THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

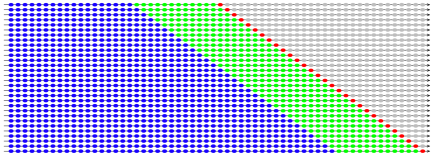
MATH5470 Final Project
Paper Replication: Empirical Asset Pricing via Machine Learning

CUI Daorong[1], LI Meng[1], HE Jiayi[1]; {dcuiab, mlieh, jhecy}@connect.ust.hk
[1] Department of Mathematics, The Hong Kong University of Science and Technology (HKUST)

Problem Overview

1. Traditional asset pricing methods have potentially severe limitations that more advanced statistical tools in machine learning can help overcome. In this project, we reproduce some of the work in Gu et al. (2020), in which they provide a comprehensive analysis of expected returns using machine learning techniques.
2. The six methods we choose are: OLS, OLS-3, PCR, PLS, RF, elastic net, and NN1 to NN5.
3. We use the out-of-sample R^2 as the performance metric and use a 'recursive strategy' (see in the following plot) when evaluating our models.
4. We discover influential covariates by ranking them according to a notion of variable importance VI_j, which is the decrease of in-sample R^2 when enforcing variable j equal to 0.
5. After data preprocessing, our dataset contains 3,502,067 effective samples, each with a numerical response variable and 920 numerical baseline covariates.

Fig. recursive scheme plot

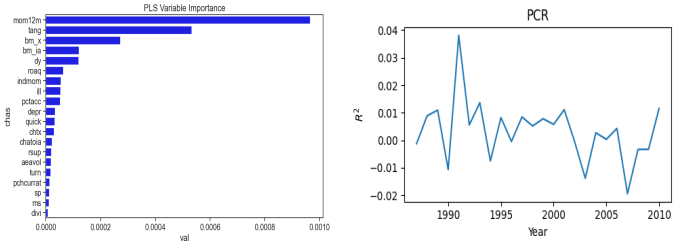


Hyperparameters for All Methods

1. OLS-3:
Huber loss $\xi = 99.9\%$ quantiles.
2. PLS:
Component: first {5, 15, …, 40}, then {1, 2, …, 5}.
3. Elastic net:
λ={1, 0.1, 0.01}; ρ={0.5, 0.3, 0.1}.
4. PCR:
n_components=15.
5. RF:
Depth=1~6; #Trees={100, 200, 300, 400, 500, …, 900};
#Features in each split in {3, 5, 10, 20, 30, 50, …}.
6. NN1 to NN5:
L1 penalty=1e-4; Batch size=10000; Epochs=50; Adam optimizer (LR=0.005); Patience=5; Ensemble=0.

Main Results

1. Some Visualizations (*VI*, time-varying R^2)



2. Monthly out-of-sample R^2 (%)
See in the following table.

3. Conclusions
We found that neural networks and PCR are the best performing methods. However, due to CPU capacity limitations, we were unable to evaluate all models for 30 years, especially the random trees, which we only ran recursively for 7 years. The most powerful predictors are associated with price trends, including return reversal and momentum, followed by measures of stock liquidity, stock volatility, and valuation ratios.

Table. Monthly out-of-sample R^2

OLS-3 (top 1,000)	-0.5876	NN1 (all)	0.2081
Elastic net	0.0081	NN2 (all)	0.0984
PLS	-0.1689	NN3 (all)	0.1684
PCR	0.3732	NN4 (all)	0.1554
RF	-0.0873	NN5 (all)	0.1548

References:
Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. The Review of Financial Studies, 33(5), 2223-2273.

Contributions: CUI Daorong (OLS, OLS-3, NN); LI Meng (elastic net, PLS); HE Jiayi (PCR, RF).

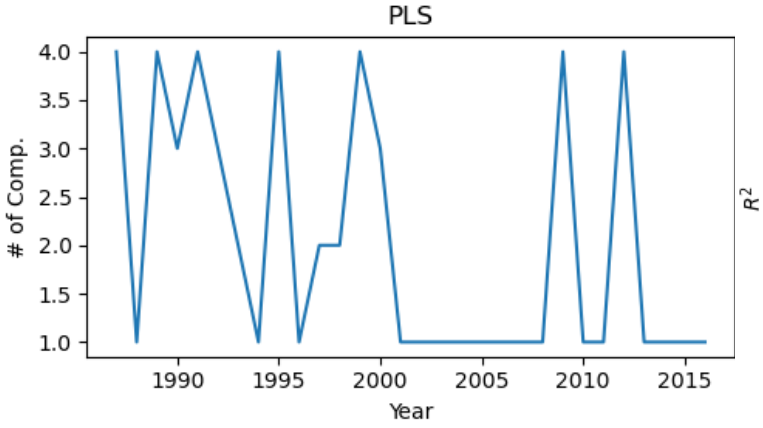
Partial Least Squares (PLS)

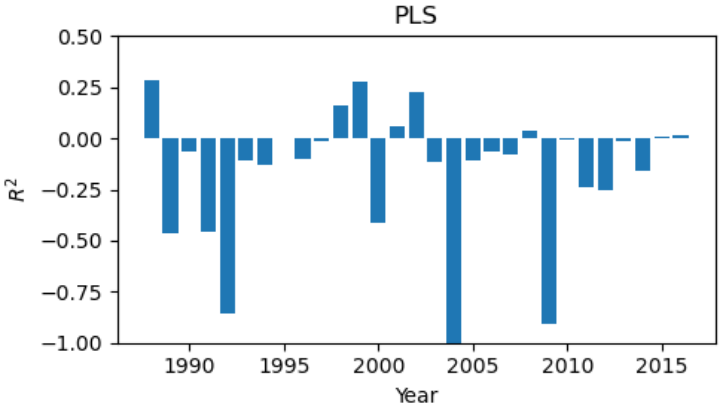
PLS looks for K-dimensional projections of Z that **maximize predicative association with the final target R**, so the j^{th} linear combination solves

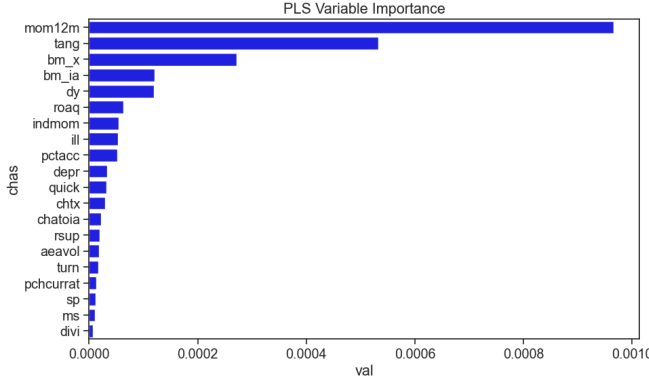
$$\omega_j = argmax_{\omega} Cov^2(R, ZW)$$

s.t.

$$\omega'\omega = 1, cov(Z\omega, Z\omega_l) = 0, l = 1, 2, \ldots, j-1.$$







Elastic Net

$$\min_{\theta} \quad \mathcal{L}(\theta) + \phi(\theta; \cdot)$$

where $$\phi(\theta; \lambda, \rho) = \lambda(1-\rho)\sum_{j=1}^{p}|\theta_j| + \frac{1}{2}\lambda\rho\sum_{j=1}^{P}\theta_j^2$$

Experiments: λ={1, 0.1, 0.01}; ρ={0.5, 0.3, 0.1}.