# Paper Replication:Empirical Asset Pricing via Machine Learning

SHANG Zhenhang, SUN Lei, QUAN Xueyang

Hong Kong University of Science and Technology

https://youtu.be/NAa_smC7X1I

# Table of Contents

# Table of Contents

# Typical Features

- Two themes in modern empirical asset pricing research: Understanding the differences of expected return among various assets; Concerning the dynamics of the overall equity risk premium.
- For the risk premium, the set of available conditioning variables is quite large.
- The uncertainty of the functional forms from high-dimensional predictors entering the risk premium.

# Potential Solutions via ML

- Risk premium measurement: the conditional expectation of the excess return realized in the future.
- Dimension reduction techniques help with reducing the degree of freedom among predictors.
- The diversity, nonlinear association approximations and parameter penalties can handle with the uncertain functional forms.

# Table of Contents

# Data Sources

- Monthly individual equity returns data of US stocks from Mar.1957 to Dec.2016.
- 30,000 stock samples in total, and 6,200 on month average.
- 94 stock company characteristic factors, 74 industry dummies (SIC classification standard) and 8 macroeconomic variables included.

# Data Splitting

- Training Set: From 1957 to 1986.
- Validation Set: From 1975 to 1986, used to tune the hyper-parameters.
- Testing set: From 1987 to 2016, used for evaluation.

# Table of Contents

# Methodology

- Simple linear as base reference and comparison.
- Penalized linear to perform shrinkage.

$$
\begin{aligned}
&\mathcal{L}(\theta;\cdot) = \mathcal{L}(\theta) + \phi(\theta;\cdot) \\
&\phi(\theta;\lambda,\rho) = \lambda(1-\rho)\textstyle\sum_{j=1}^{P}|\theta_j| + \frac{1}{2}\lambda\rho\sum_{j=1}^{P}\theta_j^2
\end{aligned}
\tag{1}
$$

# Methodology

- Dimension reduction via PCR and PLS.

$$
\begin{gathered}
w_j = \arg\max_w \operatorname{Var}(Zw), \quad \text{s.t.} \quad w'w = 1, \\
\operatorname{Cov}\left(Zw, Zw_l\right) = 0, \quad l = 1, 2, \ldots, j-1
\end{gathered} \tag{2}
$$

$$
\begin{gathered}
w_j = \arg\max_w \operatorname{Cov}^2(R, Zw), \quad \text{s.t.} \quad w'w = 1, \\
\operatorname{Cov}\left(Zw, Zw_l\right) = 0, \quad l = 1, 2, \ldots, j-1
\end{gathered} \tag{3}
$$

# Methodology

- Simple linear as base reference and comparison.
- Penalized linear to perform shrinkage.
- Dimension reduction via PCR and PLS.
- Generalized linear for non-parametric result.
- Boosted regression tree.

# Table of Contents

# Performance Evaluation

Out-of-sample $R^2$:

$$R^2_{\text{oos}} = 1 - \frac{\sum_{(i,t)\in\mathcal{T}_3} \left(r_{i,t+1} - \widehat{r}_{i,t+1}\right)^2}{\sum_{(i,t)\in\mathcal{T}_3} r^2_{i,t+1}} \tag{4}$$

# The Cross Section of Individual Stocks (Monthly)

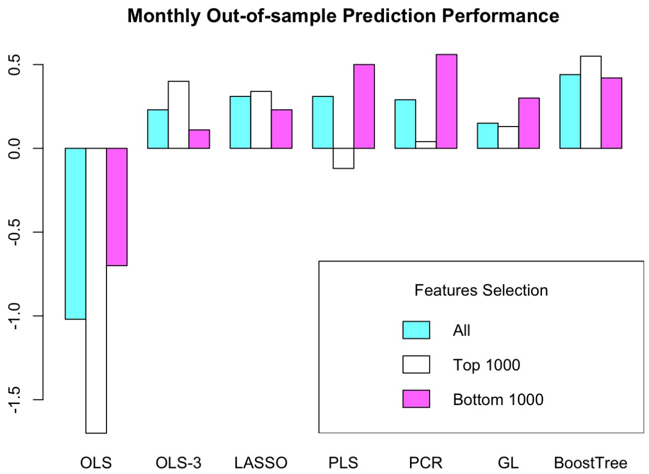



Figure 1: Monthly with OLS

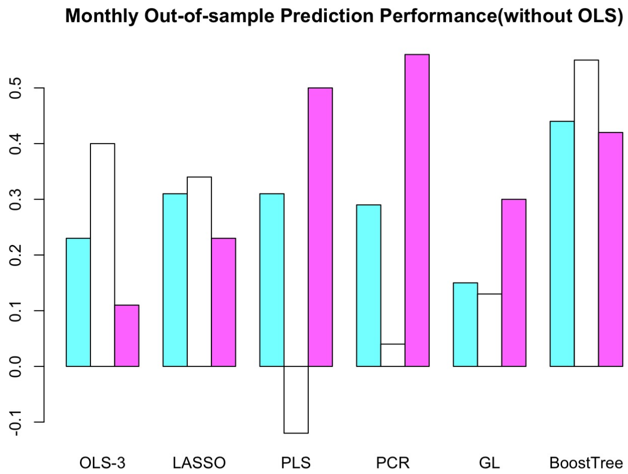
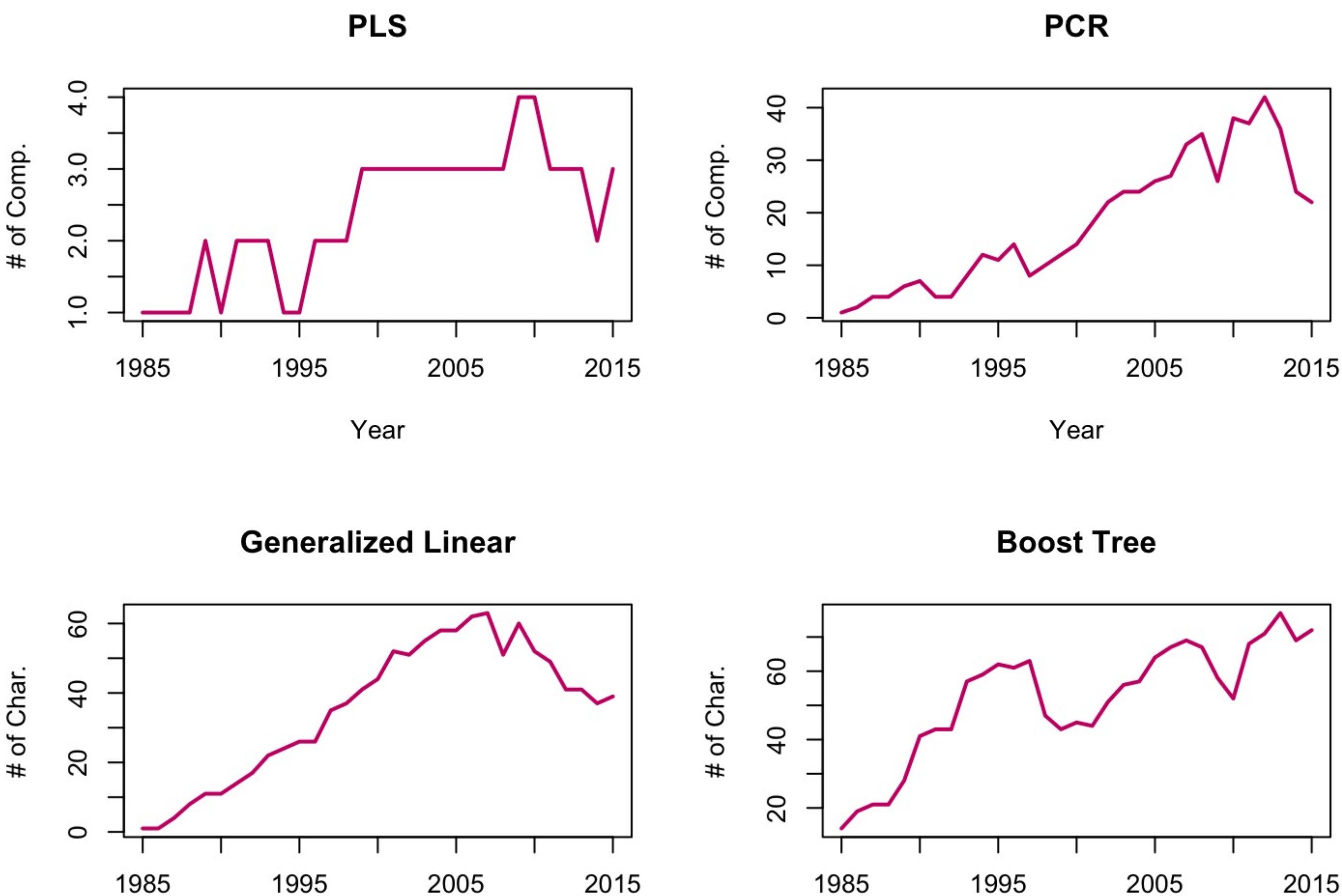



Figure 2: Monthly without OLS

# Time-varying Model Complexity



# The Cross Section of Individual Stocks (Annually)

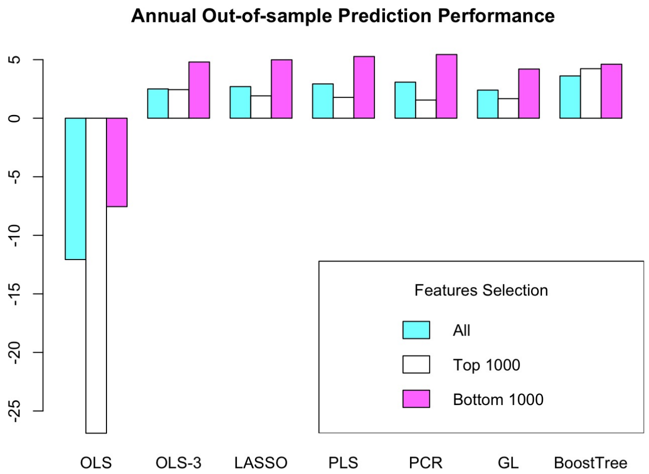



Figure 4: Annually with OLS

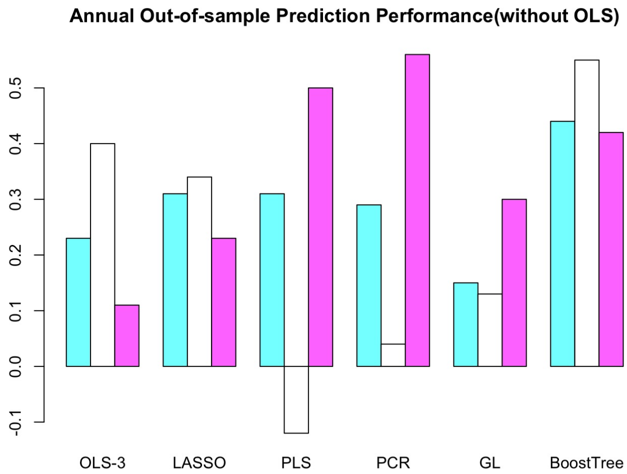
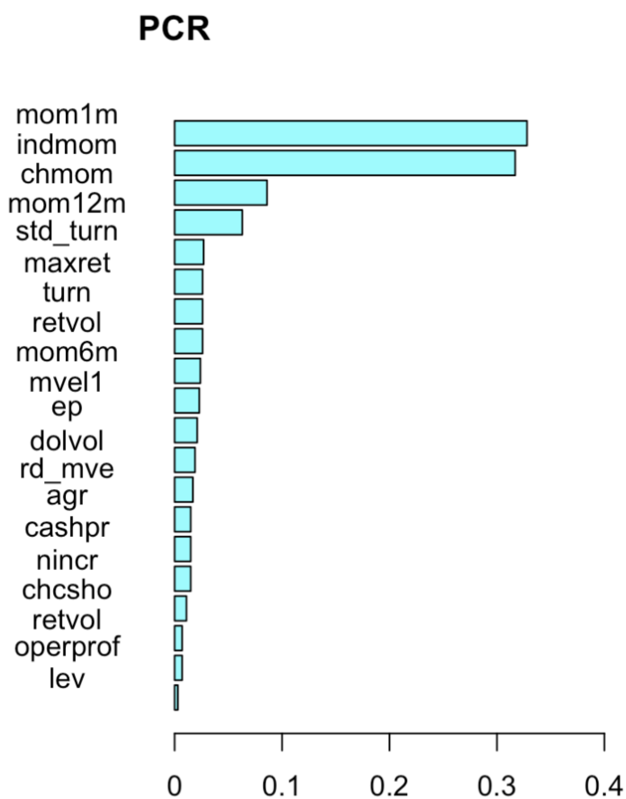



Figure 5: Annually without OLS

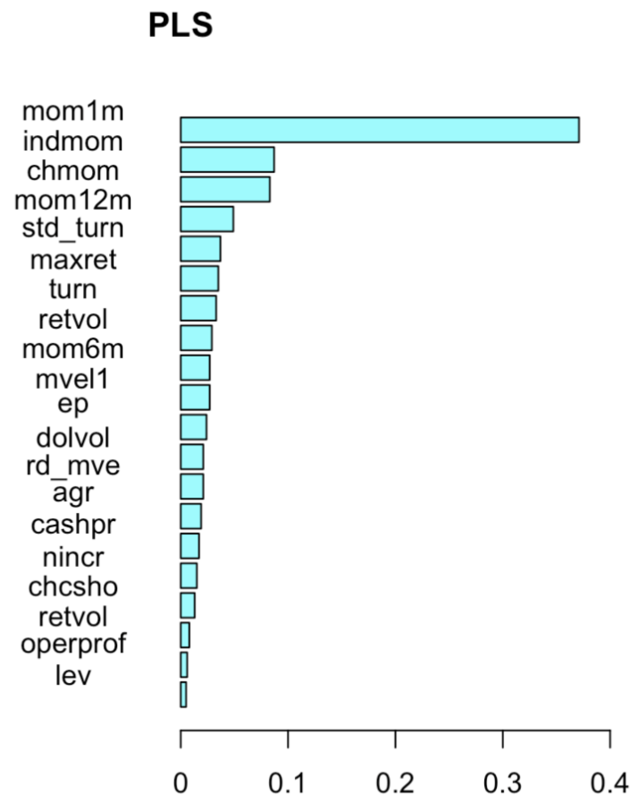
# Variable Importance

- When keeping all other variables unchanged and setting all values of the variable $j$ to 0, observe the reduction of the panel predictive $R^2$;
- Calculate the sum of squared partial derivatives (SSD) with respect to an input variable $j$.

# Variable Importance By Model I

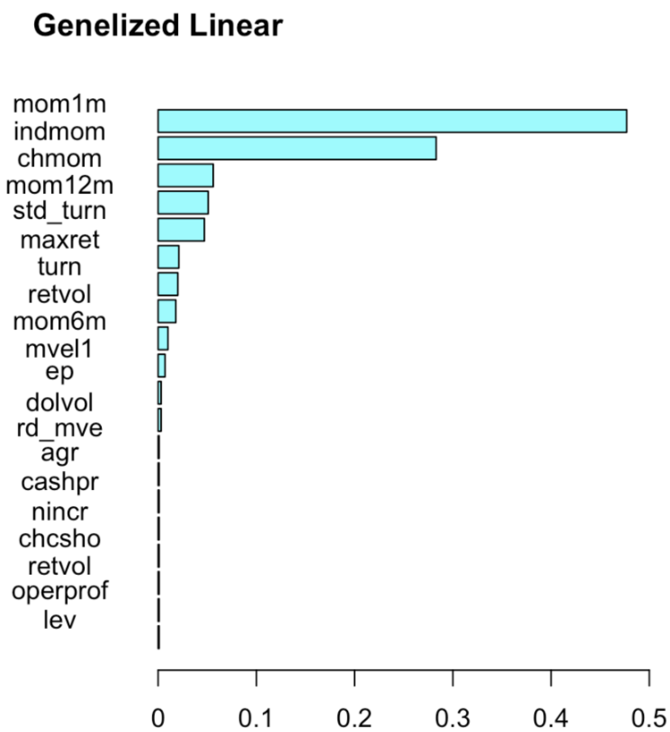



(a) PCR

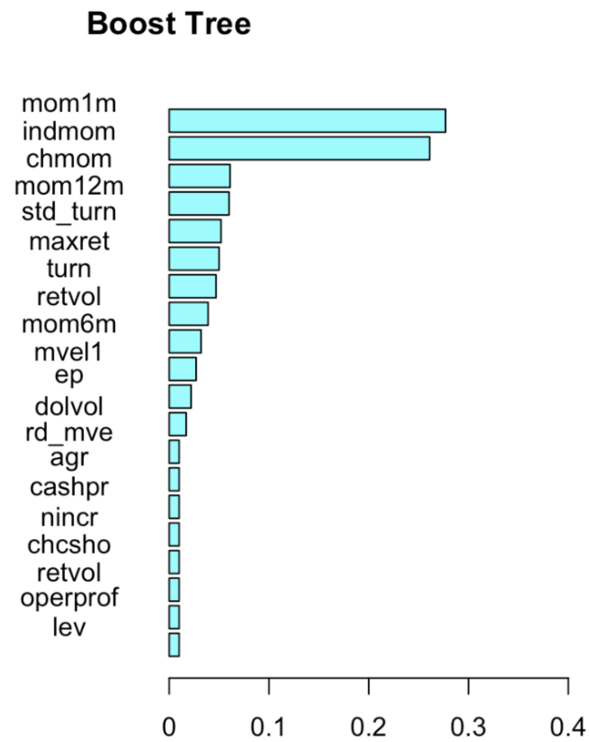



(b) PLS

# Variable Importance By Model II




(c) Generalized Linear




(d) Boost Tree

# Conclusion

- Tree models provide the best prediction performances.
- Ranking for variable importance:
  - Recent price trends
  - Liquidity
  - Risk measures
  - Valuation ratios and fundamental signals

# Table of Contents

# Reference

[1] Shihao Gu, Bryan Kelly and Dacheng Xiu (2020). Empirical Asset Pricing via Machine Learning, *The Review of Financial Studies*, vol.33, no.5, 2223-2273.

[2] de Jong, Sijmen (1993). Simpls: An Alternative Approach to Partial Least Squares Regression, *Chemometrics and Intelligent Laboratory Systems* 18, 251-263.

[3] Diebold, Francis X., and Roberto S. Mariano (1955). Comparing Predictive Accuracy, *Journal of Business & Economic Statistics* 13, 134-144.

[4] Breima Breiman, Leo, Jerome Friedman, Charles J Stone, and Richard A Olshen (1984). *Classification and regression trees*, (CRC press).

[5] B¨uhlmann, Peter, and Bin Yu (2003). Boosting with the l2 loss, *Journal of the American Statistical Association* 98, 324-339.

[6] Fama, Eugene F, and Kenneth R French (1993). Common risk factors in the returns on stocks and bonds, *Journal of financial economics* 33, 3-56.

# Contribution

- SHANG Zhenhang
  - Code in python for PLS, PCR, Generalized Linear replication and visualization.
  - Write PPT
  - Presentation
- SUN Lei
  - Code in python for OLS, OLS-3, Penalized Linear and Boost Tree replications and the integrate all the replicated model.
  - Write PPT
  - Presentation
- QUAN Xueyang
  - Write report
  - Write PPT
  - Presentation