The Factor Zoo and the Replication Crisis

Dennis Hoffmann

Abstract

Empirical asset pricing research has produced hundreds of firm characteristics claimed to predict cross-sectional stock returns. John Cochrane called this proliferation the "factor zoo" in his 2011 presidential address to the American Finance Association, and subsequent work has raised serious questions about how many of these factors are genuine. This article surveys the growth of the factor zoo, the replication crisis debate, the methodological foundations of multiple testing, and the emerging machine-learning approaches that aim to compress hundreds of candidate factors into a tractable set of priced risk exposures.

Contents

1 Introduction

Since the Capital Asset Pricing Model reduced expected returns to a single market factor (Sharpe, 1964), the cross-section of expected stock returns has been one of the most intensively studied areas in financial economics. Early anomalies such as the size effect (Banz, 1981) and the value premium (Fama and French, 1993) challenged the CAPM and motivated multi-factor models. By 2016, researchers had documented at least 316 distinct characteristics that purportedly predict returns (Harvey et al., 2016), and the count has continued to grow.

This proliferation raises a fundamental question: how many of these factors represent genuine economic phenomena, and how many are statistical artifacts of data mining? The answer matters not only for academic theory but also for practitioners who must decide which signals to trust when constructing portfolios. For an overview of how factor signals are translated into portfolio weights, see our companion article on quantitative portfolio construction.

2 The Growth of the Factor Zoo

2.1 A Brief History of Factor Discovery

The CAPM (Sharpe, 1964) posited that expected returns are determined solely by exposure to market risk. The first cracks appeared with the size effect: Banz (1981) showed that small-capitalization stocks earned higher average returns than the CAPM predicted. Fama and French (1993) formalized the value and size effects into the three-factor model (market, SMB, HML), which became the workhorse of empirical asset pricing for over a decade.

Momentum (Jegadeesh and Titman, 1993; Carhart, 1997) and profitability (Novy-Marx, 2013) followed, each documented as a robust anomaly in its own right.

The pace of factor discovery accelerated through the 2000s and 2010s. Harvey et al. (2016) catalogued 316 factors proposed in top finance and economics journals through roughly 2012, noting that 59 new factors were discovered between 2010 and 2012 alone. In his 2011 AFA presidential address, Cochrane (2011) gave the phenomenon its memorable name: "We also thought that the cross-section of expected returns came from the CAPM. Now we have a zoo of new factors."

2.2 Competing Factor Models

As the zoo grew, several research groups proposed parsimonious factor models to organize it. Table 1 summarizes the major contenders.

Table 1: Major Factor Models

Model	Factors	Year	Key Reference
CAPM	Market	1964	Sharpe (1964)
Fama–French 3	Market, SMB, HML	1993	Fama and French (1993)
Carhart 4	+Momentum (UMD)	1997	Carhart (1997)
Fama–French 5	+Profitability (RMW), Investment (CMA)	2015	Fama and French (2015)
q-Factor	Market, Size, I/A, ROE	2015	Hou et al. (2015)
Mispricing	Market, SMB, MGMT, PERF	2017	Stambaugh and Yuan (201

No single model dominates across all test assets, and spanning tests frequently reject each model's ability to price the others' factors. The q-factor model (Hou et al., 2015) is grounded in the investment CAPM and, according to its authors, largely subsumes the Fama–French five-factor model in spanning tests, though this claim remains contested. The mispricing factor model (Stambaugh and Yuan, 2017) takes an explicitly behavioral perspective. This unresolved competition underscores the difficulty of separating risk premia from mispricing with a small number of factors.

3 The Replication Crisis

The volume of proposed factors inevitably raised concerns about false discoveries. Several large-scale replication studies have attempted to quantify the problem, arriving at strikingly different conclusions.

3.1 Pessimistic Evidence

Hou et al. (2020) conducted the largest replication exercise at the time, testing 452 anomalies from the literature. Using value-weighted returns with NYSE breakpoints to mitigate the influence of microcap stocks, they found that 65% of anomalies fail to clear even the conventional single-test hurdle of $|t| = 1.96$. At the multiple-testing-adjusted threshold of $|t| = 2.78$, the failure rate rises to 82%. The trading frictions literature fared worst: 96% of 106 anomalies (102 of 106) failed to replicate. Their conclusion was stark: "Capital markets are more efficient than previously recognized."

McLean and Pontiff (2016) took a different approach, studying 97 published predictors before and after publication. They found that portfolio returns were 26% lower out of sample (before publication) and 58% lower after publication. The 26% decline provides an upper bound on the extent of data mining, while the additional 32% decline is attributed to investors learning about and trading away the anomalies.

3.2 Optimistic Evidence

Jensen et al. (2023) reached a more favorable verdict. Testing 153 characteristics across 93 countries, they found that the majority of factors can be replicated, that factors cluster into 13 economically meaningful themes, and that the evidence is strengthened rather than weakened by the large number of observed factors when evaluated through a Bayesian lens. Their comprehensive dataset is publicly available at jkpfactors.com, covering global factor returns.

Chen and Zimmermann (2022) similarly found that 98% of the 161 characteristics that were clearly significant in the original papers produce long-short portfolios with t-statistics above 1.96 when replicated faithfully. A regression of reproduced t-statistics on original t-statistics yields a slope of 0.88 and an R^2 of 82%, suggesting that while effect sizes may be somewhat overstated, the literature is more credible than the pessimistic assessments imply.

3.3 Reconciling the Views

The apparent contradiction between 35% and 98% replication rates reflects both methodological choices and a deeper conceptual disagreement about what "replication" means. At the methodological level, Hou et al. (2020) use NYSE breakpoints and value-weighted returns, which effectively remove microcap stocks that drive many anomalies but are difficult to trade in practice. Chen and Zimmermann (2022) replicate more closely to the original methodology of each paper, preserving microcap exposure. Jensen et al. (2023) use a Bayesian framework that accounts for the multiple testing problem differently from frequentist t-statistic thresholds.

But the disagreement also runs deeper. Hou et al. (2020) are testing whether factors survive in investable universes under standardized methodology, a form of scientific replication that asks whether the economic phenomenon is real and exploitable. Chen and Zimmermann (2022) are testing whether published results can be faithfully reproduced, a form of statistical replication that asks whether the original authors' computations were correct. These are different epistemological standards, and the gap between 35% and 98% partly reflects which standard each study applies. Linnainmaa and Roberts (2018) add a temporal dimension to this debate, showing that many accounting-based anomalies documented in post-1963 data do not survive in the pre-COMPUSTAT era (1926–1963), suggesting they may be artifacts of data mining rather than persistent economic phenomena.

Table 2 summarizes the key studies.

Table 2: Replication Studies Compared

Study	Factors	Rate	Threshold	Key Finding
Hou et al. (2020)	452	35%	$\|t\| > 1.96$	Microcap-mitigated; anomalies fail
McLean and Pontiff (2016)	97	n/a*	Return magnitude	26% OOS decline, post-publication decli
Chen and Zimmermann (2022)	161	98%	$\|t\| > 1.96$	Close-to-original replic
Jensen et al. (2023)	153	Majority	Bayesian	13 themes, 93 countrie

*McLean and Pontiff measure return magnitude decline, not statistical significance rates; their metric is not directly comparable to the other studies.

The reconciliation carries practical significance: factors that survive after

removing microcaps and applying stricter thresholds are more likely to be investable. The optimistic studies, by contrast, confirm that the academic literature is not fundamentally broken, even if many factors have limited practical relevance.

4 Methodological Foundations

4.1 The Multiple Testing Problem

When hundreds of hypotheses are tested, some will appear significant by chance. At the conventional 5% significance level, testing 400 independent hypotheses would produce approximately 20 "discoveries" even if none were real. Harvey et al. (2016) argued that the traditional t-statistic threshold of 1.96 is inadequate and proposed a minimum hurdle of $t > 3.0$.

Table 3 shows how multiple testing corrections raise the bar.

Table 3: Statistical Thresholds Under Multiple Testing

Method	Implied t-stat	Controls	Reference
Conventional	1.96	Single test	—
BHY (316 factors, 5%)	2.78	False discovery rate	Harvey et al. (2016)
BHY (est. total tests, 5%)	3.18	FDR + missing factors	Harvey et al. (2016)
Holm (316 factors)	3.64	Family-wise error rate	Harvey et al. (2016)
Bonferroni (316 factors)	3.78	Family-wise error rate	Harvey et al. (2016)
Practical minimum	3.0	Approximate FDR	Harvey et al. (2016)

Harvey and Liu (2020) extended this work with a double-bootstrap method that calibrates both false discovery (Type I) and missed discovery (Type II) errors simultaneously, highlighting the inherent trade-off: as the threshold rises, fewer false factors pass, but more genuine ones are missed. Harvey and Liu (2021) proposed a bootstrap framework that tests individual stocks directly rather than relying on portfolio sorts, providing a natural control for the multiple testing problem.

4.2 Publication Bias and Data Snooping

The multiple testing problem is compounded by publication bias. Journals preferentially publish significant results, leaving null findings in the "file

drawer." Harvey et al. (2016) estimated that 71% of all factors tried are missing from the published record, implying that the effective number of tests far exceeds the 316 published ones. Lo and MacKinlay (1990) were among the first to formalize data-snooping biases in asset pricing tests, showing that grouping stocks by a characteristic discovered in the same dataset inflates apparent predictability. White (2000) generalized this insight into a statistical framework for testing whether the best model found through specification search has genuine predictive superiority, or whether the result reflects data snooping.

Together, multiple testing and publication bias create an environment in which even honest researchers, making individually reasonable choices, can collectively generate a body of literature with a high false discovery rate.

5 Taming the Zoo

5.1 Shrinkage and Regularization

Rather than picking individual "winning" factors, several approaches compress the zoo into a low-dimensional structure. Kozak et al. (2020) construct a stochastic discount factor using Bayesian shrinkage, showing that a small number of principal components, not individual characteristics, explains the cross-section. Their key insight is that the quest for a sparse characteristics-based factor model is misguided: the pricing information is distributed across many correlated characteristics, and principal components capture this more efficiently than any small set of named factors.

Feng et al. (2020) take a different approach with double-selection LASSO, which evaluates new factors while controlling for omitted variable bias from the existing high-dimensional factor set. Across 135 factors, most new proposals are found to be redundant, though a few (notably profitability) have genuine marginal explanatory power.

5.2 Machine Learning Approaches

Machine learning methods offer a more flexible framework for extracting the latent factor structure. Kelly et al. (2019) introduced Instrumented Principal Component Analysis (IPCA), which allows for latent factors with time-varying loadings by using observable characteristics as instruments. Five

IPCA factors explain the cross-section significantly more accurately than existing factor models, and among a large collection of characteristics, only about ten are statistically significant at the 1% level.

Gu et al. (2020) conducted a comprehensive comparison of machine learning methods for return prediction, including LASSO, elastic net, random forests, gradient boosting, and neural networks. Trees and neural networks performed best, with gains traced to their ability to capture nonlinear interactions among predictors. All methods agreed on the dominant signals: variations of momentum, liquidity, and volatility.

Chen et al. (2024) extended this to deep neural networks with a no-arbitrage criterion function and an adversarial approach to constructing the most informative test assets. Their model explains 8% of total individual stock return variation (roughly twice the benchmark) and 23% of expected returns at the individual stock level.

5.3 How Many Factors Are Enough?

A practical question for both researchers and portfolio managers is the effective dimensionality of the factor space. Green et al. (2017) tested 94 characteristics simultaneously and found that only a handful provide independent information about average monthly returns once the others are controlled for. Swade et al. (2024) addressed the spanning question directly, starting from 153 published U.S. equity factors and finding that approximately 15 factors are sufficient to span the entire factor zoo. Common three-to-five factor models are insufficient, but the full zoo of 153 factors is highly redundant. Notably, the specific factor representatives that span the zoo vary over time, underscoring the importance of continuous factor evaluation rather than a fixed factor set.

Table 4 summarizes the dimension-reduction approaches.

6 Implications for Research and Practice

The factor zoo debate carries direct implications for how researchers evaluate new signals and how practitioners build portfolios.

For researchers, the literature suggests several guidelines. New factor discoveries should clear a t-statistic threshold of at least 3.0 (Harvey et al., 2016). Out-of-sample validation, especially in international markets, is essential for

Table 4: Dimensionality Reduction Approaches

Approach	Method	Effective Factors	Key Reference
PCA + shrinkage	Bayesian SDF	Few PCs	Kozak et al. (2020)
Double-selection LASSO	Penalized regression	Variable	Feng et al. (2020)
IPCA	Instrumented PCA	5 latent	Kelly et al. (2019)
Neural networks	Deep learning	Learned	Chen et al. (2024)
Spanning tests	Factor rotation	$\sim$15	Swade et al. (2024)

establishing robustness (Jensen et al., 2023). Economic motivation, not just statistical significance, should underpin any claimed factor: those grounded in theory (such as the investment CAPM behind the q-factor model) tend to survive replication better than purely empirical discoveries. Open replication databases such as those provided by Chen and Zimmermann (2022) and Jensen et al. (2023) make independent verification more practical than ever.

For practitioners, the key insight is that sparse factor models with three to five factors are insufficient to capture the cross-section of expected returns, but the full zoo is heavily redundant. Machine learning methods that extract the underlying latent structure, such as IPCA (Kelly et al., 2019) or deep factor models (Chen et al., 2024), offer a principled alternative to hand-picking factors, though they are not immune to overfitting and require careful out-of-sample validation. The factor zoo debate also shapes the signal-to-weights pipeline: the choice of which signals to trust is the prerequisite question before any portfolio construction decision.

7 Conclusion

The factor zoo grew from a handful of anomalies in the 1980s to over 400 proposed factors by the 2020s. The replication crisis debate has clarified that the literature is neither fundamentally broken nor entirely reliable: the truth depends heavily on methodological choices about microcap inclusion, t-statistic thresholds, and the treatment of multiple testing. Pessimistic estimates suggest that only 18–35% of anomalies survive strict scrutiny ($|t| > 2.78$ and $|t| > 1.96$, respectively) (Hou et al., 2020), while optimistic estimates place the replication rate at 98% when factors are tested faithfully against their

original methodology (Chen and Zimmermann, 2022).

The field has responded with increasingly sophisticated tools. Multiple testing adjustments raise the bar for new discoveries, shrinkage methods compress the cross-section into principal components, and machine learning extracts latent factor structures directly from data. The emerging consensus is that the effective dimensionality of the factor space lies somewhere around 10 to 15 factors, far fewer than the zoo's 400-plus inhabitants but more than the three to five of traditional models.

For the applied researcher and the quantitative portfolio manager alike, the factor zoo and the replication crisis are not merely academic curiosities. They determine which signals deserve a place in the portfolio construction pipeline and, ultimately, which risks are worth bearing.

References

Banz, R. W. (1981). The relationship between return and market value of common stocks. *Journal of Financial Economics*, 9(1):3–18.

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of Finance*, 52(1):57–82.

Chen, A. Y. and Zimmermann, T. (2022). Open source cross-sectional asset pricing. *Critical Finance Review*, 11(2):207–264.

Chen, L., Pelger, M., and Zhu, J. (2024). Deep learning in asset pricing. *Management Science*, 70(2):714–750.

Cochrane, J. H. (2011). Presidential address: Discount rates. *The Journal of Finance*, 66(4):1047–1108.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22.

Feng, G., Giglio, S., and Xiu, D. (2020). Taming the factor zoo: A test of new factors. *The Journal of Finance*, 75(3):1327–1370.

Green, J., Hand, J. R. M., and Zhang, X. F. (2017). The characteristics that provide independent information about average u.s. monthly stock returns. *The Review of Financial Studies*, 30(12):4389–4436.

Gu, S., Kelly, B. T., and Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273.

Harvey, C. R. and Liu, Y. (2020). False (and missed) discoveries in financial economics. *The Journal of Finance*, 75(5):2503–2553.

Harvey, C. R. and Liu, Y. (2021). Lucky factors. *Journal of Financial Economics*, 141(2):413–435.

Harvey, C. R., Liu, Y., and Zhu, H. (2016). ... and the cross-section of expected returns. *The Review of Financial Studies*, 29(1):5–68.

Hou, K., Xue, C., and Zhang, L. (2015). Digesting anomalies: An investment approach. *The Review of Financial Studies*, 28(3):650–705.

Hou, K., Xue, C., and Zhang, L. (2020). Replicating anomalies. *The Review of Financial Studies*, 33(5):2019–2133.

Jegadeesh, N. and Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, 48(1):65–91.

Jensen, T. I., Kelly, B. T., and Pedersen, L. H. (2023). Is there a replication crisis in finance? *The Journal of Finance*, 78(5):2465–2518.

Kelly, B. T., Pruitt, S., and Su, Y. (2019). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*, 134(3):501–524.

Kozak, S., Nagel, S., and Santosh, S. (2020). Shrinking the cross-section. *Journal of Financial Economics*, 135(2):271–292.

Linnainmaa, J. T. and Roberts, M. R. (2018). The history of the cross-section of stock returns. *The Review of Financial Studies*, 31(7):2606–2649.

Lo, A. W. and MacKinlay, A. C. (1990). Data-snooping biases in tests of financial asset pricing models. *The Review of Financial Studies*, 3(3):431–467.

McLean, R. D. and Pontiff, J. (2016). Does academic research destroy stock return predictability? *The Journal of Finance*, 71(1):5–32.

Novy-Marx, R. (2013). The other side of value: The gross profitability premium. *Journal of Financial Economics*, 108(1):1–28.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, 19(3):425–442.

Stambaugh, R. F. and Yuan, Y. (2017). Mispricing factors. *The Review of Financial Studies*, 30(4):1270–1315.

Swade, A., Hanauer, M. X., Lohre, H., and Blitz, D. (2024). Factor zoo (.zip). *The Journal of Portfolio Management*, 50(3):11–31.

White, H. (2000). A reality check for data snooping. *Econometrica*, 68(5):1097–1126.