Abstract

Like physics, modern financial theory has highly mathematicized itself but such practice brings several unneglectable challenges. Firstly, there is an accusation, namely *physics envy*, against expressing some fundamental financial concepts in terms of mathematics, which has been seen as an unwarranted push for reductionism but not a scientific setting for the complex reality, for example, the arguing from practitioners about the efficient market hypothesis (EMH) and the rational man hypothesis. Secondly, most financial research uses historical observation, instead of experiment, for hypothesis formulation, modeling and testing, but observations in society are always evolving with the evolvement of society itself. For example, market emotion is not measurable until the last two decades while the sentiment index now acts as an important criterion for practical investment decision-making. To refine financial theory, big data and machine learning (ML) might provide an opportunity. Since machine learning is capable of extracting shielded information from complicated and multistructural data, such as SNS and satellite images, the shielded information can not only help test some fundamental concepts in the current financial theory for improving our understanding of the real financial market but also envision some ignored factors and relations for financial modeling. Moreover, Fintech, such as ML for arbitrage opportunity searching, has gradually been applied in the financial industry, and it is believed that such progress will possibly reshape the financial market in form and even in substance. Therefore, it is worthy of applying and developing ML methods in the financial market as well as keeping observing the changes that ML could bring into the financial market.

In Chapter 2, in an attempt to test whether the raw message and the account follower number of stock tweets can help predict daily closing price movement, BERT-based natural language processing (NLP) is implemented. Compared with the random guess, the backtesting result shows that such tweet information can't bring significant improvement for the daily price movement prediction and it suggests that the daily EMH still holds with such information.

In Chapter 3, I propose the *density optimization* which is based on the DBSCAN clustering to investigate whether the dense subset generated by *density optimization* can outperform the original portfolio. The backtesting result proves the outperformance of the dense subset under the Sharpe ratio measure in most cases, but the dense subset fails to achieve stable outperformance under the Jenson alpha measure. By introducing a modification based on the property of the dense subset generated by *density optimization*, the model variant *fractional density optimization* generates the fractional dense subset that outperforms the original portfolio when the original portfolio has small duration and high winning ratio in most cases under the Jenson alpha measure.

In Chapter 4, the relationship between participants' reliance on algorithms, their familiarity with the task, and the performance level of the algorithm is investigated. Experiment results show that when participants could freely decide on their final forecast after observing the one produced by the algorithm (a condition found to mitigate algorithm aversion), the average degree of reliance on high and low performing algorithms did not significantly differ for participants with little experience in the task. Experienced participants relied less on the algorithm than inexperienced participants, regardless of its performance level. The reliance on the low performing algorithm was positive even when participants could infer that they outperformed the algorithm. Indeed, participants would have done better without relying on the low performing algorithm at all. The results suggest that, at least in some domains, excessive reliance on algorithms, rather than algorithm aversion, should be a concern.