The Factor Zoo: Machine Learning, Multiple Testing, and Return Predictability

He Mengqi
Anhui Polytechnic University, China

ABSTRACT

The asset pricing 'factor zoo'—the proliferation of hundreds of purportedly significant stock return predictors published in academic finance journals—has created a credibility crisis for empirical asset pricing. Harvey, Liu, and Zhu (2016) in the Review of Financial Studies documented that over 300 factors had been published by 2014, raising serious concerns about data snooping (the inflation of apparent statistical significance from testing many hypotheses on overlapping datasets) and multiple testing (the expected false discovery rate when many hypotheses are tested without adjustment). The typical t-statistic threshold of 2.0 used to claim statistical significance is woefully inadequate when hundreds of factors are tested: at $t > 2.0$, 5% of true null hypotheses are rejected by chance, and with 300 factors this implies 15 spurious 'discoveries.' McLean and Pontiff (2016) document that published factor returns decay substantially after publication—suggesting many factors are overfit in-sample. This paper reviews the factor proliferation problem, multiple testing corrections, machine learning approaches to factor discovery, and the standards required for claiming valid return predictability.

Keywords: Factor Zoo; Multiple Testing; Data Snooping; Machine Learning; Asset Pricing; Return Predictability; Harvey; Replication Crisis

1. INTRODUCTION

Academic finance's four-decade project of documenting return predictors has produced a large literature that is increasingly questioned on statistical grounds. The academic incentive structure—publication requires statistically significant results, and finance journals use standard $t > 2.0$ thresholds—combined with the practice of testing many predictors on overlapping samples creates conditions for spurious discoveries. The 'file drawer problem' (unpublished null results) means the published literature overrepresents statistically significant findings, and the reuse of CRSP stock return data across hundreds of studies on overlapping samples creates correlation between test statistics that further inflates the false discovery rate.

Harvey, Liu, and Zhu (2016) in the Review of Financial Studies (29(1): 5–68) argued that the appropriate t-statistic threshold should be approximately 3.0 for newly proposed factors—higher than the conventional 2.0—to maintain reasonable false discovery rates given the multiple testing problem. Applying this standard would disqualify many published factors. Hou, Xue, and Zhang (2020) in the Review of Financial Studies attempted to replicate 452 published anomalies and could not replicate approximately half at conventional significance levels, with many more failing to replicate out-of-sample.

2. THE MULTIPLE TESTING PROBLEM

2.1 Data Snooping Bias

Lo and MacKinlay (1990) in the Review of Financial Studies formalized data snooping bias in the context of financial anomalies: when researchers test many specifications on the same dataset, the best-performing specification will show an inflated t-statistic relative to its true population value. White's (2000) reality check and Romano and Wolf's (2005) stepwise multiple testing procedures provide bootstrap-based corrections for multiple testing in time series, but these are rarely applied in published anomaly research. Sullivan, Timmermann, and White (1999) apply these methods to published technical trading rules and find that most purportedly significant rules are indistinguishable from chance once multiple testing is properly accounted for.

2.2 Fama-French Factor Thresholds

Fama and French (2018) in the Journal of Finance propose that factors should meet four criteria to be taken seriously: (1) A reasonable economic explanation for why the factor should predict returns; (2) Persistence: the pattern should exist across different time periods; (3) Pervasiveness: the pattern should exist across different markets and asset classes; (4) Robustness: the pattern should survive to different reasonable definitions of the variable. Applying these criteria dramatically reduces the set of credible factors from hundreds to approximately a dozen well-documented anomalies.

3. MACHINE LEARNING AND FACTOR DISCOVERY

Table 1. Machine learning approaches in empirical asset pricing.

ML Approach	Application	Advantage / Risk
LASSO / Ridge regression	Variable selection from large factor set with penalization reducing overfitting	Reduces dimensionality; still requires train/test split discipline
Random forests	Nonlinear factor interactions; feature importance measures	Captures nonlinearities; requires large samples; less interpretable
Neural networks (GAN)	Chen, Kelly & Xiu (2021): deep neural network for asset pricing	State-of-the-art return prediction; 'black box' interpretation challenge
Double/debiased ML	Corrects regularization bias for inference on specific factors	Valid inference after variable selection; computationally intensive

4. STANDARDS FOR CREDIBLE FACTOR IDENTIFICATION

The replication crisis in asset pricing has motivated development of higher standards for factor credibility. Pre-registration—registering factor hypotheses and test designs before accessing data—is gaining traction in financial economics through the Social Science Replication Project and similar initiatives. Out-of-sample testing in international data (Harvey and Liu, 2020 use international data as an out-of-sample test for U.S. factors) provides a natural test bed that avoids most U.S. data snooping. Bayesian approaches (Harvey and Liu, 2016) explicitly incorporate prior beliefs about the probability that any given factor is true, producing more conservative posterior factor credibility assessments that account for the multiple testing problem.

5. CONCLUSION

The factor zoo crisis reflects a genuine scientific problem in empirical asset pricing: a combination of multiple testing, data snooping, publication bias, and insufficient out-of-sample validation has inflated the apparent evidence for return predictability. The appropriate response is not abandonment of factor research but methodological reform: higher t-statistic thresholds for newly proposed factors, mandatory out-of-sample testing, international replication, and pre-registration of factor tests. Machine learning methods offer genuine return prediction improvements when properly cross-validated and combined with economic theory, but require careful application to avoid introducing new forms of overfitting. Future research should develop standardized, publicly accessible factor testing protocols that reduce the variance of factor credibility claims and enable more reliable cumulative knowledge building in empirical asset pricing.

REFERENCES

[1] Fama, E. F., & French, K. R. (2018). Choosing factors. Journal of Financial Economics, 128(2), 234–252.
[2] Harvey, C. R., Liu, Y., & Zhu, H. (2016). ... and the cross-section of expected returns. Review of Financial Studies, 29(1), 5–68.
[3] Hou, K., Xue, C., & Zhang, L. (2020). Replicating anomalies. Review of Financial Studies, 33(5), 2019–2133.
[4] McLean, R. D., & Pontiff, J. (2016). Does academic research destroy stock return predictability? Journal of Finance, 71(1), 5–45.