LUCK VERSUS SKILL AND FACTOR SELECTION

. . .

Campbell R. Harvey and Yan Liu

INTRODUCTION

In the universe of thousands of mutual funds, a substantial number will outperform their benchmarks purely by luck. Fama and French (2010) develop an innovative approach that presents substantial progress on the economically important problem of distinguishing luck from skill in performance evaluation. Their methods can also be applied to distinguishing luck from performance in other important areas of finance. This article represents one such application to the reliability of return-predicting factors or characteristics.

LUCK

Fama and French address a classic problem in statistics called multiple testing. When many tests are conducted, some will appear "significant" by luck, using statistics designed to measure the chance of one, and only one, test coming out well.

Suppose every one of the 3,000 mutual funds has a true alpha of zero. Under this assumption, the *t*-statistics of the estimated alphas approximately follow standard normal distributions. Therefore, by pure chance, the maximum alpha *t*-statistic will be 3.6 on average, and 1% of the funds should show alpha *t*-statistics greater than 2.3. Though an individual fund only has a 1% chance of showing a 2.3 *t*-statistic, and therefore declaring its return "statistically significant," it is nearly certain that 30 funds with 2.3 *t*-statistics will exist just due to luck. It would clearly be a mistake to proclaim that finding such funds proves the presence of skill or market inefficiency.

In principle, multiple testing is easily handled. If the investigator is willing to write down the multiple testing procedure—choose the best of 100 mutual funds, say—then one can easily work out the probability distribution of the multiple test—the distribution of the greatest alpha *t*-statistic in 100 funds. The trouble is that the investigator is seldom so explicit about the multiple testing process. The profession as a whole—the tendency for only "significant" results to be published, for example—is even less explicit. Statistics which correct for multiple testing are therefore difficult to calculate. Thus, empirical work rarely corrects explicitly for multiple testing. The fact that published finance empirical work often fails out of sample is not surprising.

PRECURSORS

Kosowski, Timmermann, Wermers, and White (2006) use a bootstrap technique to study the distribution of alpha *t*-statistics under the null that funds have no true alpha. They deserve credit for advancing the "how many funds should we see outperforming" test rather than the more conventional tests for skill that look for persistence in performance or alphas in portfolios of funds sorted on the basis of some ex ante measure of skill. Their technique allows us to make statements about whether mutual funds exhibit skill in general, without the researcher needing to specify how to find the good funds. It therefore is less useful in evaluating the skill of a particular fund manager or helping investors in their quest to find the good funds.

They find that mutual fund "stars" are still stars after their adjustment: there are more successful funds than there should be, just due to chance.

The Kosowski et al. (2006) approach allows us to see if the best funds, with large individual *t*-statistics, do better than they should due to chance. Kosowski et al. thus generalize White (2000) in that White looks at the distribution of the maximum test statistics while Kosowski et al. calculate the whole distribution.

FAMA AND FRENCH

However, Kosowski et al. independently bootstrap each fund's returns. This method ignores potential correlation between the return residuals. If, beyond factor exposures, fund A and fund B both take the same "idiosyncratic" bets, then it is less likely that the better of the two has a large alpha *t*-statistic than would be the case if their residuals were uncorrelated. In the extreme case that all residuals are perfectly correlated, we essentially have only one fund and there is no need for multiple testing adjustment at all.

If the average correlation among the residuals is positive, then it seems that the multiple testing adjusted threshold should be lower than what Kosowski et al. suggest. However, funds often take opposite (idiosyncratic) bets against each other, and this tendency varies over time. In that case, there will be more high *t*-statistic funds due to chance than the Kosowski et al. approach calculates.

Fama and French solve this problem by bootstrapping the residuals across all funds, which is a key innovation. For example, suppose we have a 60 × 3,000 panel of fund returns in which rows are dates (months) and columns are funds. Fama and French resample an entire row. Resampling by row accommodates arbitrary cross-correlation of the residual risks. Fama's insight on this issue reaches back to the Fama-MacBeth procedure, which also allows arbitrary cross-sectional correlation. (One can block-bootstrap adjacent time periods as well, to maintain temporal correlations or include lags in the underlying regressions. However, time-series autocorrelations of fund returns are small, so this modification likely has little practical effect.)

Using this bootstrap and a four-factor performance attribution model, Fama and French find, among other things, that the top 1% fund's empirical alpha *t*-statistic of 2.5 is only very slightly above the first percentile of "luck" alphas (simulated p-value = 36.96%). Therefore, even for the top funds, there is little to no significant outperformance.

BUILDING ON THESE INSIGHTS

Fama and French and Kosowski et al. do not provide an overall test statistic. What is the probability of seeing the entire distribution of observed alphas, or one by some metric more extreme, if all the underlying alphas are zero? It's fun to look at the upper 1%, or upper 5%, but an arbitrariness lies in that choice. To produce an overall statistic, we would need to understand the joint properties of the values at each point of the distribution. What is the joint distribution of, say, the number of funds that exceed *t*-statistics of 2.0 and the number that exceed *t*-statistics of 3.0? (Such a test would likely reject the null that all alphas are zero, but because of the puzzlingly large number of funds with negative alphas, not positive alphas!)

Both Fama and French and Kosowski et al. also assume the joint null that none of the funds has skill. They do not tell us what would happen if a portion of these funds actually do have skill. Fama and French add a table of "injected alpha" predictions—they assume that true alphas are random draws from a normal distribution with a zero mean and find the variance estimate that best fits the realized alphas. But they do not reverse engineer the distribution of injected alpha that best matches the data, nor offer measures of uncertainty about this distribution. Reverse engineering the implied distribution of injected alpha is a straightforward calculation, and we encourage researchers to do it even though Fama and French did not.

As a result, Fama and French do not tell us just how many funds have skill, and how much they have, or whether a particular fund has significant skill or not. These concerns inspired a sequence of follow-up papers. To study the joint properties of the tests, recent work draws on the statistics literature that proposes the false discovery rate, which is defined as the expected number of false discoveries among all discoveries. In particular, Barras, Scaillet, and Wermers (2010) use the false discovery rate to correct for the bias in estimating the proportion of skilled funds. Harvey, Liu, and Zhu (2016) use false discovery rate as the multiple testing counterpart of the Type I error and provide multiple testing adjusted benchmark *t*-ratios. To evaluate the performance of tests under alternative hypotheses, Barras, Scaillet, and Wermers group the universe of mutual funds into three and use the distributional information in the *t*-statistics to estimate the proportion of false discoveries. Ferson and Chen (2014) refine their method by removing the perfect power assumption (i.e., the test will never confuse a good fund with a fund that has a zero or negative mean return) as well as incorporating the possibility that a bad fund (i.e., a fund that has a negative mean return) can be falsely identified as a good fund (i.e., a fund that has a positive mean return), and vice versa. Harvey, Liu, and Zhu parametrically model the distribution of mean returns for true discoveries in a multiple testing framework. By explicitly modeling the distribution of the test statistics under both the null and the alternative hypotheses, they are able to make inferences on the fraction of true discoveries.

APPLICATION TO OTHER AREAS OF FINANCE

Test multiplicity haunts finance research. For the fund evaluation literature, it is termed *luck versus skill*. For the return predictability and risk factor literature, it is called *data snooping* or *data mining*: some variables will appear to predict returns, and some factors (portfolios of stocks) will appear to have large average returns, based on one-at-a-time test statistics, if one looks at many candidates and chooses the best ones.

The fact that we do not observe the tests that are tried but failed is an important source of multiple-test bias. Harvey, Liu, and Zhu (2016) take this bias into account by explicitly modeling the missing data process. They assume that researchers drop all insignificant tests—tests that have *t*-statistics below a certain threshold. Following this assumption, observed *t*-statistics follow a truncated distribution, and they back out the underlying number of unreported tests. They estimate that more than four times the number of observed tests are missing (i.e., unpublished) for the academic literature on the cross-section of expected returns. That is, we observe only one in five tests.

Building on Fama and French's insight, however, we must also adjust test statistics for the fact that all variables share economy-wide shocks in fundamentals or liquidity conditions.

The return prediction literature faces an additional problem: the left-hand side variables are the same while researchers try multiple right-hand side variables. Time-series research tries to forecast the same, usually market return, with many right-hand side variables. In cross-sectional estimates, either in the form of Fama-MacBeth regressions, Gibbons-Ross-Shanken panel regressions, or portfolios formed on the basis of predictor variables, the cross-section of stock returns is the same left-hand side variable as one tries hundreds of forecasting variables.

FAMA-FRENCH APPLIED TO FACTOR SELECTION

Fama and French's technique can help us overcome these deep problems of empirical asset pricing. To illustrate, we apply Fama and French's technique to address multiple comparison problems in factor selection. Factor returns are long-short zero-cost investment strategies based on return predicting variables. Fama and French's "high minus low" (HML)—long (buy) value stocks with high book-to-market ratios, and short growth stocks with low book-to-market ratios—is the classic example. There is not much of a jump between evaluating the viability of a long-short investment strategy and evaluating the viability of a particular investment manager.

We study the Standard and Poor's (S&P) Capital IQ database of "alphas." S&P sells data on the historical performance of synthetic long-short strategies. We focus on 484 strategies for the US equity market from 1985 to 2014. These strategies are cataloged into eight groups based on the types of risks that they are exposed to (e.g., market risk) or the nature of the forecasting variables (e.g., characteristics). The database has a good coverage of well-known return signals, including CAPM beta (Capital Efficiency Group), size (Size Group), value (Valuation Group), and momentum (Momentum Group).

Following Fama and French, we first calculate the *t*-statistics for the average return (raw returns without factor adjustment) of each long-short strategy, and we rank these *t*-ratios. We then demean the long-short portfolio returns and bootstrap their return observations. In bootstrapping, we keep the entire cross-section intact for each resampled time period, again following Fama and French.

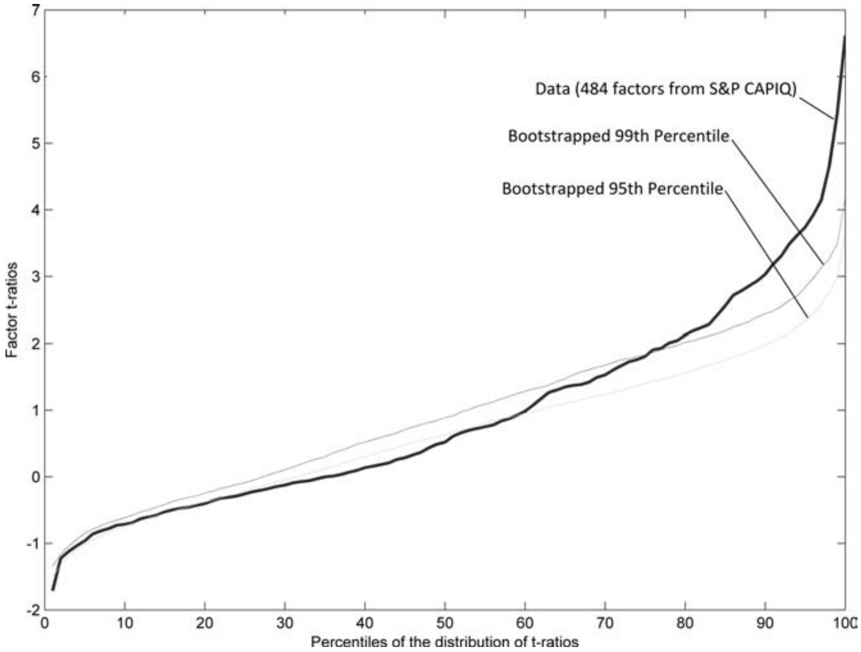

Figure 1. Factor tests for Capital IQ data

We obtain 484 (pre-transaction costs) long-short strategy returns from S&P Capital IQ. The curve labeled "Data" shows the *t*-statistic percentiles for these strategies. After demeaning each of the 484 strategies, the other two curves show the 99th and the 95th percentiles of the bootstrapped *t*-statistic percentiles, respectively.

Figure 1 presents the results. The line labeled "Data" shows the empirical *t*-ratio percentiles for the Capital IQ strategies. The x-axis presents the percentile of the 484 strategies in their rank by average return *t*-statistic. The y-axis shows the value of the alpha *t*-statistic for each fund. The dashed lines show the 99th and 95th percentiles in the bootstrapped percentiles of the no-mean-return distribution.

The top percentiles of the data are generally well above their 99th and 95th percentile bounds. In particular, the maximum *t*-ratio for the data is 6.62 and the 99th percentile for the simulated maximum *t*-ratios is 4.13; the 90th percentile for the data is 3.11 and the 99th percentile for the simulated 90th percentiles is 2.45. Based on the graph, we have confidence to reject the null hypothesis that none of the 484 factors has a positive expected return.

These results for the S&P Capital IQ factors contrast with Fama and French's mutual fund results. First, Fama and French find scant evidence that mutual funds outperform their four-factor benchmark. Figure 1 suggests that a number of the Capital IQ factors appear to outperform—that is, generate a mean return that is significantly positive. This is perhaps not surprising given that there is considerable evidence that some of these strategies (such as the equity premium and the return to value-oriented investing) generate positive returns. However, the bulk of the Capital IQ returns are less widely known long-short strategies. Second, Fama and French find significant underperformance (negative alphas) for the worst performing fund managers. There is no evidence of significant underperformance in the Capital IQ data.

However, our illustrative example makes no effort to overcome survivorship and selection bias. S&P may have culled significant underperformers and included in-sample successful strategies (backfill bias). Our point is to demonstrate the widespread usefulness of Fama and French's statistical method. Their research also includes meticulous data work to include all the dead funds and overcome survivor bias, selection bias, backfill bias, and so forth.

The Fama-French method does not tell us which of the factors have positive expected returns. For example, the 80th percentile of *t*-statistics seems to lie above the 99th percentile confidence band. We should not conclude that each strategy in the upper 20% is significant.

We also do not observe all the factors that have been tried. Although the IQ sample covers almost 500 factors, thousands more could have been tried but not reported. Harvey, Liu, and Zhu (2016) provide some tools to deal with this problem. They provide a truncated distributional framework to back out the number of missing predictor variables.

A CUTOFF

A simple tool for evaluating the effects of multiple testing, and that allows us to consider the significance of a given fund or strategy return, would be very useful. To that end, we now find a *t*-ratio cutoff that classifies a statistically significant factor, correcting for multiple testing.

Consider the mutual fund evaluation problem. In every bootstrap iteration, we calculate *t*-ratios for, say, 3,000 funds. We save the maximum *t*-ratio. We iterate 10,000 times and get the distribution of the maximum *t*-ratio. We then look at the max *t*-ratio among the fund managers. If the *t*-ratio of the best-performing fund manager is larger than the 95th percentile of the max of the bootstrapped distribution (under the null of no skill), we declare the manager "skilled."

The maximum may be sensitive to outliers. Tail percentiles may be better. However, there is a tradeoff. The max has more power to reject the null while percentiles are less affected by extreme observations. Here we use max as an illustration.

Next, for the manager declared skilled, we alter the null distribution to include that manager's alpha. Hence, 2,999 funds have a zero mean return and one fund will have a positive mean. We repeat the bootstrap. We compare the distribution of the second highest *t*-ratio to the *t*-ratio of the second best performing manager. If the manager is again better than the 95th percentile of the null, we declare that manager to be skilled, and insert her alpha into the null and continue. If the manager does not exceed the 95th percentile, we stop.

Essentially, at each step of our method, our null hypothesis is that some managers have skill, their levels of skill are set at the in-sample estimates, and the rest of the managers have a skill of zero. We use the *t*-statistic of the best performing manager among those that have not been identified as skillful to test this null. The first time the null is rejected, we record the *t*-statistic. This is the *t*-statistic threshold that our method identifies. Managers that have a *t*-statistic above this threshold are declared to have skill.

We can apply the same idea to factors. We find the fraction of statistically significant factors by sequentially injecting the in-sample means of the factors we declare as significant. In particular, we add back the means of the top p percent of factors. We then bootstrap to generate the distribution of the cross-section of *t*-ratios. In essence, this distribution is based on the hypothesis that the top p percent of factors are significant, their means equal their in-sample fit, and the remaining $1 - p$ percent of factors have a mean of zero. To test this hypothesis, we compare the $(1 - p)$th percentile of the bootstrapped distribution with the $(1 - p)$th percentile of *t*-ratios for the real data. We sequentially increase the level of p until the null hypothesis is not rejected. Our estimate of the fraction of significant factors will be p.

Figure 2 presents the results. When we inject the top 5% of factors into the null ($q = 5.0$, top panel), the 95th percentile of *t*-ratios of the real data is above the 95th percentile of the simulated 95th percentiles. Hence, inserting the means of the top 5% of factors into the null is not enough to explain the observed 95th percentile of *t*-ratios. Hence, we would declare the top 5% as "significant." We gradually increase q. When $q = 13.0$ (i.e., bottom panel), the 87th percentile of *t*-ratios of the real data meets the 95th percentile of the simulated 87th percentiles. The corresponding *t*-ratio cutoff is 2.71. We therefore conclude that 63 ($= 484 \times 13\%$) out of the 484 factors are "significant."

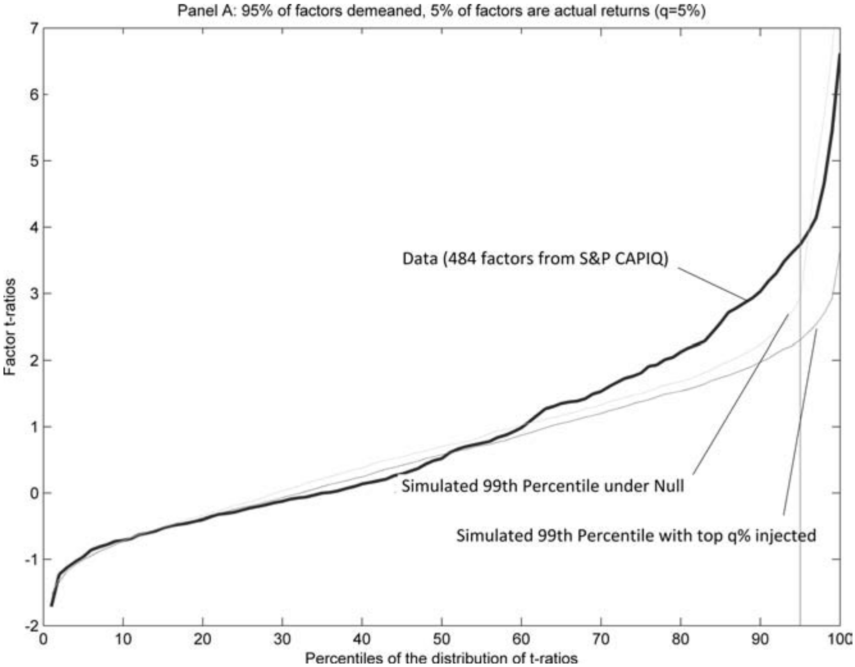

Figure 2. Estimating the fraction of significant factors

We obtain 484 (pre-transaction costs) long-short strategy returns from S&P Capital IQ. The curves labeled "Data" in both panels show the *t*-statistic percentiles for these strategies. After demeaning each of the 484 strategies, a curve in each panel shows the 99th percentiles of the bootstrapped *t*-statistic percentiles. In Panel A, we demean the bottom 95% of strategies (and keep the top 5% intact), and the third curve in the top panel shows the 99th percentiles of the bootstrapped *t*-statistic percentiles of this new sample, which we refer to as "simulated 99th percentile with top 5% injected." In panel B, we demean the bottom 87% of strategies (and keep the top 13% intact), and the third curve in the bottom panel shows the 99th percentiles of the bootstrapped *t*-statistic percentiles.

Again, this analysis is contingent on the integrity of the data provided by S&P Capital IQ. In addition, transactions costs are not included. Given the prominence of a few existing factors (e.g., market, size, value, and momentum), sometimes we are interested in the incremental alpha of a strategy after adjusting for prominent factors. But our method is easily extended to accommodate this.

Inspired by this example, Harvey and Liu (2015) propose a general method to estimate regression models in the presence of multiple testing. They first

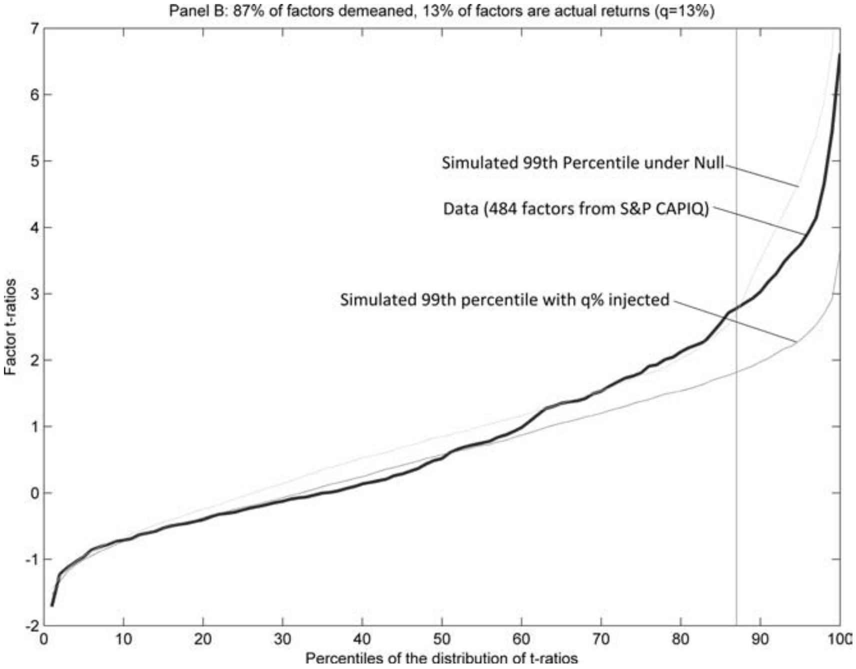

Figure 2. (Continued)

orthogonalize the right-hand variables so that they have no in-sample explanatory power. This is similar to Fama and French. They then bootstrap to test whether a "true" variable exists among a candidate set of right-hand side variables. To control for multiple testing, they use the max statistic, similar to White (2000). The final step of their method is to recursively identify the set of significant right-hand side variables, one variable at a time. Their framework is general enough to encompass most regressions models in asset pricing applications. In particular, they show how their method applies to predictive regressions, Gibbons-Ross-Shanken panel regressions, and Fama-MacBeth regressions. Harvey and Liu (2016) probe the multiple testing issue from a different angle. They propose a structural approach that allows us to pool information from the cross-section to adjust the inference on a particular entity within the cross-section.

CONCLUSION

Fama and French provide an innovative approach to distinguish between luck and skill in investment manager performance. We show that their insights touch a wide swath of financial research. In empirical research in finance, we often have many candidate variables or a large number of economic agents/units that might provide an answer to an important economic question. Two issues hinder such investigations. First, inference is often incorrectly drawn from a statistic designed for independent rather than multiple testing. Second, financial variables are correlated in complicated ways. The large number of potential return-forecasting variables makes the accurate modeling of the correlation structure among these variables almost impossible. Fama and French deal with both of these hurdles. While they focus their application on investment manager evaluation, their insights apply to a broader spectrum of research in financial economics.

***

The authors appreciate the detailed comments of John Cochrane and Tobias Moskowitz.

REFERENCES

Barras, L., O. Scaillet, and R. Wermers. 2010. "False Discoveries in Mutual Fund Performance: Measuring Luck in Estimated Alphas." *Journal of Finance* 65, 179–216.

Fama, E. F., and J. D. MacBeth. 1973. "Risk, Return, and Equilibrium: Empirical Tests." *Journal of Political Economy* 81, 607–636.

Fama, E. F., and K. R. French. 2010. "Luck versus Skill in the Cross-Section of Mutual Fund Returns." *Journal of Finance* 65, 1915–1947.

Ferson, W. E., and Y. Chen. 2014. "How Many Good and Bad Fund Managers Are There, Really?" Working paper, University of Southern California.

Gibbons, M. R., S. A. Ross, and J. Shanken. 1989. "A Test of the Efficiency of a Given Portfolio." *Econometrica* 57, 1121–1152.

Harvey, C. R., Y. Liu, and H. Zhu. 2016. " . . . And the Cross-Section of Expected Returns." *Review of Financial Studies* 29, 5–68.

Harvey, C. R., and Y. Liu. 2015. "Lucky Factors." Working paper, Duke University.

Harvey, C. R., and Y. Liu. 2016. "Rethinking Performance Evaluation." Working paper, Duke University.

Kosowski, R., A. Timmermann, R. Wermers, and H. White. 2006. "Can Mutual Fund 'Stars' Really Pick Stocks? New Evidence from a Bootstrap Analysis." *Journal of Finance* 61, 2551–2595.

White, H. 2000. "A Reality Check for Data Snooping." *Econometrica* 68, 1097–1126.