risk-return characteristics. The overall portfolio volatility is largely determined by the covariance properties of forecast errors. So, even if R^2 improves out of sample, there is no guarantee that the performance of portfolio will also improve out-of-sample. Standardize the variance of forecast variables to avoid heteroscedasticity and correlation problems. The cross-sectional variance of each predictor variable is the same, the variables are uncorrelated with each other, at this time, whether regularization is applied to machine learning methods or not will not affect the risk-return characteristics of the portfolio.

IV. CONCLUSION

At present, application of machine learning in asset pricing is still in the early stage of exploration, the research focus mainly on return prediction or closely related issues. With the introduction of machine learning tools, innovation means using new data or algorithms. using machine learning tools to predict fundamental financial indicators or mining the nonlinear relationship between factors and returns has become a research hotspot. Some innovations in the latest literature mainly focus on searching for new alternative data that can become revenue sources (such as public opinion data, patent data, news data, etc.), discovering new anomalies, improving existing predictive variables. However, there is still a lot of work to be done to integrate machine learning

methods with asset pricing theory and empirical methods closely. There are two promising directions for machine learning methods in the future: one is to empirically estimate the asset demand system based on investors' detailed portfolio holding data, the other is to analyze expectation data of investors. Both directions generally belong to the field of asset demand analysis.

ACKNOWLEDGEMENT

The research was supported by Dongguan Science and Technology Bureau (Guang Dong Province, China) under grants 20211800900692 and by the "Intelligent Financial Course Group" under grants of Curriculum Office from Dongguan City College (Guang Dong Province, China).

REFERENCES

[1] Markowitz, H. M. (1952). Portfolio selection. The Journal of Finance, 7(1),77-91.

[2] Fama, E. F., and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. Journal of Financial Economics, 33(1), 3-56.

[3] Basu, S. (1997). Investment performance of common stocks in relation to their price-earnings ratios: a test of the efficient market hypothesis. Journal of Finance, 32(3), 663-682.

[4] Banz (1981). The relationship between return and market value of common Stocks. Journal of Financial Economics , (9) , 3-18.

[5] Jegadeesh, N., and Titman, S.(1993). Returns to buying winners and selling losers: Implication for stock efficiency. Journal of Finance, 48(1),65-91.

[6] Carhart (1997). On persistence in mutual fund performance. Journal of Finance, 52(1), 57-82.

[7] Novy-Marx, R. (2013). The other side of value: The gross profitability premium. Journal of financial Economics, 108(1), 1-28.

[8] Fama, E. F.& French, K. R. 2015. A five-factor asset pricing model. Journal of Financial Economics.116(1):1-22.

[9] Miller, M. H., & Modigliani, F. (1961). Dividend policy, growth, and the valuation of shares. Journal of Business, 34(4), 411-433.

[10] Hou, K., Xue, C., & Zhang, L. (2015). Digesting anomalies: An investment approach. Review of Financial Studies, 28(3), 650-705.

[11] Liu, Stambaugh, & Yuan (2019). Size and value in China. Journal of Financial Economics, 134(1), 48-69.

[12] Daniel, K. D., Hirshleifer, D. A., & Sun, L. (2020). Short-and long-horizon behavioral factors. Review of Financial Studies, 33(4), 1673-1736.

[13] Wolpert, D. H. (1996). The Lack of a Priori Distinctions Between Learning Algorithms. Neural Computation, 8,1341-1390.