A Census of the Factor Zoo*

Campbell R. Harvey
Duke University, Durham, NC 27708 USA
National Bureau of Economic Research, Cambridge, MA 02912 USA

Yan Liu
Purdue University, West Lafayette, IN 47906 USA

ABSTRACT

The rate of factor production in the academic research is out of control. We document over 400 factors published in top journals. Surely, many of them are false. We explore the incentives that lead to factor mining and explore reasons why many of the factors are simply lucky findings. The backtested results published in academic outlets are routinely cited to support commercial products. As a consequence, investors develop exaggerated expectations based on inflated backtested results and are then disappointed by the live trading experience. We provide a comprehensive census of factors published in top academic journals through January 2019. We also offer a link to a Google sheet that has detailed information on each factor, including citation information and download links. Finally, we propose a citizen science project that allows researchers to add to our database both published papers as well as working papers.

Keywords: Overfitting, Backtesting, Data mining, Multiple testing, Factor investing, Value investing, Momentum.

JEL: G10, G11, G12, G20, G23, G40

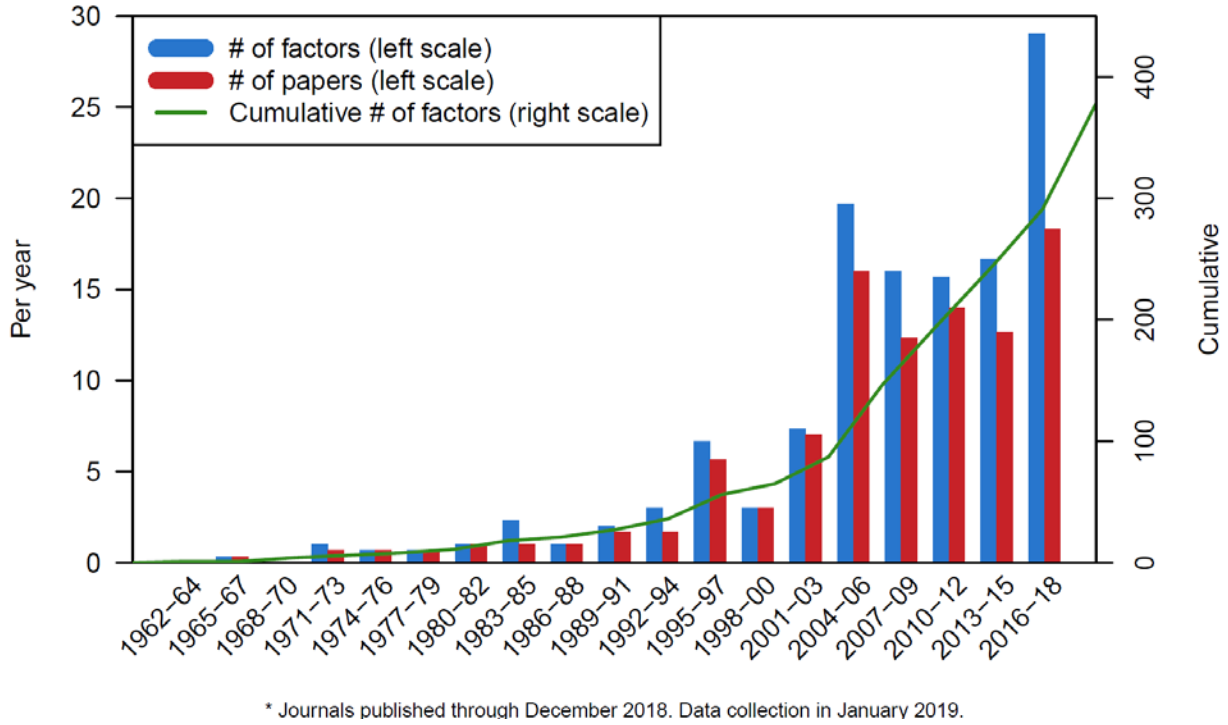
*Corresponding author: Campbell R. Harvey, cam.harvey@duke.edu. Initially posted to SSRN, February 25, 2019. We appreciate the research assistance of Mory Elsaify and the editorial assistance of Kay Jaitly.

Introduction

The "factor zoo" seems an appropriate metaphor for the growing number of investment factors proposed by both academics and practitioners. An initial census of the zoo was carried out by Harvey, Liu, and Zhu (2016), who detail over 300 factors published in top academic journals based on data through 2012. They also show that almost all of the past research fails to take into account the multiple testing problem: with so many factors tried, some will appear "significant" purely by chance.

The purpose of our paper is threefold. First, we provide an update to Harvey, Liu, and Zhu (2016) and detail 382 factors published in top academic journals, as shown in **Figure 1**. Second, we provide a link to a Google sheet that lists each of the factors as well as links and citation information. Third, we propose a citizen science project in which researchers can offer additional factors (from both published papers and working papers) which we will vet and add to the Google sheet.

Figure 1. Out of control factor production - through January 2019



* Journals published through December 2018. Data collection in January 2019.

The Issues

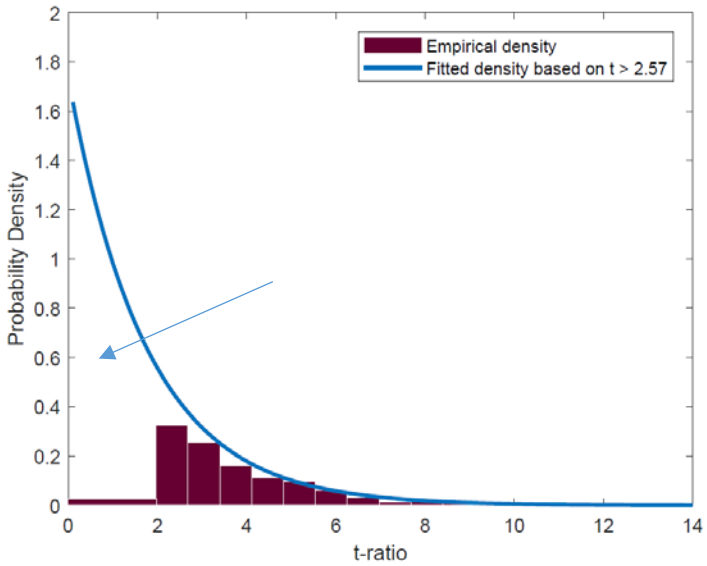
Harvey (2017) details the following dilemma. Academic journals overwhelmingly publish papers with positive results that support the hypothesis being tested. Papers with positive results tend to be cited more than papers with negative results. Journal quality is often proxied by impact factors, which measure the number of citations the papers published by the journal get. Journal editors want higher impact factors. Authors figure this out. To maximize the chance a paper is published, the paper needs a positive result. Hence, the data mining begins.

Consequently, a number of problems can arise. First, as researchers, we must take multiple testing into account when assessing statistical significance. We usually focus on an acceptable rate of false positives (e.g., a 5% level) and will often declare a factor significant at the 5% level (e.g., two standard errors from zero, aka *two sigma*). This works for a single try, but if we test for instance, 20 different factors, one will likely be two sigma—purely by chance. If we accept this factor as a true factor, the error rate will not be 5%, but closer to 60%. Hence, it is crucial to impose a higher hurdle for declaring a factor significant. Two sigma is far too weak a threshold and leads to an unacceptably large number of false positives.

Second, while it makes sense that we should increase the threshold reflecting the number of tests, what is this number? Whereas we can count the factors published in academic journals, we cannot count the factors that do not make it into journals. Harvey (2017) describes the "file drawer" effect. In this case, a researcher invests time in a project and realizes the factor will not exceed the usual significance levels. While a possibility exists that the academic researcher could publish the negative result, the researcher realizes that the paper will generate very little interest in terms of citations. The researcher is faced with the following dilemma. Should she expend the effort to finish the project and invest her time in the peer review process at potentially multiple journals, or walk away from the project (i.e., put it in the file drawer)? Perhaps, most importantly, working on a project with a negative result is not "fun." So aside from the decreased probability of publishing a paper with negative results, researchers often file away papers they are not excited about.

The file drawer effect means that we are not aware of all the factors that have been tested This is evident from the severe truncation of the distribution of *t*-statistics presented by Harvey, Liu, and Zhu (2016). In **Figure 2**, we observe that the middle and left-hand side of the distribution is missing. Therefore, any correction for the number of factors tested should be viewed as conservative.

Figure 2. Truncated distribution of *t*-statistics for factor studies, 1963–2018



Third, many multiple testing corrections are suggested in the literature. The simplest is the Bonferroni correction. Suppose we try 50 factors and find that one is approximately three sigma with a *p*-value of 0.01. Three sigma is impressive under a single test (usually we look for a *p*-value < 0.05)—but we did 50 tests. The Bonferroni correction simply multiplies the *p*-value by the number of tests. So the Bonferroni-adjusted *p*-value is 0.50, which is much larger than our usual 0.05. **Figure 3** shows three corrections detailed by Harvey, Liu, and Zhu (2016), including the Bonferroni (in blue).

Figure 3. Historical testing thresholds controlling for cumulative factor discovery, 1963–2018

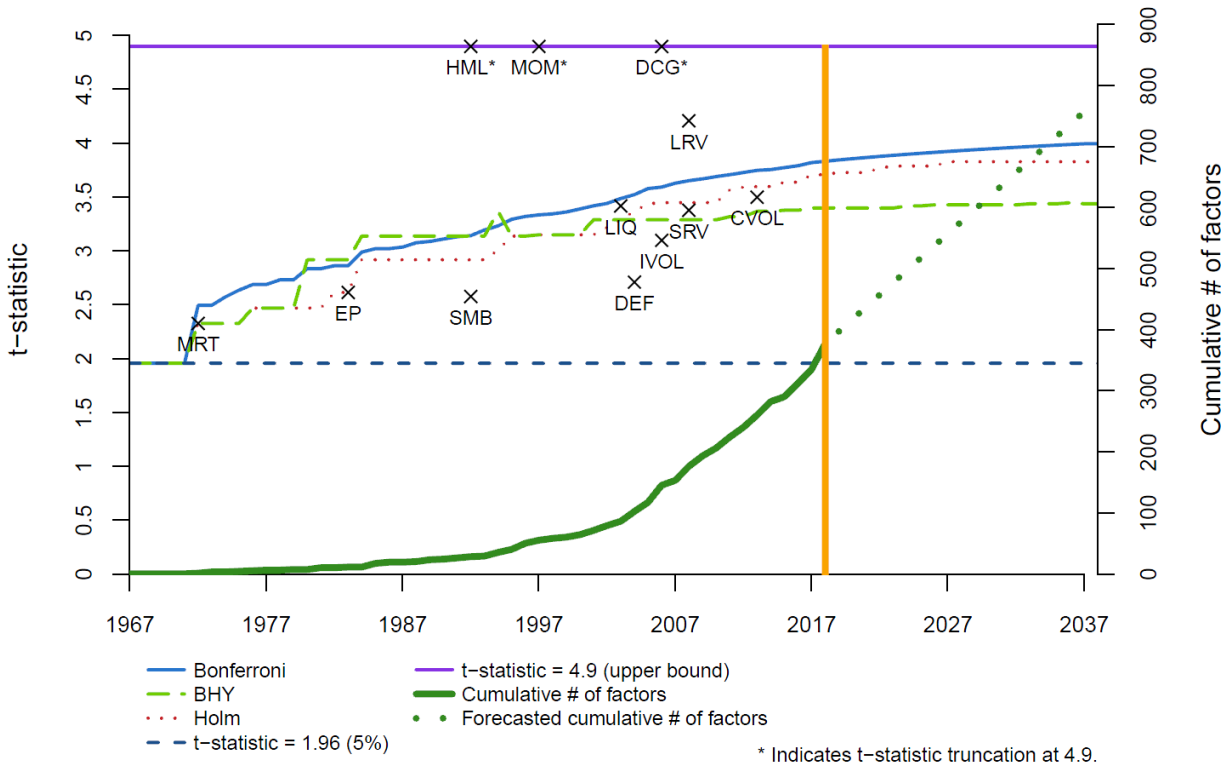


Figure 3 shows the historical number of factors discovered (in green, right axis). The blue dashed line is the two sigma (*t*-statistic = 2.0), which is common for establishing significance for a single test. In the figure, we present three approaches to multiple testing correction, including the Bonferroni. Each starts at 2.0 in the 1960s, because only a single test was available at this time. As more factors were discovered, the threshold increases.

The figure also shows some of the popular factor discoveries. Interestingly, the Fama and French three-factor model was published in 1992. Given the factors discovered at that time, the value factor, or HML, comfortably exceeds the threshold for each multiple testing correction. The size factor, or SMB, has a different fate, however. It falls short of every hurdle. This suggests that the three-factor model might have been a two-factor model if multiple testing corrections were taken into account. Figure 3 also extrapolates the factor production through 2037 as well as the multiple testing thresholds.

Using a two-sigma cutoff obviously leads to an error rate that is too high (also see Harvey and Liu, 2018a). Each of the multiple testing corrections increases the number of sigmas and therefore reduces the error rate. This is good, but leads to two related problems. Suppose our goal is to be in the 5% range for false discoveries. With no correction, we could easily be in the 60% range. With correction, what error rate should we expect? Is it 30%, which is far better than 60%, but still unacceptable given that our goal is 5%? Is it 0.1%, which is below our target and may be unacceptable for another important reason?

Why would a 0.1% error rate be unacceptable? This threshold greatly limits our false positives—which is good. But the situation is more complicated. With such a stringent threshold, we will discard many true factors. That is, we will miss many discoveries. Think of it this way. We could reject every single factor. As a result, there will be no false positives, because there are no positives. In doing so, however, we will also miss factors that are true sources of a risk premium. This is exactly the issue Harvey and Liu (2018b) address in developing a data-driven approach to balance the false discoveries and the missed discoveries.

A fourth concern is even more difficult to deal with. Even if we know exactly how many factors have been tried, it is naïve to apply the same adjustment to each factor. Here is an example. Suppose Researcher A develops an economic model from first principles. An implication of the model is that a particular formulation of a factor should impact the cross-section of expected returns. Such a formation has not been tested before. Researcher A goes to the data and tests this particular factor formation and finds that it is "significant." Researcher B has a much different strategy. This researcher uses no theory or economic foundation. Researcher B is trained in data science rather than in economics. He tests various combinations and permutations of CRSP and Computstat data and "discovers" a new factor. Should we treat the discovery from Researcher A (first principles) and Researcher B (data mined) identically? We think not. We believe, as Harvey (2017) advocates, injecting some prior beliefs into the decision making is critically necessary.

Even after addressing all of these issues, a fifth remains. It is routine in academic publications to ignore transaction costs when reporting factor premia. This is a serious mistake, particularly for high-turnover factors, such as cross-sectional momentum or factors that require shorting of small- and micro-cap securities (see, e.g., Asness and Frazzini, 2013, and Novy-Marx and Velikov, 2016). Allowing for reasonable transaction costs has a similar effect as increasing the threshold for significance: far fewer factors seem real.

The Factor Census

For any census, certain choices need to be made for classification. How we define factor is one of these choices. We take a very broad view consistent with the literature. We have two main classifications. The first is "common" and the second is "characteristics." The common factor definition is the traditional definition; for example, a common factor is the market return. Each asset has an exposure to the common factor; in the case of market return, the exposure is beta. If the factor is useful, low-exposure assets have low expected

returns, and high-exposure firms have high expected returns; that is, the exposures are able to explain the cross-section of expected returns.

Another class of studies looks at individual firm characteristics to see if they explain the cross-section of expected returns or if portfolios of these characteristics—such as going long assets with a high value of the characteristic and short assets with a low value of characteristic—produce significant risk-adjusted returns.

We also drill down into each of the general categories. In the common factor category we present six subcategories (examples): 1) financial (market return); 2) macro (unexpected inflation); 3) microstructure (market liquidity); 4) behavioral (market sentiment); 5) accounting (HML); and 6) other.

The characteristics category has only five subcategories because the macro category does not make sense. The subcategories (examples) are: 1) financial (idiosyncratic volatility); 2) microstructure (asset transaction cost); 3) behavioral (asset media coverage); 4) accounting (asset price-to-earnings ratios); and 5) other.

The Google Sheet

The title of the main Google sheet is "Sorted by Year." This sheet contains the columns listed below. Importantly, we include both theory and empirical papers. In addition, we list the factors that each paper tests, although some of the factors may not be unique and the non-unique factors are not counted.

- Column A is the year the paper was published.
- Columns B and C are counters for the common factors and individual factors.
- Column D is the short-form name of the factor (e.g., idiosyncratic volatility).
- Column E gives brief details on the formation of the factor. Theory papers have their own flag.
- Column F gives both the category and subcategories (e.g., common, financial).
- Column G names the journal the paper is published in. Our census (with some exceptions) mainly focuses on the top five general purpose economics journals, the top three finance journals, the next two ranked finance journals, and the top three accounting journals.
- Column H details the short references for in-text citations (e.g., Sharpe, 1964).
- Column I lists the full reference.
- Column J provides a hyperlink to the published version of the paper.
- Column K provides the *t*-statistic.
- Column L details the sample.
- Column M allows for an alternative *t*-statistic, because different testing methods can be used.
- Column N provides space for notes.

Four additional Google sheets are also included:

- "Common Sorted by Year" provides common factors sorted by year.
- "Individual Sorted by Year" provides individual characteristic factors sorted by year.
- "Working Papers" are factors in papers that are not yet published.

- "T-Statistics" is the summary of the *t*-statistics from all of the papers.

The link to the Google sheet[1] with the Factor Census is https://tinyurl.com/y23ozzkc

The link to the updating form[2] is http://tinyurl.com/y2xq65tt. The updating form is the citizen science project. We ask for the same information that is included in the main Google sheet. We welcome your own papers or papers by others that we might have missed. Working papers or published papers are welcome.

Please participate in this citizen science project.

REFERENCES

Asness, Clifford, and Andrea Frazzini. 2013. "The Devil in HML's Details." *Journal of Portfolio Management*, vol. 39, no. 4 (Summer):49–68.

Harvey, Campbell R. 2017. "Presidential Address: The Scientific Outlook in Financial Economics." *Journal of Finance,* vol. 72, no. 4:1399–1440.

Harvey, Campbell R., and Yan Liu. 2018a. "Lucky Factors." *Journal of Financial Economics*, forthcoming. https://ssrn.com/abstract=2528780

Harvey, Campbell R., and Yan Liu. 2018b. "False (and Missed) Discoveries in Financial Economics." *Journal of Finance,* forthcoming. https://ssrn.com/abstract=3073799.

Harvey, Campbell, Yan Liu, and Heqing Zhu. 2016. ". . . and the Cross-Section of Expected Returns." *Review of Financial Studies*, vol. 29, no. 1 (January):5–68.

Novy-Marx, Robert, and Mihail Velikov. 2016. "A Taxonomy of Anomalies and Their Trading Costs." *Review of Financial Studies*, vol. 29, no. 1:104–147.

[1] Full link is https://docs.google.com/spreadsheets/d/1mws1bU56ZAc8aK7Dvz696LknM0Vp4Rojc3n61q2-keY/edit?usp=sharing

[2] Full link is https://docs.google.com/forms/d/e/1FAIpQLSflDTiqr2bFsP5tqW7XStIp0-ikwHqK94dbCoMshLfesvKL2g/viewform?vc=0&c=0&w=1