Empirical Asset Pricing via Machine Learning Q&A

Data

1. Q: Where can I download the data?

   A: You can go to Prof. Dacheng Xiu's homepage (dachxiu.chicagobooth.edu) and find the data download link under the paper title.

2. Q: What's data preprocessing method you use?

   A: We do cross-section rank transformation on all firm characteristics for each month (set values to [-1, 1]). The response variable is excess return (raw return minus risk-free rate).

3. Q: How to deal with missing data?

   A: We replace all missing values of firm characteristics with 0. We remove the samples with missing returns.

Methodology

1. Q: What is pooled OLS?

   A: Stack all cross-section and time-series samples and do one overall regression.

2. Q: What's the denominator of the R^2 definition?

   A: For panel R^2, we use 0 prediction as R^2 denominator. For all portfolios R^2 and factors R^2, we use historical average returns as R^2 denominator. Because 0 prediction is a better estimation for individual stocks, while historical average returns are better for portfolios. The denominator of R^2 definition is "the benchmark model", and R^2 is the percentage of MSE improvement scale.

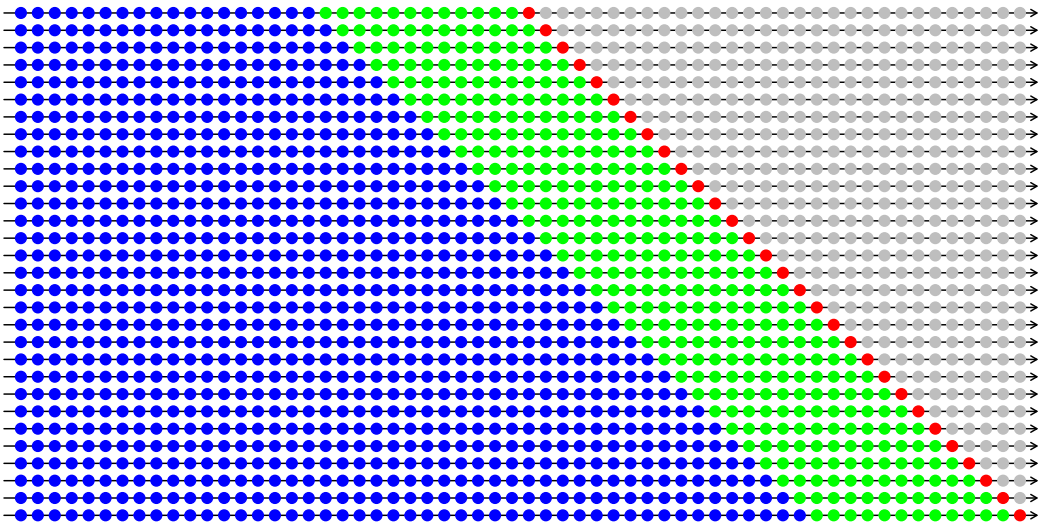
3. Q: What is the R^2 variable importance definition?

   A: The R^2-VIP is the decrease of in-sample R^2 when inforcing one variable equal to 0.

4. Q: How to split training/validation/testing samples?

   A: We use a recursive strategy when training the model. See the following plot. There are total 30 years in OOS period (1987-2016). We re-train our models by year. E.g. For OOS year 1987, the training period is 18 years (1957-1974), the testing period is 12 years (1975-1986). For next OOS year 1988, we increase training period by one year and keep the length of testing

period (but rolling forward by one year). So that there are 19 years in training and 12 years in testing.



5. Q: How to choose the tuning parameters?

   A: Please check the Internet Appendix Table A.5. We use grid search for λ in regressions.

Portfolios

1. Q: What is the long-short portfolio in the paper?

   A: At the end of month T, we calculate one-month-ahead out-of-sample stock return predictions for each method. We then sort stocks into deciles based on each model's forecasts. And we buy the highest expected return stocks (decile 10) and sells the lowest (decile 1). At the end of month $T+1$, we can calculate the realized returns of the portfolios (buy side and sell side respectively).

2. Q: What if the portfolios exclude micro-cap tickers?

   A: We do calculate the long-short portfolio excluding bottom 20% micro-cap tickers. The best SR is 1.69. Please check Table A.10 in the Internet Appendix.

Simulation

1. Q: Where can I download the simulation codes?

   A: You can go to Prof. Dacheng Xiu's homepage (dachxiu.chicagobooth.edu) and find the github link under the paper title. We provide a part of our simulation codes.