Risk Premium Measurement in China Stock Market: Based on Machine Learning Algorithms

(Research Proposal)

Xingkong Wei[1], Huihang Wu[1], Xiaoyan Zhang[1] and Xinran Zhang[1]

1: PBC School of Finance, Tsinghua University

Introduction:

China has the world's second-largest stock market, and it has become an important part of the global capital market. However, the Chinese stock market risk premia is not well understood by global investors. There are two main challenges.

First, risk premia are notoriously difficult to measure. The collection of candidate conditioning variables for the risk premium is large and the relationship between the two is nonlinear. Hou et al. (2019) list 452 firm characteristics in the anomalies literature. Traditional methods break down when the predictors number increase and machine learning methods grow fast in this literature. Kelly & Pruitt (2015) use dimension reduction methods (such as principal component analysis and partial least squares regression) to estimate stock risk premia; Chinco et al. (2018) apply LASSO models to high frequency return forecasts; Gu et al. (2018) find that tree models and neural network models work well. Second, China has unique political and economic environments, which are quite different from those in the US and other developed economies, such as IPO constraints and suspension rules. To deal with this issue, Liu et al. (2019) explore and develop factor models in China by cutting the smallest 30% of stocks in A share market.

In this paper, we try to provide a new benchmark of measuring risk premia via machine learning algorithms in the context of China.

Objectives:

- Can machine learning algorithms work well in Chinese stock market?
- How to adjust our model setups for Chinese stock market?
- What kind of characteristics are important in China?
- How large are economic gains from machine learning forecast?

Research Strategy:

- Model Setup

Following Gu, Kelly, and Xiu (2018) setups, we describe an asset's excess return as an

additive prediction error model:

$$r_{i,t+1} = E_t(r_{i,t+1}) + \epsilon_{i,t+1}$$

Where

$$E_t(r_{i,t+1}) = g^*(z_{i,t})$$

Stocks are indexed as i = 1,…, N and months by t = 1,…, T.

$r_{i,t+1}$ is stock i realized return in time t+1.$E_t(r_{i,t+1})$ is expected return for stock i in time t. $z_{i,t}$ is stock i firm characteristics in time t. $g^*(.)$ is flexible function of these predictors.

- Machine learning algorithms and the computing language

We try to perform a comparative analysis of methods in the machine learning repertoire, including principal component analysis, partial least squares regression, support vector regression, boosted regression trees, random forests, and neural networks. We build our models on PYTHON, an interpreted, high-level, general-purpose programming language.

- Data

We obtain monthly total individual equity returns and financial statements from WIND (Wind Information Inc.) for all firms listed in the Shanghai and Shenzhen Stock Exchanges. Our sample begins in January 1991 and ends in December 2017. Following Hou et al. (2019), we construct 214 firm characteristics based on data availability.

References

Chinco, A., Clark Joseph, A. D. and Ye, M., 2018, "Sparse Signals in the Cross-Section of Returns", *The Journal of Finance*, 74: 449-492.

Gu, S., Kelly, B. T. and Xiu, D., 2018, "Empirical Asset Pricing Via Machine Learning", *Working Paper*.

Hou, K., Xue, C. and Zhang, L., 2019, "Replicating Anomalies", *The Review of Financial Studies*, (forthcoming).

Kelly, B. and Pruitt, S., 2015, "The Three-Pass Regression Filter: A New Approach to Forecasting Using Many Predictors", *Journal of Econometrics*, 186(2): 294-316.

Liu, J., Stambaugh, R. F. and Yuan, Y., 2019, "Size and Value in China", *Journal of Financial Economics*, (forthcoming).