False (and Missed) Discoveries in Financial Economics

Campbell R. Harvey

Duke University, Durham, NC 27708 USA
National Bureau of Economic Research, Cambridge, MA 02138 USA

Yan Liu*

Texas A&M University, College Station, TX 77843 USA

Current version: August 14, 2019

Abstract

The multiple testing problem plagues many important issues in finance such as fund and factor selection. Many look good purely by luck. There are a number of statistical techniques to control for multiplicity that reduce Type I errors — but it is unknown by how much. We propose a new way to calibrate both Type I and Type II errors. We start with the researcher's prior belief on the proportion of managers that are skilled. Using a double bootstrap method, we then establish a t-statistic hurdle that is associated with a specific false discovery rate (e.g., 5%). We also establish a t-statistic hurdle that is associated with a certain acceptable ratio of misses to false discoveries (Type II error scaled by Type I error) — effectively allowing for differential costs of the two types of mistakes. Evaluating current methods within our framework, we find that they lack the power to detect outperforming managers.

Keywords: Type I, Type II, Multiple testing, False discoveries, Odds ratio, Power, Mutual funds, Anomalies, Bayesian, Factors, Backtesting, Factor Zoo

* Current Version: August 14, 2019. First posted on SSRN: November 19, 2017. Send correspondence to: Yan Liu, Mays Business School, Texas A&M University, College Station, TX 77843. Phone: +1 919.428.1118, E-mail: y-liu@mays.tamu.edu. We appreciate the comments of Stefan Nagel, two anonymous referees as well as Laurent Barras, Claude Erb, Juhani Linnainmaa, Michael Weber and seminar participants at Rice University, University of Southern California, University of California at Irvine, Hanken School of Economics, Man-Numeric, Research Affiliates and the 2018 Western Finance Association meetings in San Diego. Kay Jaitly provided editorial assistance.

1 Introduction

In manager selection (or equivalently the selection of factors or trading strategies), investors can make two types of mistakes. The first involves selecting a manager who turns out to be unskilled: this is a Type I error, or a false positive.[1] The second error is not selecting or missing a manager the investor thought was unskilled, but was not: this is a Type II error, or a false negative. The latter error is rarely quantified in finance. Indeed, an economically important variation exists within the errors that should be taken into account; that is, all false positives are not equal, because one manager might slightly underperform and another manager might have a large negative return. Further, it makes sense to treat Type I and Type II errors asymmetrically: the cost of a Type II error is likely different from the cost of a Type I error, and both costs depend on the specific decision at hand. For example, suppose an investor wants to pick managers with the criterion that Type I errors are five times more costly than Type II errors. Current tools do not allow for such a selection criterion.

More fundamentally, it is difficult to simply characterize Type I errors because of multiple testing. Using a single-hypothesis testing criterion (two standard errors from zero) will lead to massive Type I errors when there are thousands of managers because many look good purely by luck. Statisticians have suggested a number of fixes that take multiple testing into account. For example, the simplest is the Bonferroni correction, which multiplies each manager's p-value by the number of managers. This type of correction does not take the covariance structure into account. Further, it is not obvious what the Type I error rate is after implementing the correction. We know the error rate is less than the single testing criteria — but how much less?

We offer a different approach. Based on actual manager data, we determine the performance threshold that delivers a particular Type I error rate (e.g., 5%). We also characterize the Type II error rate associated with our optimized Type I error rate. It is then straightforward to scale the Type II error by the Type I error and solve for the cutoff that produces the desired trade-off of false negatives and false positives.

Our focus on both Type I and Type II errors echoes recent studies in economics that highlight the importance of examining test power. For example, by surveying a large number of studies, Ioannidis, Stanley, and Doucouliagos (2017) show that many research areas have

[1] Throughout our paper, we follow the convention of the empirical literature on performance evaluation to associate manager skill with the magnitude of alpha (e.g., skilled managers are those that generate positive alphas). Our notion of manager skill is thus different from that in Berk and Green (2004) where skilled managers generate a zero net alpha in equilibrium.

90% of their results under-powered, leading to an exaggeration of the results. Ziliak and McCloskey (2004) find that only 8% of papers published in the *American Economic Review* in the 1990s consider test power. The issue of test power adds one more challenge to prevalent research practices in economics research (Leamer, 1983, De Long and Lang, 1992, Ioannidis and Doucouliagos, 2013, Harvey and Liu, 2013, Harvey, Liu, and Zhu, 2016, Harvey, 2017).

Why is test power important to research in financial economics? On the one hand, when the main finding of the research is the non-existence of an effect (i.e., the null hypothesis is not rejected), test power directly affects the credibility of the finding because it determines the probability of not rejecting the null hypothesis when the effect is true. For example, in one of our applications, we show that existing studies lack power in detecting outperforming mutual funds. On the other hand, when the main finding is the rejection of the null hypothesis (i.e., the main hypothesis), this finding often has to survive against alternative hypotheses (i.e., alternative explanations for the main finding). A low test power for alternative explanations generates a high Type I error rate for the main hypothesis (Ioannidis, 2005).

Our paper addresses the issue of test power in the context of multiple tests. Our contribution is threefold. First, we introduce a framework that allows for an intuitive definition of test power. Second, we offer a double bootstrap approach to flexibly (depending on the particular dataset) estimate test power. Lastly, we illustrate how the consideration of test power materially changes our interpretation of some important research findings in the current literature.

In a single-hypothesis test, the Type II error rate at a particular parameter value (i.e., performance metric for the manager) is calculated as the probability of failing to reject the null hypothesis at this value. In multiple tests, the calculation of the Type II error rate is less straightforward because, instead of a single parameter value, we need to specify a vector of non-zero parameters, with each parameter corresponding to a single test under the alternative hypothesis.

We propose a simple idea to estimate the Type II error rate. Assuming that a fraction of p_0 of managers have skill, we adjust the data so that p_0 of managers have skill (with their skill level set at the in-sample estimate), and the remaining fraction of $1 - p_0$ of managers have no skill (with their skill level set to be a zero excess return). By bootstrapping from these adjusted data, we evaluate the Type II error rate through simulations. Our method thus circumvents the difficulty of specifying the high-dimensional parameter vector under the alternative hypothesis. We set the parameter vector at what we consider a reasonable value — the in-sample estimate corresponding to a certain p_0. In essence, we treat p_0 as a sufficient statistic, which helps estimate the Type II error rate. We interpret p_0 both from a frequentist and a Bayesian perspective.

Our paper is related to bootstrap approach in performance evaluation proposed by Kosowski, Timmermann, Wermers, and White (KTWW, 2006) and Fama and French (2010).[2] These papers use a single bootstrap approach to adjust for multiple testing. In particular, under the assumption of no skill for all funds ($p_0 = 0$), they demean the data to create a "pseudo" sample, Y_0, for which $p_0 = 0$ exactly holds true in sample. They then bootstrap Y_0 to test the overall hypothesis that all funds have a zero alpha. Because we are interested in both the Type I and Type II error rates committed by a given testing procedure (including those of KTWW and Fama and French), our method uses two rounds of bootstrapping. For example, to estimate the Type I error rate committed by Fama and French, we first bootstrap Y_0 to create a perturbation, Y_i, for which the null hypothesis is true. We then apply Fama and French (i.e., second bootstrap) to each Y_i and record the testing outcome ($h_i = 1$ if rejection). We estimate the Type I error rate as the averaged h_i. The Type II error rate can be estimated in a similar fashion.

We focus on two empirical applications to illustrate how our framework helps address important issues related to Type I and Type II errors associated with multiple tests.

We first apply our method to the selection of two sets of investment factors. The first database includes hundreds of backtested factor returns. For a given p_0, our method allows us to measure the Type I and Type II error rates for these factors; the ability to do so allows investors to make choices that strike a balance between Type I and Type II errors. When p_0 is uncertain, we use our method to evaluate the performance of existing multiple-testing adjustments. Investors are able to select the adjustment that works well regardless of the value of p_0 or for a range of values of p_0. Indeed, our application shows that usually there is an ordering (best to worst) of multiple-testing methods in performance, regardless of the values of p_0.

The second database includes around 18,000 anomaly strategies, which are constructed and studied by Yan and Zheng (2017). Relying on the Fama and French (2010) approach to adjust for multiple testing, Yan and Zheng (2017) claim that a large fraction of the 18,000 anomalies in their data are true and they attribute anomaly returns to mispricing. We use our model to estimate the error rates of their approach and produce results that are inconsistent with their narrative.

[2] See Harvey and Liu (2018a) for another application of the bootstrap approach to the test of factor models.

We then apply our approach to performance evaluation, revisiting the problem of determining whether mutual fund managers have skill. In particular, we use our double bootstrap technique to estimate the Type I and Type II error rates committed by the popular Fama and French (2010) approach. We find their approach lacks power to detect outperforming funds. Even when a significant fraction of funds are outperforming, and moreover, endowed with the large magnitude of returns that these funds have in the actual sample, the Fama and French method may still declare, with a high probability, a zero alpha across all funds. Our result provides pause for their conclusions regarding mutual fund performance and helps reconcile the difference between KTWW (2006) and Fama and French (2010).

Our paper is not alone in raising the issue of power in performance evaluation. Ferson and Chen (2017), Andrikogiannopoulou and Papakonstantinou (2018), and Barras, Scaillet, and Wermers (2018) focus on the power that results from applying the false discovery rate approach in estimating the fraction of outperforming funds of Barras, Scaillet, and Wermers (2010). Our paper differs by proposing a non-parametric bootstrap-based approach to systematically evaluate test power. We also apply our method to a wide range of issues in financial economics, including the selection of investment strategies, identification of equity market anomalies, and evaluation of mutual fund managers. Chordia, Goyal, and Saretto (2018) study an even larger collection of anomalies than do Yan and Zheng (2017) and use several existing multiple-testing adjustment methods to estimate the fraction of true anomalies. In contrast to Chordia, Goyal, and Saretto, we focus on the comparison of different methods using our bootstrap-based approach. In particular, we show exactly what went wrong with the inference of Yan and Zheng (2017).

Our paper also contributes to the growing literature in finance that applies multiple-testing techniques to related areas in financial economics (see, e.g., Harvey, Liu, and Zhu, 2016, Harvey, 2017, Chordia, Goyal, and Saretto, 2018, Barras, 2018, Giglio, Liao, and Xiu, 2018). One obstacle for this literature is that, despite the large number of available methods developed by the statistics literature, it is unclear which method is most suitable for a given data set. Our paper provides a systematic approach that offers data-driven advice on the relative performance of multiple testing adjustment methods.

Our paper is organized as follows. In the second section, we present our method. In the next section, we apply our method to the selection of investment strategies and performance evaluation for mutual funds. Some concluding remarks are offered in the final section.

2 Method

2.1 Motivation: A Single Hypothesis Test

Suppose we have a single hypothesis to test and the test is about the mean of a univariate variable X. For a given testing procedure (e.g., the sample mean t test), there are at least two metrics that are important for gauging the performance of the procedure. One is the Type I error rate, which is the probability of incorrectly rejecting the null when it is true (a false positive); the other is the Type II error rate, which is probability of incorrectly declaring insignificance when the alternative hypothesis is true (a false negative). The Type II error rate is also linked to test power (i.e., power $= 1-$ Type II error rate), which is probability of correctly rejecting the null when the alternative hypothesis is true.

In a typical single hypothesis test, we try to control the Type I error rate at a certain level (e.g., 5% significance) while seeking methods that generate a low Type II error rate or, equivalently, a high test power. To evaluate the Type I error rate, we assume that the null is true (i.e., $\mu_0 = 0$) and calculate the probability of a false discovery. For the Type II error rate, we assume a certain level (e.g., $\mu_0 = 5\%$) for the parameter of interest (i.e., the mean of X) and calculate the probability of a false negative as a function of μ_0.

In the context of multiple hypothesis testing, the evaluation of Type I and Type II error rates is less straightforward for several reasons. First, for the definition of the Type I error rate, the overall hypothesis that the null hypothesis holds for each individual test may be too special to be realistic for certain applications. As a result, we often need alternative definitions of Type I error rates that apply even when some of the null hypotheses may not be true.[3] For the Type II error rate, its calculation in general depends on the set of parameters of interest, which is a high-dimensional vector as we have multiple tests. As a result, it is not clear what value for this high-dimensional vector is most relevant for the calculation of the Type II error rate.

Given this difficulty in determining the Type II error rate, current multiple testing adjustments often only focus on the Type I errors. For example, Fama and French (2010) look at mutual fund performance and test the overall null hypothesis of a zero alpha for all funds. They do not assess the performance of their method when the alternative is true, i.e., the probability of incorrectly declaring insignificant alphas across all funds when some funds display skill. As another example, the multiple testing adjustments studied in Harvey, Liu, and

[3] One example is the false discovery rate, as we shall see later.

Zhu (2016) focus on the Type I error defined by either the FWER (family-wise error rate, the probability of making at least one false discovery) or the FDR (false discovery rate, the expected fraction of false discoveries among all discoveries). Whether or not these methods have good performance in terms of Type I error rates and what the implied Type II error rates are remain open questions.

Second, while we can often calculate Type I and Type II error rates analytically under certain assumptions for a single hypothesis test, such assumptions become increasingly untenable when we have many tests. For example, it is difficult to model cross-sectional dependence when we have a large collection of tests.

Third, when there are multiple tests, even the definitions of Type I and Type II error rates become less straightforward. While traditional multiple testing techniques apply to certain definitions of Type I error rates such as FWER or FDR, we are interested in a general approach that allows us to evaluate different measures of the severity of false positives and false negatives. For example, while the family-wise error rate from the statistics literature has been applied in Harvey, Liu, and Zhu (2016) to evaluate strategies based on anomalies, an odds ratio that weighs the number of false discoveries against the number of misses may be more informative for the selection of investment strategies as it may be more consistent with the manager's objective function.

Motivated by these concerns, in the next section we provide a general framework that allows us to evaluate error rates when there are multiple tests. Our framework is characterized by three features. First, we propose a simple metric to summarize information in the parameters of interest and use it to evaluate Type I and Type II error rates. In essence, this metric reduces the dimensionality of the parameters of interest and allows us to evaluate error rates around what we consider a reasonable set of parameter values. Second, we evaluate error rates through a bootstrap method, which allows us to capture cross-sectional dependence nonparametrically. Third, our method is quite flexible in terms of how we define the severity of false positives and false negatives, making it possible for us to evaluate error rate definitions that are appropriate for a diverse set of finance applications.

2.2 Bootstrapped Error Rates under Multiple Tests

To ease the exposition, we describe our method in the context of testing the performance of many trading strategies. Suppose we have N strategies and D time periods. We arrange the data into a $D \times N$ data matrix X_0.

Suppose one believes that a fraction, p_0, of the N strategies are true. We provide a simulation-based framework to evaluate error rates related to multiple hypothesis testing corresponding to a given p_0.

There are several ways to interpret p_0. When $p_0 = 0$, no strategy is believed to be true, which is the overall null hypothesis of a zero return across all strategies. This hypothesis can be economically important. For example, KTWW and Fama and French (2010) examine this hypothesis for mutual funds to test market efficiency. We discuss this hypothesis in detail in the next section when we apply our method to Fama and French (2010).

When $p_0 > 0$, some strategies are believed to be true, then p_0 can be thought of a plug-in parameter — similar to the role of μ_0 in a single test as detailed in the previous section — that helps us measure the error rates in the presence of multiple tests. As we discussed previously in the context of multiple tests, in general one needs to make assumptions on the values of the population statistics (e.g., the mean return of a strategy) of all the strategies that are believed to be true in order to determine the error rates. However, we argue that in our framework p_0 is a single summary statistic that allows us to effectively evaluate error rates without having to condition on the values of the population statistics. As a result, we provide a simple way to extend the error rate analysis for a single hypothesis test to multiple hypothesis tests.

Note that by choosing a certain p_0, investors are implicitly taking a stand on the plausible economic magnitudes of alternative hypotheses. For example, investors may have a belief that strategies with a mean return above 5% are likely to be true, resulting in a corresponding p_0. While this is one way of rationalizing the choice of p_0 in our framework, our model allows us to evaluate the implications of not only p_0 or the 5% cutoff but the entire distribution of alternatives on error rates.

Parameter p_0 also has a Bayesian interpretation. Harvey, Liu, and Zhu (2016) present a stylized Bayesian framework for multiple hypothesis testing where multiple testing adjustment is achieved indirectly through the likelihood function. On the other hand, Harvey (2017) recommends the use of the minimum Bayes factor, which builds on Bayesian hypothesis testing but abstracts from the prior specification by focusing on the prior that generates the minimum Bayes factor. Our treatment of p_0 lies in between Harvey, Liu, and Zhu (2016) and Harvey (2017) in the sense that while we do not go as far as making assumptions on the prior distribution of p_0 that is later fed into the full-blown Bayesian framework as in Harvey, Liu, and Zhu (2016), we do deviate from the Harvey (2017) assumption of a degenerate prior (i.e., the point mass that concentrates on the parameter value that generates the minimum

Bayes factor) by exploring how error rates respond to changes in p_0. While we do not attempt to pursue a complete Bayesian solution to multiple hypothesis testing,[4] the sensitivity of error rates to changes in p_0 that we highlight in this paper are important ingredients to both Bayesian and frequentist hypothesis testing.

Although the choice of p_0 is inherently subjective, we offer several guidelines as to the selection of an appropriate p_0. First, an examination of the summary statistics of the data can help narrow down the range of reasonable priors.[5] For example, about 22% of strategies in our CAPIQ sample (see later sections for details on this data) have a t-statistic above 2.0. This suggests that, at 5% significance level, p_0 is likely lower than 22% given the need for a multiple testing adjustment. Second, it may be a good idea to apply prior knowledge to elicit p_0. For example, researchers with a focus on quantitative asset management may have an estimate of the success rate of finding a profitable investment strategy based on past experience. Such an estimate can guide their choice of p_0. Finally, while researchers in principle can pick any p_0 in our model, we present a simulation framework in later sections that helps gauge the potential loss from applying different priors.

For a given p_0, our method starts by choosing $p_0 \times N$ strategies that are deemed as true. A simple way to choose these strategies is to first rank the strategies by their t-statistics and then choose the top $p_0 \times N$ with the highest t-statistics. While this approach is consistent with the idea that strategies with higher t-statistics are more likely to be true, it ignores the sampling uncertainty in ranking the strategies. To take this certainty into account, we perturb the data and rank the strategies based on the perturbed data. In particular, we bootstrap the time periods and create an alternative panel of returns, X_i (note that the original data matrix is X_0). For X_i, we rank its strategies based on their t-statistics. For the top $p_0 \times N$ strategies with the highest t-statistics, we find the corresponding strategies in X_0. We adjust these strategies so that their in-sample means are the same as the means for the top $p_0 \times N$ strategies in X_i.[6] We denote the data matrix for these adjusted strategies as $X_{0,1}^{(i)}$. For the remaining strategies in X_0, we adjust them so they have a zero in-sample mean (denote the data matrix for these adjusted strategies as $X_{0,0}^{(i)}$). Finally, we arrange $X_{0,1}^{(i)}$ and $X_{0,0}^{(i)}$ into a new data matrix Y_i by concatenating the two data matrices. This Y_i will be

[4] Storey (2003) provides a Bayesian interpretation of the positive false discovery rate. Scott and Berger (2006) have a general discussion of Bayesian multiple testing. Harvey, Liu, and Zhu (2016) discuss some of the challenges of addressing multiple testing within a Bayesian framework. Harvey, Liu, Polson, and Xu (2019) present a full-blown Bayesian framework to test market efficient?

[5] See Harvey (2017) for examples on Bayes factors that are related to data-driven priors.

[6] Alternatively, if a factor model is used for risk adjustment, we could adjust the intercepts of these strategies after estimating a factor model so that the adjusted intercepts are the same as those for the top $p_0 \times N$ strategies in X_i.

the hypothetical data that we use to perform our follow-up error rate analysis, for which we know exactly which strategies are believed to be true and which strategies are false.

Our idea of constructing a "pseudo" sample under the alternative hypothesis (i.e., some strategies are true) is motivated by the bootstrap approach proposed by the mutual fund literature. In particular, KTWW perform a bootstrap analysis at the individual fund level to select "star" funds. Fama and French (2010) look at the cross-sectional distribution of fund performance to control for multiple testing. Both papers rely on the idea of constructing a "pseudo" sample of fund returns for which the null hypothesis of zero performance is known to be true. We follow them by constructing a similar sample for which some of the alternative hypotheses are known to be true, with their corresponding parameter values (i.e., strategy means) set at plausible values, i.e., their in-sample means associated with our first-stage bootstrap.

Notice that due to sampling uncertainty, what constitutes alternative hypotheses in our first-stage bootstrap may not correspond to the true alternative hypotheses for the underlying data generating process. In particular, strategies with a true zero mean return in population may generate an inflated mean after the first-stage bootstrap and are thus falsely classified as alternative hypotheses. While the existing literature offers several methods to shrink the in-sample means (e.g., Jones and Shanken, 2005, Andrikogiannopoulou and Papakonstantinou, 2016, and Harvey and Liu, 2018), we do not embed them into our current paper. In a simulation study in which we evaluate the overall performance of our approach, we treat the misclassification of hypotheses as one potential source that affects our model performance. Under realistic assumptions of the data generating process, we show that our method performs well despite the potential misclassification of alternative hypotheses.

For Y_i, we bootstrap the time periods J times to evaluate the error rates for a statistical procedure, such as a fixed t-statistic threshold (e.g., a conventional t-statistic threshold of 2.0) or a range of multiple testing approaches detailed in Harvey, Liu, and Zhu (2016). By construction, we know which strategies are believed to be true (and false) in Y_i, which allows us to summarize the testing outcomes for the j-th bootstrap iteration with a vector $\bar{O}^{i,j} = (TN^{i,j}, FP^{i,j}, FN^{i,j}, TP^{i,j})'$, where $TN^{i,j}$ is the number of tests that correctly identify a false strategy as false (true negative), $FP^{i,j}$ is the number of tests that incorrectly identify a false strategy as true (false positive), $FN^{i,j}$ is the number of tests that incorrectly identify a true strategy as false (false negative), and $TP^{i,j}$ is the number of tests that correctly identify a true strategy as true (true positive). Notice that for brevity we suppress the dependence of $\bar{O}^{i,j}$ on the significance threshold (i.e., either a fixed t-statistic threshold or the threshold generated by a data-dependent testing procedure). Table 1 illustrates these four summary statistics using a contingency table.

Table 1: **Classifying Testing Outcomes**

Decision	Null (H_0)	Alternative (H_1)
Reject	False positive (Type I error) $(FP^{i,j})$	True positive $(TP^{i,j})$
Accept	True negative $(TN^{i,j})$	False negative (Type II error) $(FN^{i,j})$

With these summary statistics, we can construct several error rates that are of interest to us. We focus on three types of error rates in our paper. The first type is motivated by the false discovery rate (see Benjamini and Hochberg, 1995, Benjamini and Yekutieli, 2001, Barras, Scaillet, and Wermers, 2010 and Harvey, Liu, and Zhu 2016) and is defined as:

$$RFDR^{i,j} = \begin{cases} \frac{FP^{i,j}}{FP^{i,j}+TP^{i,j}}, & \text{if } FP^{i,j}+TP^{i,j} > 0, \\ 0, & \text{if } FP^{i,j}+TP^{i,j} = 0, \end{cases}$$

where 'RFDR' stands for the *realized false discovery rate*, which is the fraction of false discoveries (i.e., $FP^{i,j}$) among all discoveries (i.e., $FP^{i,j} + TP^{i,j}$). The expected value of $RFDR$ extends the Type I error rate in a single hypothesis test to multiple tests.

The second type, motivated by the false discovery rate, aims to capture the rate of misses and is defined as:

$$RMISS^{i,j} = \begin{cases} \frac{FN^{i,j}}{FN^{i,j}+TN^{i,j}}, & \text{if } FN^{i,j}+TN^{i,j} > 0, \\ 0, & \text{if } FN^{i,j}+TN^{i,j} = 0, \end{cases}$$

where 'RMISS' stands for the *realized rate of misses* (sometimes referred to as the false omission rate or false non-discovery rate), which is the fraction of misses (i.e., $FN^{i,j}$) among all tests that are declared insignificant (i.e., $FN^{i,j} + TN^{i,j}$). The expected value of $RMISS$ extends the Type II error rate in a single hypothesis test to multiple tests.[7]

[7] Alternative error rate definitions include *precision* p (defined as the ratio of the number of correct positives to the number of all predicted positives, i.e., $TP^{i,j}/(FP^{i,j} + TP^{i,j})$) and *recall* r (defined as the ratio of the number of true positives to the number of strategies that should be identified as positive, i.e., $TP^{i,j}/(TP^{i,j} + FN^{i,j})$, and is also known as the hit rate or true positive rate). One can also define the false discovery rate as the expected fraction of false discoveries among all tests for which the null is true, which more closely corresponds to the Type I error rate definition for a single test.

Finally, similar to the concept of odds ratio in Bayesian analysis, we define the ratio of false discoveries to misses as:

$$RRATIO^{i,j} = \begin{cases} \frac{FP^{i,j}}{FN^{i,j}}, \text{ if } FN^{i,j} > 0, \\ 0, \quad \text{ if } FN^{i,j} = 0, \end{cases}$$

where 'RRATIO' stands for the *realized ratio of false discoveries over misses*, which is simply the ratio of false discoveries (i.e., $FP^{i,j}$) over misses (i.e., $FN^{i,j}$).[8]

Notice that by using summary statistics that count the number of occurrences for different types of testing outcomes, we are restricting ourselves to error rate definitions that only depend on the number of occurrences. Alternative definitions of error rates that may involve the magnitudes of the effects being tested (e.g., an error rate that puts a higher weight on a missed strategy with a higher Sharpe ratio) can also be studied in our framework.[9]

Finally, we take in account the sampling uncertainty in ranking the strategies and the uncertainty in generating the realized error rates for each particular ranking by averaging across both i and j. Suppose we perturb the data I times, and for each time we generate J bootstrapped random samples. We have:

$$\begin{aligned} TYPE1 &= \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} RFDR^{i,j}, \\ TYPE2 &= \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} RMISS^{i,j}, \\ ORATIO &= \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} RRATIO^{i,j}. \end{aligned}$$

We refer to $TYPE1$ as the Type I error rate, $TYPE2$ as the Type II error rate, and $ORATIO$ as the odds ratio (between false discoveries and misses) in our paper. Notice that similar to $\bar{O}^{i,j}$, our estimated $TYPE1$, $TYPE2$ and $ORATIO$ implicitly depend on the significance threshold.

There are several advantages of $ORATIO$ compared to $TYPE1$ and $TYPE2$. First, $ORATIO$ links Type I error and Type II error together by quantifying the chance of a false discovery

[8] In most of our applications, there are always misses so the difference between our current definition (i.e., $RRATIO^{i,j} = 0$ if $FN^{i,j} = 0$) and alternative definitions, such as excluding simulation runs for which $FN^{i,j} = 0$, is small.

[9] See DeGroot (1975), DeGroot and Schervish (2011, chapter 9), and Beneish (1997, 1999).

per miss. For example, if an investor believes that the cost of a Type I error is ten times that of a Type II error, then the optimal $ORATIO$ should be 1/10. Second, the measure takes the magnitude of p_0 into account. When p_0 is very small, $TYPE2$ is usually much smaller than $TYPE1$. However, this mainly reflects the rare occurrence of the alternative hypothesis and does not necessarily imply the good performance of the model in controlling $TYPE2$. In this case, $ORATIO$ may be a more informed metric in balancing Type I and Type II errors. While we do not attempt to specify the relative weight between $TYPE1$ and $TYPE2$ (which likely requires the specification of a loss function that weighs Type I errors against Type II errors), we use $ORATIO$ as a heuristic metric to weigh Type I errors against Type II errors.

To summarize, we follow the steps below to evaluate the error rates:

Step I Bootstrap the time periods and let the bootstrapped panel of returns be X_i. For X_i, obtain the corresponding $1 \times N$ vector of t-statistics t_i;

Step II Rank the components in t_i. For the top p_0 of strategies in t_i, find the corresponding strategies in the original data X_0. Adjust these strategies so they have the same means as those for the top p_0 of strategies ranked by t_i in X_i. Denote the data matrix for the adjusted strategies as $X_{0,1}^{(i)}$. For the remaining strategies in X_0, adjust them so they have a zero mean in-sample (denote the corresponding data matrix as $X_{0,0}^{(i)}$). Arrange $X_{0,1}^{(i)}$ and $X_{0,0}^{(i)}$ into a new data Y_i;

Step III Bootstrap the time periods J times. For each bootstrapped sample, calculate the realized error rates (or odds ratio) for Y_i, denoted as $f_{i,j}$ (f stands for a generic error rate that is a function of the testing outcomes);

Step IV Repeat Step I-III I times. Calculate the final bootstrapped error rate as $\frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} f_{i,j}$.

Again, keep in mind that the calculation of the realized error rate (i.e., $f_{i,j}$) in Step III requires the specification of the significance threshold (or a particular data-dependent testing procedure). As a result, the final bootstrapped error rates produced by our model are (implicitly) functions of the threshold.

2.3 Type II Error Rates under Multiple Tests

While our definition of the Type I error rate (i.e., the FDR) is intuitive and fairly standard in the finance and statistics literature, our definition of the Type II error rate (i.e., the false omission rate) deserves further clarification.

First of all, the way we define the Type II error rate is analogous to how we define the FDR. This can be seen by noting that while the FDR uses the total number of positives (i.e., TP + FP) as the denominator, our Type II error rate uses the total number of negatives (i.e., TN + FN). Several existing papers in statistics also promote the use of the false omission rate — also referred to as the false non-discovery rate — to measure the Type II error rate in the context of multiple testing (e.g., Genovese and Wasserman, 2002, Sarkar, 2006).

Many are familiar with the usual definition of the Type II error rate under single testing, which translates into the expected fraction of false negatives out of the total number of alternatives (i.e., the false negative rate, $\frac{FN}{(FN+TP)}$) in multiple testing. While we can easily accommodate this particular definition, we believe that the false omission rate is more appropriate under multiple testing and more suited to situations that are relevant to finance.

Note that these alternative definitions of the Type II error rates are transformations of each other and the Type I error rate, so we do not lose information by focusing on the false omission rate. For a given dataset, the number of alternatives (i.e., $FN + TP$) and the number of nulls (i.e., $TN + FP$) are both fixed. There are only two unknowns among the four numbers. So any Type II error rate definition will be implied by the Type I error rate and the false omission rate.[10]

The Type II error rate is linked to power. As a result, our Type II error rate implies a different interpretation of power that is best demonstrated with an example. Suppose there are 100 managers and five are skilled. Suppose our test procedure correctly declares all 95 unskilled managers as unskilled and identifies three out of five managers as skilled. One way to define the Type II error would be 2/5, implying that 60% ($= 1 - 3/5$) are correctly identified, which corresponds to the usual definition of power in a single test. Now suppose we increase the total number of managers to 1,000 but keep the same number of skilled managers, five, and the same number of identified skilled managers, three. This would imply the same Type II error rate as before (i.e., 2/5), making the two examples indistinguishable from a Type II error perspective. However, it seems far more impressive for a testing method

[10] To be precise, the realized error rates are explicit functions of the two unknowns. So this is only approximately true after taking expectations.

to correctly identify three out of five managers among 1,000 managers rather than 100. This is exactly what is our definition of the Type II error rate, which produces an error of 2/97 ($= 2/(95+2)$ for 100 tests) versus 2/997 ($= 2/(995+2)$ for 1,000 tests).

While we focus on the false omission rate in our paper as the Type II error rate definition, our bootstrap-based approach can easily accommodate alternative definitions such as the false negative rate or even definitions that have differential weights on individual errors. We view this as an important advantage of our framework in comparison to the existing statistics literature (e.g., Genovese and Wasserman, 2002, Sarkar, 2006).

2.4 Bootstrapped Error Rates for Fama and French (2010)

Fama and French (2010) focus on the overall null hypothesis of zero alpha across mutual funds to test a version of the efficient market hypothesis. Therefore, the relevant error rates in their context are the probability of rejecting this overall null hypothesis when it is true (Type I error) and the probability of not rejecting this null hypothesis when some funds have the ability to generate a positive alpha (Type II error). We apply our framework in the previous section to find the corresponding Type I (denoted as $TYPE1_{ff}$) and Type II (denoted as $TYPE2_{ff}$) error rates.

We first focus on Type I error. This corresponds to the case of $p_0 = 0$ in our framework described in the previous section. Let the original data be X_0. Similar to Fama and French (2010), we subtract the in-sample alpha estimate from each fund's returns to generate a "pseudo" sample of funds whose alphas are precisely zero. Let this sample be Y_0. We treat Y_0 as the population of fund returns for which the overall null hypothesis of a zero alpha across all funds is true.

Fama and French (2010) bootstrap Y_0 to generate distributions of the cross-section of t-statistics and compare these bootstrapped distributions to the actual distribution to make inference. Similar to their idea (and the earlier idea in KTWW), if Y_0 can be treated as the "pseudo" sample of fund returns under the null hypothesis, then any bootstrapped version of Y_0 can also be regarded as the "pseudo" sample. This provides the basis for our evaluation of the Type I error rate for Fama and French (2010). We bootstrap Y_0 many times to generate alternative samples for which the overall null hypothesis of a zero alpha across funds is true. For each sample, we apply Fama and French (2010) to see if the null hypothesis is (falsely) rejected. We take the average rejection rate as the Type I error rate.

In particular, we bootstrap the time periods to perturb Y_0. This is where our approach departs from Fama and French (2010) in that while they make a one-shot decision for Y_0, we need perturbations of Y_0 to simulate the error rates committed by the Fama-French approach. Let the perturbed data be Y_i. Notice that due to sampling uncertainty, fund alphas are no longer zero for Y_i, although the overall null hypothesis is still true since Y_i is obtained by simply perturbing Y_0. For Y_i, we perform the Fama-French testing approach and let the testing outcome be h_i, where h_i equals one if we reject the null hypothesis and zero otherwise. Finally, we perturb Y_0 many times and calculate the empirical Type I error rate (i.e., $TYPE1_{ff}$) as the averaged h_i.

We follow a procedure that is similar to the one described in the previous section to generate the Type II error rate. In particular, for a given p_0 of the fraction of funds with a positive alpha, we bootstrap the time periods and identify the top p_0 of funds with the highest t-statistics for alphas. We find the corresponding funds in the original data and adjust them so they have the same alphas as those for the top p_0 of funds in the bootstrapped sample (denote the data matrix for the adjusted strategies as $X_{1,0}^{(i)}$). At the same time, we adjust the returns of the remaining funds so that they have an alpha of zero (denote the corresponding data matrix as $X_{0,0}^{(i)}$) . By joining $X_{1,0}^{(i)}$ with $X_{0,0}^{(i)}$, we obtain a new panel, based on which we apply the Fama-French approach and record the testing outcome as $l_i = 1$ if not rejecting the null hypothesis (and zero otherwise). The empirical Type II error rate (i.e., $TYPE2_{ff}$) is calculated as the averaged l_i.

In summary, while Fama and French (2010) make a one-time decision about whether or not to reject the overall null hypothesis for a given set of fund returns, our approach allows us to calibrate the error rates committed by Fama and French (2010) by repeatedly applying their method to bootstrapped data that are generated from the original data.

2.5 Discussion

While traditional frequentist single hypothesis testing focuses on the Type I error rate, more powerful frequentist approaches (e.g., likelihood-ratio test) and Bayesian hypothesis testing take into account both the Type I and Type II error rates. However, as we have mentioned, in the context of multiple testing, the evaluation of the Type II error rate is difficult for at least two reasons. First, there are a large number of alternative hypotheses making this a potentially intractable multidimensional problem. Second, the multidimensional nature of the data, in particular the dependence among tests, makes the inference on the joint distribution of error rates across tests difficult. Our framework provides solutions to both problems.

Given the large number of alternative hypotheses, it seems inappropriate to focus on any particular parameter vector as the alternative hypothesis. Our double-bootstrap approach simplifies the specification of alternative hypotheses by grouping similar alternative hypotheses. In particular, all alternative hypotheses that correspond to the case in which a fraction of p_0 of the hypotheses are true (and the rest are false and hence set at the null hypotheses) are classified into a single H_1 that is associated with p_0.

While our grouping procedure helps achieve dimension reduction of the space of alternative hypotheses, there are still a large number of possible alternative hypotheses that are associated with a given p_0. However, we argue that many of these alternative hypotheses are unlikely to be true for the underlying data generating process and hence irrelevant for hypothesis testing. First, hypotheses with a low in-sample t-statistic are less likely to be true than hypotheses with a high in-sample t-statistics. Second, the true parameter value for each effect under consideration is not arbitrary. It can be estimated from the data. Our bootstrap-based approach takes both of these points into account. In particular, our ranking of t-statistics based on the bootstrapped sample in Step I addresses the first issue, and our assignment of true parameter values to a fraction of p_0 tests that are deemed as true after Step I addresses the second issue.

After we fix p_0 and find a parameter vector that is consistent with p_0 through the first-stage bootstrap, our second-stage bootstrap allows us to evaluate any function of error rates (both Type I and Type II) nonparametrically. Importantly, bootstrapping provides a convenient approach to take data dependence into account, as argued in Fama and French (2010).

Lastly, our framework allows us to make data specific recommendations for the statistical cutoffs, as well as to evaluate the performance of any particular multiple testing adjustment. In our applications, we provide examples of both. Given the large number of multiple testing methods that are available and the potential variation of the performance of a given method across different datasets, we view it as necessary to be able to generate data specific statistical cutoffs that exactly achieve a pre-specified Type I or Type II (or a combination of both) error constraint. For example, Harvey, Liu, and Zhu (2016) apply several well-known multiple testing methods to anomaly data to control for the false discovery rate (FDR). However, whether these or other methods can achieve the pre-specified false discovery rate for these anomaly data is unknown. Our double-bootstrap approach allows us to accurately calculate the FDR conditional on the t-statistic cutoff and p_0. Consequently, we can choose the t-statistic cutoff that exactly achieves a pre-specified FDR for values of p_0 that are deemed plausible.

While our paper focuses on estimating statistical objectives such as the false discovery rate, one may attempt to apply our approach to economic objectives such as the Sharpe ratio of a portfolio of strategies or funds. However, there several reasons to exercise caution against such an attempt.

First, there seems to be a disconnect between statistical objectives and economic objectives in the finance literature. This tension traces back to two strands of literature. One is the performance evaluation literature where researchers or practitioners, facing thousands of funds to select from, focus on simple statistics (i.e., statistical objectives) such as the false discovery rate. The other strand is the mean-variance literature where combinations of assets are considered to generate efficient portfolios. Our framework is more applicable to the performance evaluation literature.[11] Following the logic of this literature, we are more interested in estimating economically motivated quantities such as the fraction of outperforming funds than trying to find the optimal combinations of funds due to either practical constraints (e.g., difficult to allocate to many funds) or computational issues (i.e., cross-section is too large to reliably estimate the covariance matrix). While certain extensions of conventional statistical objectives in multiple testing (e.g., differential weights on the two types of error rates) are allowed in our framework, these extensions should still be considered rudimentary from the standpoint of economic objectives.

Second, statistical objectives and economic objectives may not be consistent with each other. For example, an individual asset with a high Sharpe ratio, which will likely be declared significant in a multiple testing framework, may not be as attractive in the mean-variance framework once its correlations with other assets are taken into account.[12] As another example, the quest for test power may not perfectly align with economic objectives as methods that have low statistical power may still have important implications from an economic perspective (Kandel and Stambaugh, 1996).

[11] Note that although models used in the performance evaluation literature can generate the risk-adjusted alpha that indicates the utility gain (under certain assumptions of the utility function) by allocating to a particular fund, they still do not account for the cross-sectional correlations in idiosyncratic fund returns that impact the allocation to a group of funds, which is the objective of the mean-variance literature.

[12] Here correlations may be caused by correlations in idiosyncratic risks that are orthogonal to systematic risk factors.

We require that p_0 be relatively small so that the top $p_0 \times N$ strategies in the first-stage bootstrap always have positive means.[13] This ensures that the population means for strategies that are deemed true in the second-stage bootstrap are positive, which is consistent with the alternative hypotheses and is required by our bootstrap approach. This is slightly stronger than requiring p_0 to be smaller than the fraction of strategies that have positive means for the original sample since we need this (i.e., p_0 is smaller than the fraction of strategies with positive means) to be the case for all bootstrapped samples.[14] In our applications, we make sure our choices of p_0 always meet this condition. We believe our restriction on p_0 is reasonable. It makes little sense to believe that every strategy with a positive return is true. Due to sampling uncertainty some zero-mean strategies will have positive returns.

3 Applications

We provide two applications of our framework. The first studies two large groups of investment strategies. We illustrate how to use our method to select outperforming strategies and compare our method to existing multiple testing techniques. Our second application revisits Fama and French (2010)'s conclusion about mutual fund performance. We show that Fama and French's technique is underpowered in that it is unable to identify truly outperforming funds. Overall, our two applications highlight the flexibility of our approach in addressing questions that are related to different aspects of the multiple testing problem.

3.1 Investment Strategies

3.1.1 Data Description: CAPIQ and 18,000 Anomalies

We start with Standard and Poor's (S&P) Capital IQ database that covers a broad set of "alpha" strategies. This database has the historical performance of synthetic long-short strategies, which are catalogued into eight groups based on the types of their risk exposures

[13] This is the case for our application since we test the one-sided hypothesis where the alternative hypothesis is that the strategy mean is positive. We do not require such a restriction on p_0 if the hypothesis test is two-sided.

[14] While imposing the constraint that p_0 being less than the fraction of strategies with a positive mean in the original sample cannot in theory rule out bootstrapped samples for which this condition is violated, we find zero occurrences of these violations in our results since p_0 is set to be below the fraction of strategies with a positive mean by a margin. One element that we believe helps prevent these violations is the stability of order statistics, which ensures that the fraction of strategies with a positive mean in bootstrapped samples is fairly close to that in the original sample.

(e.g., market risk) or the nature of the forecasting variables (e.g., firm characteristics). Many well-known investment strategies are covered by this database, e.g., CAPM beta (Capital Efficiency Group), value (Valuation Group), and momentum (Momentum Group). For our purposes, we study 484 strategies (i.e., long-short strategies) for the U.S. equity market from 1985 to 2014.[15]

The CAPIQ data contains a large number of 'significant' investment strategies based on single hypothesis tests, which distinguishes it from the other datasets that we consider later (i.e., the anomaly data or mutual funds). This is not surprising as CAPIQ is biased towards a select group of strategies that are known to perform well. As such, it is best to treat these data as an illustration of how our method helps address the concern about Type II errors. One caveat for using the CAPIQ data (or other data that include a collection of investment strategies, such as Hou, Xue, and Zhang, 2017) is the selection bias in the data that is mainly due to publication bias. We do not attempt to address publication bias in this paper. See Harvey, Liu, and Zhu (2016) for a parametric approach in dealing with publication bias.

The second dataset is the 18,113 anomalies studied in Yan and Zheng (2017). Yan and Zheng construct a comprehensive set of anomalies based on firm characteristics.[16] Using the Fama and French (2010) approach to adjust for multiple testing, they claim that a large portion of anomalies are true and declare that there is widespread mispricing. We revisit the inference problem faced by Yan and Zheng (2017). Our goal is to use our framework to calibrate the error rates for their approach and offer insights on how many anomalies are true for their data — and we reach a different conclusion.

Overall, the two datasets resemble the two extreme cases for a pool of investment strategies that researchers or investors seek to analyze. While CAPIQ data consist of a select set of backtested strategies, the 18,000 anomalies include a large collection of primitive strategies obtained through a data mining exercise.

3.1.2 Preliminary Data Analysis

We use the t-statistic to measure the statistical significance of an investment strategy. For the CAPIQ data, we simply use the t-statistic of the strategy return.[17] For the 18,113 anomalies

[15] We thank S&P CapitalIQ for making these data available to us. Harvey and Liu (2017) also use the CAPIQ data to study factor selection.

[16] We thank Sterling Yan for providing us with their data on anomaly returns.

[17] Our results for the CAPIQ data are similar if we adjust strategy returns using, say, the Fama-French-Carhart four-factor model.

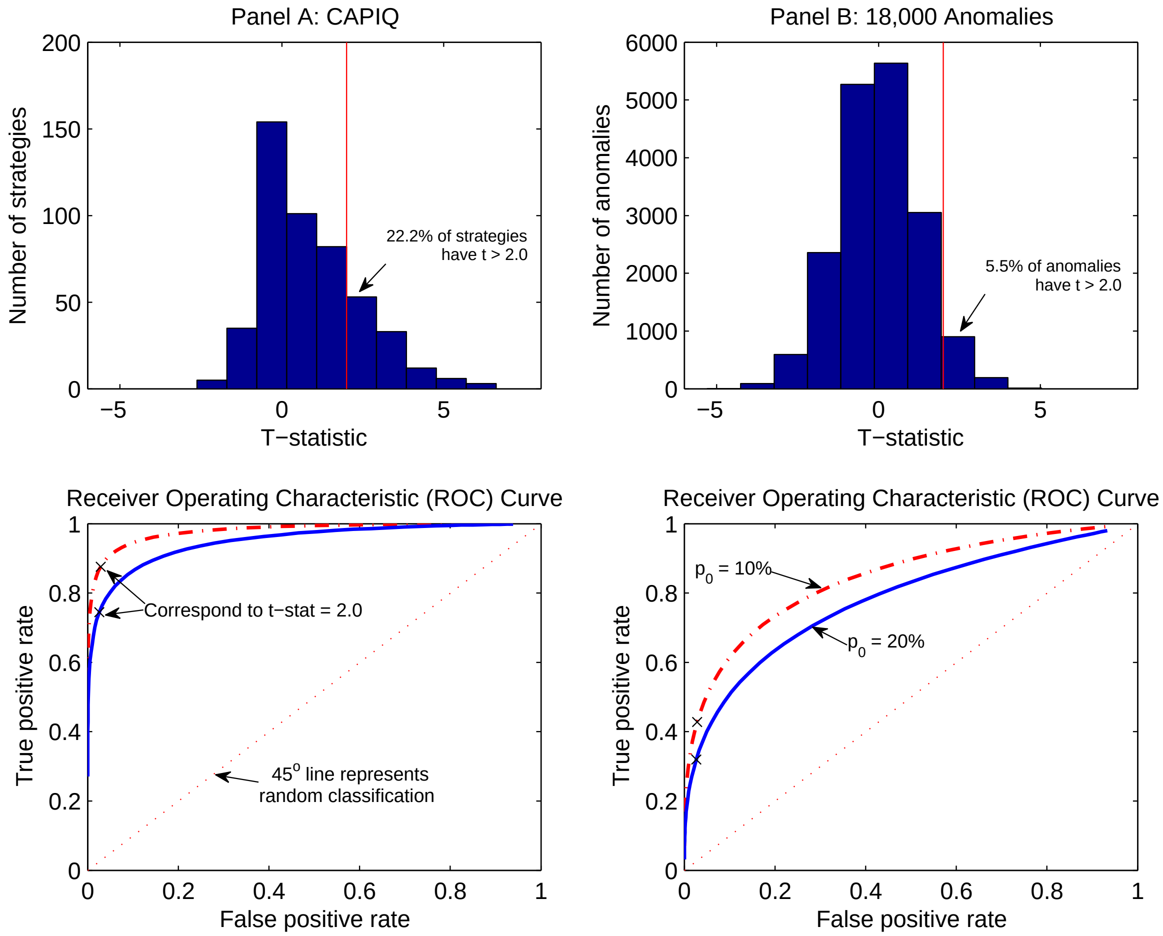
in Yan and Zheng (2017), we follow their procedure and calculate excess returns with respect to the Fama-French-Carhart four-factor model.[18] Yan and Zheng construct their strategies with firm characteristics and these strategies likely have high exposures to existing factors. While we focus on simple t-statistics to describe our framework in the previous section, our bootstrap-based method can be easily adapted to incorporate any benchmark factors.[19]

In Figure 1, the top figure in Panel A (Panel B) shows the t-statistic distribution for the CAPIQ (18,000 anomalies) data. For the CAPIQ data, the distribution is skewed to the right. The fraction of strategies that have a t-statistic above 2.0 is 22.1% (=107/484). For the 18,000 anomalies, the distribution is roughly symmetric around zero and has a fraction of 5.5% (=989/18,113) of t-statistics that fall above 2.0. These statistics are consistent with how these two datasets are assembled.

[18] Yan and Zheng (2017) use several specifications of the benchmark model, including the Fama-French-Carhart four-factor model. We only focus on the Fama-French-Carhart four-factor model to save space as well as to illustrate how benchmark models can be easily incorporated into our framework. Our results are qualitatively similar if we use alternative benchmark models.

[19] To preserve cross-sectional correlations, factors need to be resampled simultaneously with strategy returns when we resample the time periods.

Figure 1: **Preliminary Data Analysis**



T-statistic Distributions and the Receiver Operating Characteristic (ROC) curves for CAPIQ and 18,000 anomalies. We plot the t-statistic distributions (the top figures in Panel A and B) for the 484 investment strategies in the CAPIQ data and the 18,113 strategies in Yan and Zheng (2017). The t-statistic is calculated as the t-statistic for the original strategy return for the CAPIQ data and the t-statistic for the Fama-French-Carhart four-factor model adjusted anomaly alpha for the 18,113 strategies. The bottom figures in Panel A and B show the ROC curves, corresponding to $p_0 = 10\%$ and 20%. The crosses on the ROC curves mark the points that correspond to a t-statistic cutoff of 2.0.

Notice that the direction of the long-short strategies for the 18,113 anomalies data is essentially randomly assigned, which explains the symmetry of the distribution of the t-statistics. This also suggests that we should perform two-sided hypothesis tests as both significant outperformance and underperformance (relative to benchmark factor returns) qualifies as evidence of anomalous returns. However, to be consistent and comparable with our results for CAPIQ, we illustrate our method through one-sided tests that only test for outperformance. Our method can be straightforwardly applied to the case of two-sided tests. Moreover, given the data mining nature of the 18,113 anomalies data (so the two tails of the t-statistic distribution for anomalies are roughly symmetric), the Type I (Type II) error rates for one-sided tests under p_0 are approximately one-half of the Type I (Type II) error rates under $2p_0$. While we only present results for one-sided tests, readers who are interested in two-sided tests can apply the above transformation to obtain the corresponding error rates for two-sided tests.

We illustrate how to use our method to create the Receiver Operating Characteristic (ROC) curve, which is an intuitive diagnostic plot to assess the performance of a classification method (e.g., a multiple testing method).[20] It plots the True Positive Rate (TPR, defined as the number of true discoveries over the total number of true strategies) against the False Positive Rate (FPR, defined as the number of false discoveries over the total number of zero-mean strategies).

Our framework allows us to use bootstrapped simulations to draw the ROC. In particular, for a given p_0, the first-round bootstrap of our method classifies all strategies into true strategies and zero-mean strategies. The second-round bootstrap then calculates the realized TPR and FPR for each t-statistic cutoff for each bootstrapped simulation. We simulate many times to generate the average TPR and FPR across simulations.

Note that our previously defined FDR is different from FPR. Although the numerator is the same (i.e., the number of false discoveries), the denominator is the total number of true strategies (FPR) versus the total number of discoveries (FDR).[21] Our framework is flexible in handling alternative error rate definitions that are deemed useful.

[20] See Fawcett (2006) and Hastie and Tibshirani (2009) for applications of the ROC method.

[21] Conceptually, FDR is a harsher error rate definition than FPR when there are a large number of false strategies and the signal-to-noise ratio is low in the data. In this case, a high t-statistic cutoff generates very few discoveries. But FDR could still be high since it is difficult to distinguish the good from the bad. In contrast, since the denominator for FPR is larger than that for FDR, FPR tends to be much lower than FDR.

Figure 1 also shows the ROC for $p_0 = 10\%$ and 20%. On the ROC graph, the ideal classification outcome is given by the point $(0, 1)$, i.e., $FPR = 0$ and $TPR = 100\%$. As a result, a ROC curve that is closer to $(0, 1)$ (or further away from the 45-degree line which denotes random classification) is deemed better. We see two patterns from Figure 1. First, the ROC curve for the CAPIQ data is better than that for the 18,000 anomalies. Second, for each dataset, a smaller p_0 results in a better ROC curve.

These two patterns reflect key features of the two datasets we analyze. The better ROC curve for CAPIQ data stems from the higher average signal-to-noise ratio in the data since it contains select investment strategies. Although the higher signal-to-noise ratio can also be seen from the distribution of t-statistics (i.e., the top part of Figure 1), the ROC curve quantifies this through FPR and TPR, which are potentially more informative metrics to examine to classify investment strategies.[22] On the other hand, a smaller p_0 results in a more selective group of strategies (e.g., the average t-statistic is higher for a smaller p_0), which explains the better classification outcome as illustrated by the ROC curve. The ROC curve highlights the tradeoff between FPR and TPR for different levels of p_0.[23]

The ROC analysis provides a tradeoff between FPR and TPR for the two datasets we study. However, the ROC framework treats Type I and Type II errors symmetrically. Our approach can easily be adapted to allow for asymmetric error costs.

3.1.3 The Selection of Investment Strategies: Having a Prior on p_0.

We first apply our method to study how the Type I and Type II error rates vary across different t-statistic thresholds. In practice, researchers often use a pre-determined fixed t-statistic threshold to perform hypothesis testing, such as 2.0 at the 5% significance level for a single hypothesis test, or a t-statistic that exceeds 3.0 based on Harvey, Liu, and Zhu (2016).[24] Using our method, we investigate the implications of these choices.

[22] Another benefit is that the generated ROC in our framework takes test correlations into account (since, similar to Fama and French (2010), our second-stage bootstrap generates the same resampled time periods across all strategies), whereas the t-statistic distribution can not.

[23] Note that it is straightforward to find the optimal FPR (and the corresponding t-statistic cutoff) associated with a certain tradeoff between FPR and TPR. For example, if we equally weight the FPR and the TPR (i.e., we try to maximize TPR − FPR), then the optimal FPR is given by the tangency point of a 45-degree tangent line to the ROC curve.

[24] Notice that it was never the intention of Harvey, Liu and Zhu (2016) to recommend the 3.0 threshold as a universal rule that applies to any dataset. In fact, the very purpose of our current paper is to show how one can use our method to calibrate the Type I and Type II error rates based on different t-statistic thresholds, through which one can obtain the "optimal" t-statistic threshold that applies to the particular dataset at hand (also see Harvey, 2017).

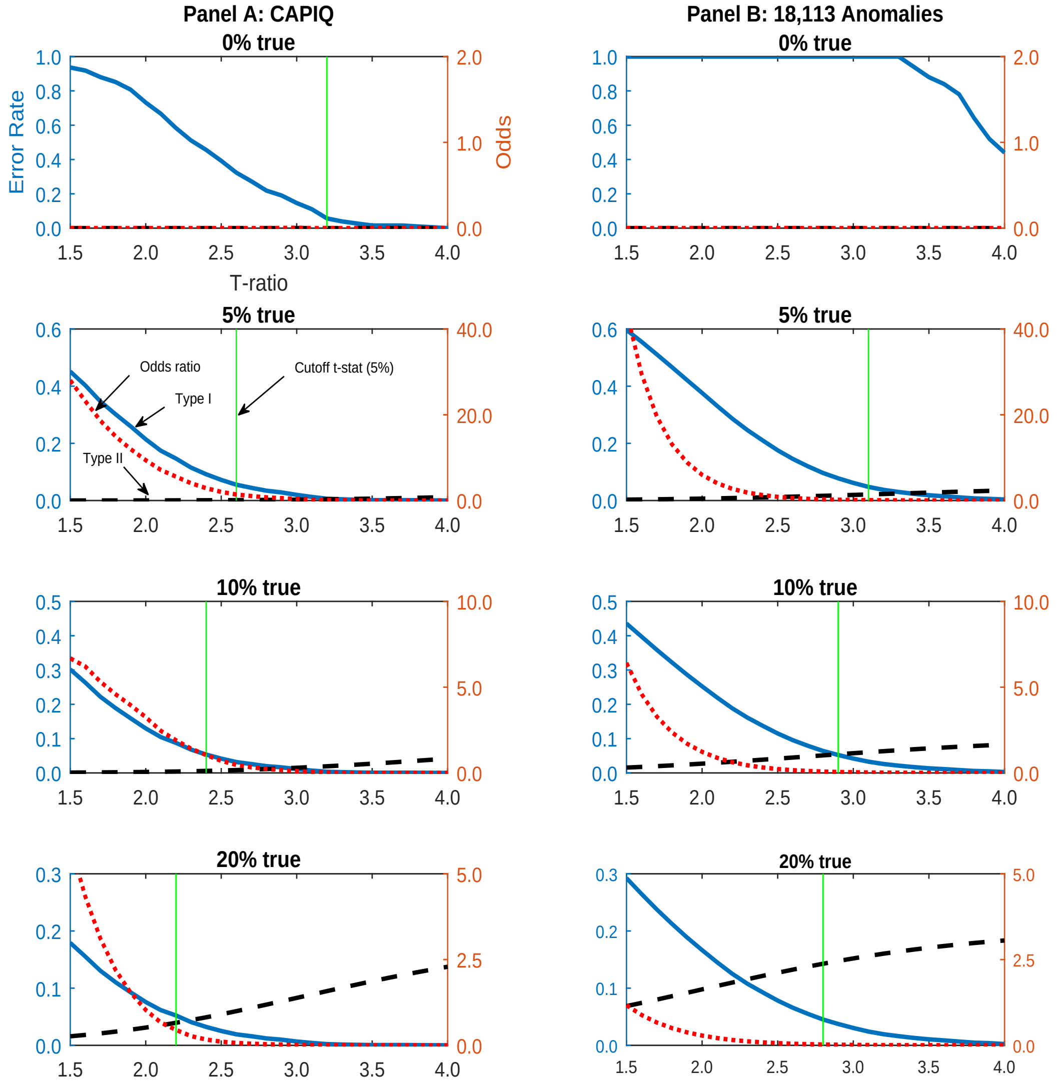
Figure 2 shows the error rates across a range of t-statistics for both the CAPIQ data and 18,000 anomalies. We see the classical tradeoff between the Type I and the Type II error rates in single hypotheses tests manifested in a multiple testing framework. When the threshold t-statistic increases, the Type I error rate (the rate of false discoveries among all discoveries) declines while the Type II error rate (the rate of misses among all non-discoveries) increases. In addition, the odds ratio, which is the ratio of false discoveries to misses, also decreases as the threshold t-statistic increases. We also highlight the threshold t-statistic that exactly achieves a 5% significance in Figure 2.[25] We see that this threshold t-statistic decreases as p_0 (the prior fraction of true strategies) increases. This makes sense as a higher prior fraction of true strategies calls for a more lenient t-statistic threshold.

How should an investment manager pick strategies based on Figure 2? First of all, the manager needs to elicit a prior on p_0, which is likely driven by both her previous experience and the data (e.g., the histogram of t-statistics we show in the last section). Suppose that the manager believes that a p_0 of 10% is plausible for the CAPIQ data. If she wants to control the Type I error rate at 5%, then she should set the t-statistic threshold at $t = 2.4$ (see Panel A in Figure 2, $p_0 = 10\%$). Based on this t-statistic threshold, 18% of strategies survive based on the original data.

Alternatively, under the same belief on p_0, suppose the investment manager is more interested in balancing Type I and Type II errors and wants to achieve an odds ratio around 1/5, that is, there are on average five misses for each false discovery (or, alternatively, she believes that the cost of a Type I error is five times that of a Type II error). Then she should set the t-statistic threshold at $t = 2.6$ if she was interested in $p_0 = 10\%$ and 15% of the strategies survive.

[25] The threshold t-statistic is 4.9 for the 18,000 anomalies when $p_0 = 0$. Since we set the range of the t-statistic (i.e., x-axis) to be from 1.5 to 4.0, we do not display this t-statistic in Figure 2.

Figure 2: **Error Rates for Fixed T-statistic Thresholds: CAPIQ and 18,113 Anomalies**



Simulated Type I and Type II error rates for CAPIQ and the 18,113 anomalies. For each p_0 of the fraction of strategies that are believed to be true, we follow our method in Section 2 and set $I = 100$ (for each i, we bootstrap to obtain the ranking of strategies and set the top p_0 as true) and $J = 1,000$ (conditional on i, for each j, we bootstrap the time periods) to run in total 100,000 (= $100 \times 1,000$) bootstrapped simulations to calculate the empirical Type I (fraction of false discoveries) and Type II (fraction of missed discoveries among all non-discoveries) error rates. On the graph, the (blue) solid line is the Type I error rate, the (black) dashed line is the Type II error rate, the (red) dotted line is the odds ratio. The left axis is the error rate that applies to the Type I and Type II error rate, whereas the right axis is the odds ratio. The vertical (green) line marks the t-statistic cutoff that corresponds to a Type I error rate of 5%. The t-statistic cutoff for the top figure in Panel B (4.9) is omitted.

Therefore, under either the Type I error rate or the odds ratio, neither 2.0 (the usual cutoff for 5% significance) nor 3.0 (based on Harvey, Liu, and Zhu, 2016) is optimal from the perspective of the investor. The preferred choices lie in between 2.0 and 3.0 for the examples we consider and depend on both p_0 and the CAPIQ data that we study.

Comparing Panel A and B in Figure 2, at $p_0 = 0$, the much higher t-statistic cutoff for the 18,113 anomalies (i.e., 4.9) than that for CAPIQ (i.e., 3.2) reflects the larger number of strategies for the anomaly data (thresholds need to be higher when more strategies are tested, i.e., the multiple testing problem is more severe). For alternative values of p_0, the higher t-statistic cutoff for the anomaly data is driven by its low signal-to-noise ratio, which is also seen from the analysis of the ROC curve.

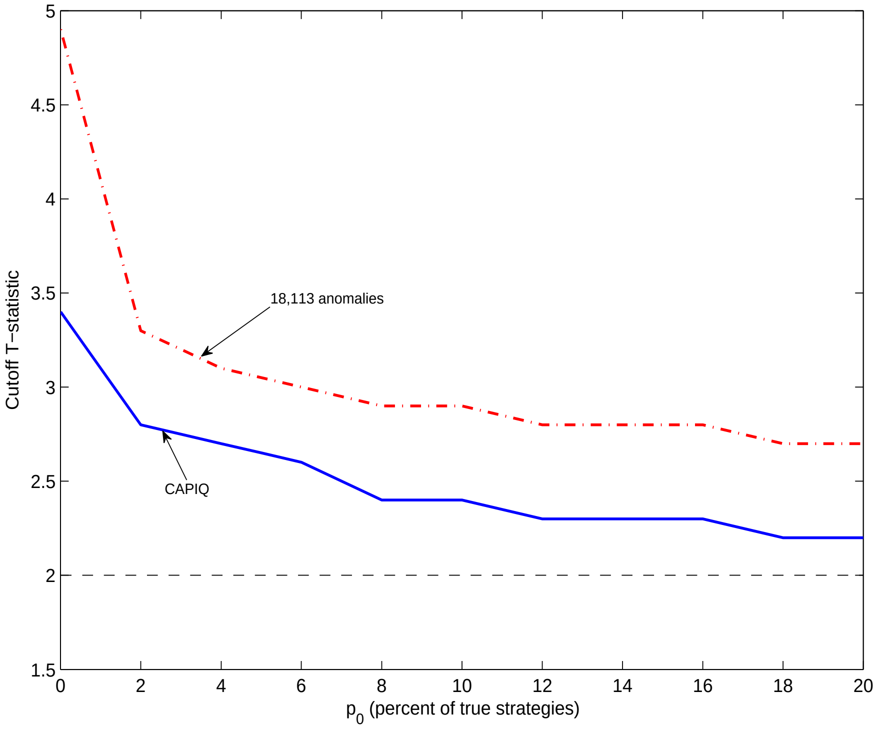
The overall message of Figure 2 is that two ingredients need to be taken into account to select the best strategies. One ingredient is investors' prior on the fraction of true strategies. We are all subjective Bayesians so it is only natural for us to try to incorporate important insights from the Bayesian approach into a multiple testing framework.[26] Our method allows us to perform a simple sensitivity analysis of the error rates to changes in the prior fraction of true strategies. Instead of trying to control the error rates under all circumstances as in the traditional frequentist hypothesis testing framework, our view is that we should incorporate our prior beliefs into the decision making. This helps us make informed decisions based on the tradeoff between the Type I and the Type II errors.

The other ingredient is the particular data at hand and the hypothesis that is being tested. While many traditional multiple testing adjustments work under general assumptions on the dependence structure in the data, they may be too conservative in that they underreject when the null hypothesis is false, leading to too many Type II errors (as we shall see later). Our bootstrap-based framework provides a convenient approach to generate t-statistic thresholds that are calibrated to the particular decision being considered and the particular data under analysis.

Figure 3 plots the cutoff t-statistics that generate a Type I error rate of 5% against p_0. In general, the cutoff t-statistic declines as p_0 becomes larger, since it is less likely for a discovery to be false when a larger fraction of strategies are true.

[26] See Harvey (2017) for the application of the Bayesian approach to hypothesis testing.

Figure 3: **Cutoff T-statistics as a Function of p_0 for CAPIQ and 18,113 Anomalies**



Cutoff t-statistics as a function of p_0 for CAPIQ investment strategies and the 18,113 anomalies. For each hypothetical level of p_0 between 0% and 20% (with 1% increments), we search for the optimal cutoff t-statistic between 1.5 and 5.0 (with 0.1 increments) that corresponds to a Type I error rate of 5%. The kinks in the graph are caused by the discrete increments in both p_0 and the range of t-statistics from which we search from.

3.1.4 The Selection of Investment Strategies: Unknown p_0

When p_0 is unknown, our framework can be used to evaluate the performance of multiple testing corrections across different values of p_0. While multiple testing methods are designed to control the false discovery rate at the appropriate level regardless of the value of p_0, their performance in finite samples and, especially, for the particular data under analysis, is unknown. Our method provides guidance on which method to use for a particular data set.

Given our focus on the false discovery rate, we implement several popular multiple testing adjustments proposed in the statistics literature that aim to control the expected false discovery rate. Our goal is not to compare all existing methods. Rather, we choose a few representative methods in terms of their theoretical properties and illustrate how to apply our framework to select the best method for the particular data being analyzed.

We want to emphasize that our data-driven approach is different from the usual simulation exercises carried out by studies that propose multiple testing methods. Most simulation studies make simple assumptions on the underlying data structure (e.g., a multivariate normal distribution with a certain correlation structure).[27] However, the usual data sets we encounter in financial economics (e.g., the cross-section of stock returns) have important features (e.g., missing data, cross-sectional dependence, tail dependence, etc.) that may make simplifying assumptions poor approximations. In our context of multiple testing, many multiple testing methods are known to be sensitive to some of these features (i.e., cross-sectional dependence and tail dependence). Therefore, it is important to customize the performance evaluation of a multiple testing method to the particular data being examined.

The first set of multiple testing adjustments that we consider are BH (Benjamini and Hochberg, 1995), which controls the expected false discovery rate under the pre-specified value if tests are mutually independent, and BY (Benjamini and Yekutieli, 2001), which controls the expected false discovery rate under arbitrary dependence of the tests.[28] As such, BH may not work if tests are correlated in a certain fashion while BY tends to be overly conservative (i.e., too few discoveries).[29]

Given our focus on test power, we also consider Storey (2002), which sometimes presents an improvement over BH and BY in terms of test power. There is a plug-in parameter in Storey (2002) (called θ) that helps replace the total number of tests in BH and BY with an estimate of the fraction of true null hypotheses. Bajgrowicz and Scaillet (2012) suggest $\theta = 0.6$. We experiment with three values for θ: $\theta = 0.4$, 0.6, and 0.8.

Lastly, we consider a strand of multiple testing methods that have strong theoretical properties in terms of the requirement on the data to achieve a pre-specified error rate.[30] In particular, we implement Romano, Shaikh, and Wolf (RSW, 2008), who propose a bootstrap-based approach that controls the false discovery rate asymptotically under arbitrary dependence in the data.[31] Our goal is to analyze the finite-sample performance of RSW for the two data sets we have. However, the implementation of RSW is computationally challenging (espe-

[27] See, e.g., the simulation studies in Romano and Wolf (2005) and Romano, Shaikh, and Wolf (2008).

[28] Benjamini and Yekutieli (2001) show that independence can be replaced by a weaker assumption that is defined as positive regression dependency as in Benjamini and Hochberg (1995).

[29] For further details on these two methods and other multiple testing adjustments as well as their applications in finance, see Harvey, Liu, and Zhu (2016) and Chordia, Goyal, and Saretto (2018).

[30] We thank Laurent Barras for bringing this literature to our attention.

[31] See Romano and Wolf (2005) for a companion approach that controls the family-wise error rate.

cially for the 18,000 anomalies) as it requires the estimation of $B \times O(M^2)$ regression models (where B is the number of bootstrapped iterations and M is the total number of tests) to derive a single t-statistic cutoff. Therefore, we randomly sample from the 18,000 anomalies to reduce the sample size and apply RSW to the sub-samples. However, this makes our results not directly comparable with those that examine the full sample. We therefore present our results on RSW in Appendix A.

Besides the aforementioned methods, we also report the Type II error rates derived under the assumption that p_0 is known ex ante and the t-statistic cutoff is fixed. In particular, for a given p_0, our previous analysis allows us to choose the t-statistical threshold such that the pre-specified Type I error rate is exactly achieved, thereby minimizing the Type II error rate of the test.[32] This minimized Type II error rate is a useful benchmark for us to gauge the performance of other multiple testing methods. Note that other multiple testing methods may generate a smaller Type II error rate than the optimal rate because their Type I error rates may exceed the pre-specified levels.[33]

Before presenting our results, we briefly report summary statistics on the correlations among test statistics for our data sets. For CAPIQ, the 10^{th} percentile, median, and 90^{th} percentile for the pairwise test correlations are -0.377, 0.081, and 0.525; the corresponding numbers for the absolute values of the pairwise correlations are 0.040, 0.234, and 0.584. For the 18,000 anomalies, these numbers are -0.132, 0.003, and 0.145 (original values), and 0.013, 0.069, and 0.186 (absolute values). While correlations should provide some information on the relative performance of different methods, other features of the data may also be important, as we shall see later in our analysis.

Table 2 and 3 report the results for CAPIQ and the 18,000 anomalies, respectively. Focusing on Table 2, BH generates false discovery rates that are always below and oftentimes close to the pre-specified significance levels, suggesting its overall good performance. In comparison, BY is too conservative in that it generates too few discoveries, resulting in higher Type II error rates compared to BH. The three Storey tests in general fail to meet the pre-

[32] Test II error rate is minimized in our framework in the following sense. Suppose we are only allowed to choose a fixed t-statistic cutoff for each assumed level of p_0. Imagine that we try to solve a constrained optimization problem where the pre-specified significance level is the Type I error rate constraint and our objective function is to minimize the Type II error rate. Given the tradeoff between the Type I and Type II error rate, the optimal value for the objective function (i.e., the Type II error rate) is achieved when the constraint is met with equality.

[33] There is also a difference in the Type I error rate between a fixed t-statistic cutoff and multiple testing methods. Using our notation from the previous section (i.e., X_i denotes a particular parameterization of the hypotheses), while multiple testing methods seek to control the false discovery rate for each parameterization (i.e., X_i) of the hypotheses (i.e., $E(FDR|X_i)$), the fixed t-statistic cutoff we used previously aims to control $E(E(FDR|X_i)|p_0)$, which averages $E(FDR|X_i)$ across different realizations of X_i for a given p_0.

specified significance levels, although the Type I error rates generated by Storey ($\theta = 0.4$) are reasonably close. Overall, BH seems to be the preferred choice for CAPIQ among the five tests we examine. Storey ($\theta = 0.4$) is also acceptable, although one has to be cautious about the somewhat higher Type I error rates than the desired levels.

The test statistic from our method (i.e., the last column in Table 2) provides gains in test power compared to BH, BY and Storey. In particular, compared to other methods that also meet the pre-specified significance levels (i.e., BH and BY as in Table 2), the Type II error rate for this test statistic is uniformly lower. For example, at $p_0 = 20\%$ and a significance level of 10%, all three Storey methods are over-sized (Type I error $> 10\%$). When we perform power comparison, we leave out the Storey methods and focus on the other two models (i.e., BH and BY). Between these two, BH is the better performing model since its Type I error rate is closer to the pre-specified significance level. We therefore use it as the benchmark to evaluate the power improvement. Comparing with BH which generates a Type II error rate of 3.5%, our model produces an error rate of 2.9%.

For the 18,113 anomalies, results in Table 3 present a different story than Table 2. In particular, BH generally performs worse than the three Storey tests in controlling the false discovery rate when p_0 is not too large (i.e., $p_0 \leq 10\%$), although the Storey tests also lead to higher Type I error rates than desired. Overall, BY should be the preferred choice to strictly control the Type I error rate at the pre-specified level, although it appears to be far too conservative. Alternatively, Storey ($\theta = 0.8$) dominates the other methods in achieving the pre-specified significance levels when $p_0 \leq 10\%$.

Our results in Appendix A show the finite-sample performance of RSW applied to our data sets. Despite the strong theoretical appeal of RSW, it often leads to a higher Type I error rate than the desired level for both data sets. In fact, compared to the aforementioned multiple testing methods we consider, RSW oftentimes generate the largest distortion in test size when p_0 is relatively small (i.e., $p_0 \leq 20\%$).

Our results highlight the data-driven nature of the performance of alternative multiple testing methods. While BH is shown to work well theoretically when tests are independent, it is not clear how departures from independence affects its performance. Interestingly, our results show that BH performs much worse for the 18,000 anomalies than for CAPIQ, although the data for the 18,000 anomalies appears more "independent" than CAPIQ based on the summary statistics for test correlations.[34] As another example, we show that Storey ($\theta = 0.4$)

[34] Note that we use the average correlation among strategies as an informal way to gauge the degree of cross-sectional dependence. However, correlation likely provides an insufficient characterization of depen-

and Storey ($\theta = 0.8$) could be the preferred choice for the two data sets we examine, whereas $\theta = 0.6$ is the value recommended by Bajgrowicz and Scaillet (2012). Finally, methods that are guaranteed to perform well asymptotically may have a poor finite-sample performance, as we show for RSW in Appendix A.

Overall, our method provides insight on which multiple testing adjustment performs the best for a given data set. Of course, it is also possible to use our method directly to achieve a level of false discovery and to optimize the power across different assumptions for p_0.

dence because certain forms of dependence may not be captured by the correlation. In our context, there may be important departures from independence in the 18,000 anomalies data that impact the performance of BH but are not reflected in the average correlation. Our results thus highlight the data-dependent nature of existing multiple testing methods.

Table 2: **Error Rates for Existing Methods: CAPIQ**

Simulated Type I and Type II error rates for CAPIQ data. For each p_0 of the fraction of strategies that are believed to be true, we follow our method in Section 2 and set $I = 100$ (for each i, we bootstrap to obtain the ranking of strategies and set the top p_0 as true) and $J = 1,000$ (conditional on i, for each j, we bootstrap the time periods) to run in total 100,000 (= $100 \times 1,000$) bootstrapped simulations to calculate the empirical Type I and Type II error rates. For a given significance level α, we set the Type I error rate at α and find the corresponding Type II error rate, which is the "optimal" Type II error rate. We also implement BH (Benjamini and Hochberg, 1995), BY (Benjamini and Yekutieli, 2001), and Storey (Storey, 2002) for our simulated data and calculate their respective Type I and Type II error rates.

		Type I					Type II					
p_0	α	BH	BY	Storey			BH	BY	Storey			$HL(opt)$*
				$(\theta = 0.4)$	$(\theta = 0.6)$	$(\theta = 0.8)$			$(\theta = 0.4)$	$(\theta = 0.6)$	$(\theta = 0.8)$	
(frac. of true)	(sig. level)											
2%	1%	0.010	0.002	0.023	0.022	0.021	0.001	0.002	0.001	0.001	0.001	0.001
	5%	0.044	0.009	0.058	0.058	0.058	0.000	0.001	0.000	0.000	0.000	0.000
	10%	0.086	0.013	0.102	0.104	0.107	0.000	0.001	0.000	0.000	0.000	0.000
5%	1%	0.008	0.002	0.018	0.018	0.017	0.006	0.011	0.005	0.005	0.005	0.005
	5%	0.047	0.007	0.060	0.060	0.060	0.003	0.006	0.002	0.002	0.002	0.002
	10%	0.091	0.012	0.102	0.105	0.113	0.002	0.005	0.001	0.001	0.002	0.001
10%	1%	0.008	0.002	0.017	0.017	0.016	0.022	0.037	0.021	0.021	0.020	0.021
	5%	0.046	0.006	0.058	0.059	0.063	0.011	0.024	0.010	0.010	0.010	0.010
	10%	0.087	0.014	0.108	0.112	0.123	0.007	0.019	0.007	0.007	0.006	0.007
20%	1%	0.008	0.002	0.016	0.016	0.016	0.082	0.117	0.088	0.077	0.074	0.079
	5%	0.042	0.006	0.061	0.064	0.071	0.049	0.088	0.044	0.043	0.041	0.047
	10%	0.079	0.012	0.117	0.123	0.142	0.035	0.074	0.030	0.029	0.028	0.029
30%	1%	0.007	0.002	0.017	0.017	0.017	0.179	0.223	0.161	0.159	0.154	0.153
	5%	0.037	0.005	0.064	0.068	0.078	0.126	0.188	0.104	0.101	0.096	0.095
	10%	0.069	0.011	0.119	0.129	0.155	0.098	0.168	0.076	0.073	0.068	0.066

*Type II calculated at optimized Type I error = α

Table 3: **Error Rates for Existing Methods: 18,000 Anomalies**

Simulated Type I and Type II error rates for the 18,000 anomalies data. For each p_0 of the fraction of strategies that are believed to be true, we follow our method in Section 2 and set $I = 100$ (for each i, we bootstrap to obtain the ranking of strategies and set the top p_0 as true) and $J = 1,000$ (conditional on i, for each j, we bootstrap the time periods) to run in total 100,000 (= $100 \times 1,000$) bootstrapped simulations to calculate the empirical Type I and Type II error rates, as well as the odds ratio. For a given significance level α, we set the Type I error rate at α and find the corresponding Type II error rate, which is the "optimal" Type II error rate. We also implement BH (Benjamini and Hochberg, 1995), BY (Benjamini and Yekutieli, 2001), and Storey (Storey, 2002) for our simulated data and calculate their respective Type I and Type II error rates.

		Type I					Type II					
p_0	α	BH	BY	Storey			BH	BY	Storey			$HL(opt)$*
				($\theta = 0.4$)	($\theta = 0.6$)	($\theta = 0.8$)			($\theta = 0.4$)	($\theta = 0.6$)	($\theta = 0.8$)	
(frac. of true)	(sig. level)											
2%	1%	0.018	0.003	0.019	0.019	0.018	0.009	0.013	0.008	0.008	0.008	0.011
	5%	0.077	0.008	0.072	0.072	0.070	0.005	0.010	0.005	0.005	0.005	0.007
	10%	0.143	0.018	0.134	0.133	0.131	0.004	0.009	0.004	0.004	0.004	0.005
5%	1%	0.017	0.002	0.018	0.018	0.017	0.030	0.040	0.030	0.030	0.030	0.031
	5%	0.072	0.008	0.069	0.069	0.067	0.020	0.033	0.020	0.020	0.020	0.021
	10%	0.133	0.017	0.125	0.124	0.122	0.015	0.030	0.016	0.016	0.016	0.017
10%	1%	0.016	0.002	0.017	0.017	0.016	0.074	0.089	0.075	0.075	0.075	0.074
	5%	0.067	0.008	0.067	0.067	0.065	0.055	0.080	0.056	0.055	0.055	0.054
	10%	0.122	0.016	0.121	0.121	0.120	0.044	0.074	0.045	0.045	0.045	0.045
20%	1%	0.014	0.002	0.016	0.016	0.016	0.166	0.187	0.169	0.169	0.169	0.171
	5%	0.058	0.008	0.063	0.063	0.063	0.135	0.175	0.138	0.137	0.137	0.143
	10%	0.105	0.014	0.115	0.116	0.117	0.115	0.167	0.116	0.116	0.115	0.121
30%	1%	0.013	0.002	0.015	0.016	0.016	0.270	0.289	0.267	0.267	0.267	0.274
	5%	0.051	0.007	0.060	0.061	0.062	0.238	0.278	0.231	0.230	0.229	0.235
	10%	0.092	0.013	0.110	0.113	0.116	0.214	0.270	0.203	0.201	0.200	0.210

*Type II calculated at optimized Type I error = α

3.1.5 Revisiting Yan and Zheng (2017)

Applying the preferred methods based on Table 3 to the 18,000 anomalies, the fraction of true strategies is found to be 0.000% (BY) and 0.015% (Storey, $\theta = 0.8$) under 5% significance level, and 0.006% (BY) and 0.091% (Storey, $\theta = 0.8$) under a 10% significance level.[35]

Our results suggest that only about 0.1% (note this is still a fairly large number, 18) of the 18,000 anomaly strategies in Yan and Zheng (2017) are classified as "true" to control the false discovery rate at 10%. In contrast, Yan and Zheng claim that "a large number" of these strategies are true based on the Fama-French approach. In particular, they suggest that the 90^{th} percentile of t-statistics is significant, implying that at least 10% of anomalies, i.e., 1,800, are generating significant positive returns. They conclude there exists widespread mispricing.[36]

Why are our results so different from Yan and Zheng (2017)? First of all, it is important to realize that the multiple testing methods we have used so far (which do not include the Fama-French approach that Yan and Zheng employ) have a different objective function than Yan and Zheng's. In particular, while Yan and Zheng are interested in finding an unbiased estimate of the fraction of true discoveries, multiple testing adjustments seek to control the false discovery rate at a certain level and there will be many omissions (true strategies missed) when the hurdle is very high. As a result, it is to be expected that the fraction of true anomalies declared by multiple testing methods will be somewhat smaller than the true value. However, this is unlikely to fully explain the stark contrast between our inference and Yan and Zheng's estimate, i.e., two orders of magnitude. Using our framework, we show that their application of the Fama-French approach is problematic.

More specifically, we show that the Fama-French approach is not designed to estimate the fraction of true strategies. As we will detail in the next section, the Fama-French approach singularly focuses on whether or not the entire population of strategies has a zero mean return. When the null hypothesis (i.e., the entire population has a zero mean) is rejected, all we know is that some strategies are true but we cannot tell how many are true. While we focus on the misapplication of the Fama-French approach in estimating the fraction of true strategies in this section, we defer the analysis of its test power to the next section.

[35] Across different values of p_0, BY dominates BH and Storey ($\theta = 0.8$) dominates the other two Storey methods in achieving the pre-specified significance levels. We therefore focus on BY and Storey ($\theta = 0.8$). The statistics in this paragraph are based on unreported results.

[36] In earlier versions of the paper, they find that even the 70^{th} percentile is significant (which is confirmed by our analysis), suggesting that the top 30% of anomalies are true.

To apply our framework, we assume that a fraction of p_0 of strategies are true for their anomalies. Next, we define the test statistic that describes the inference approach that is implicitly used in Yan and Zheng (2017). Let $Frac_i$ be the $(100-i)^{th}$ percentile of the cross-section of t-statistics and let $p(Frac_i)$ be the p-value associated with $Frac_i$ based on the bootstrapping approach in Fama and French (2010). Define $Frac$ as the maximum I in $(0, 0.4)$ such that $p(Frac_i) \leq 0.05$ is true for all $i \in (0, I)$.[37] In other words, I is the maximum fraction such that all percentiles above $100(1-I)$ are rejected at the 5% level based on the Fama-French approach. We set the upper bound at 0.4 because it is unlikely that more than 40% of their anomalies generate significantly positive returns given the distribution of t-statistics shown in Figure 1. Changing the upper bound to alternative values does not qualitatively affect our results.

Table 4: **Diagnosing Yan and Zheng (2017): Summary Statistics on the Fraction of Anomalies Identified**

Simulated fraction of anomalies identified by following the Fama-French approach in Yan and Zheng (2017). For each p_0 of the fraction of strategies that are believed to be true, we follow our method in Section 2 and set $I = 100$ (for each i, we bootstrap to obtain the ranking of strategies and set the top p_0 as true) and $J = 100$ (conditional on i, for each j, we bootstrap the time periods) to run in total 10,000 $(= 100 \times 1,00)$ bootstrapped simulations. For each bootstrap sample, we perform the Fama-French test and find the $Frac$, which is the maximum fraction (i) such that all percentiles above $100(1-i)$ (with 1% increment) are rejected at the 5% level. We set the upper bound of i at 40%.

	Summary Statistics for $Frac$			
p_0	Mean (%)	Stdev.(%)	Prob($Frac \geq 5\%$)	Prob($Frac \geq 10\%$)
0.5%	2.06	7.31	0.18	0.15
1.0%	4.04	9.04	0.23	0.21
2.0%	7.30	10.27	0.44	0.31
5.0%	20.19	13.36	0.99	0.83
10.0%	35.68	7.55	1.00	1.00
15.0%	39.62	1.76	1.00	1.00

Table 4 shows the summary statistics on $Frac$ for different levels of p_0. The simulated means of $Frac$ all exceed the assumed levels of p_0, suggesting that the approach taken by Yan and Zheng (2017) is biased upward in estimating the fraction of true anomalies. Focusing on the last column (i.e., the probability for $Frac$ to exceed 10%), when p_0 is only 0.5%, there is a

[37] More formally, $Frac = \max_{I\in(0,0.4)}\{\max\{p(Frac_i)\}_{i\leq I} \leq 0.05\}$.

15% probability for the Fama-French test statistic to declare all top 10% of anomalies true. In fact, if p_0 were 10% (as claimed by Yan and Zheng, 2017), the Fama-French test statistic would frequently declare more than 30% of anomalies true (as seen from a mean statistic of 35.68 that is close to the 40% upper bound).

The intuition for the above results is straightforward. Suppose $p_0 = 10\%$. Due to sampling uncertainty, not all 10% of true strategies will be ranked above the 90^{th} percentile. Given the large number of tests for more than 18,000 anomalies, strategies with a zero population mean may be lucky and effectively crowd out true strategies by being ranked higher than the 90^{th}-percentile in t-statistic for a given sample. As such, the 10% of true strategies affect not only the 90^{th} percentile of t-statistics and beyond, but also possibly lower percentiles. As a result, the Fama-French approach may detect significant deviations from the null hypothesis for a lower percentile, thereby overestimating the magnitude of the fraction of true strategies.

To summarize, Yan and Zheng (2017) misapplied the Fama-French approach to reach the conclusion that a large number (i.e., more than 1,800) of anomalies are true. Using multiple testing methods that are most powerful in detecting true anomalies (while controlling the false discovery rate at 10%), we find that a very small number (i.e., 18) are significant. While a further investigation of the economic underpinnings of these significant anomalies may further narrow the list of true anomalies, our results cast doubt on the possibility of discovering true anomalies through a pure data mining exercise, such as the one carried out in Yan and Zheng (2017).

3.2 A Simulation Study

We now perform a simulation study to evaluate the performance of our proposed method. Similar to our application when p_0 is unknown, we focus on the ability of our method in correctly ranking the performance of existing methods. We are particularly interested in how the cross-sectional alpha (or mean return) distribution among true strategies affects our results. We therefore examine a variety of cross-sectional distributions, which we denote as F.

We also aim to highlight the Type I versus Type II error tradeoff among existing multiple testing methods. However, since hypothesis testing is essentially a constrained optimization problem where we face Type I error constraint (i.e., Type I error rate is no greater than the pre-specified significance level) while seeking to maximize test power (i.e., one minus the Type II error rate), a tighter constraint (e.g., the pre-specified significance level is exactly met) leads to a lower Type II error rate. Therefore, it is sufficient to show how our method helps select the best-performing method in terms of meeting the Type I error rate constraint.

We set up our simulation study within the CAPIQ data.[38] To preserve the cross-sectional dependence in the data, we simulate strategy returns based on the actual CAPIQ data. We fix the fraction of true strategies at 10% throughout our analysis. However, our specification of p_0 does not have to equal 10% (i.e., correctly specified). We also study alternative values of p_0 to examine our model performance when p_0 is potentially misspecified.

We explore several specifications for F. They all take the form of a Gamma distribution with mean μ_0 and standard deviation σ_0.[39] We entertain three values for μ_0 ($\mu_0 = 2.5\%$, 5% and 10%) and three values for σ_0 ($\sigma_0 = 0$, 2.5%, and 5%).[40] (For example, $(\mu_0 = 5.0\%, \sigma_0 = 0)$ denotes a constant (i.e., a point mass distribution) at $\mu_0 = 5.0\%$.) In total, we study nine specifications for F.

There are several advantages to model the cross-sectional distribution F with a Gamma distribution. First, we focus on one-sided tests in this section so F should have a positive support. Second, moments are available analytically so we can easily change the mean (variance) while holding the variance (mean) constant to study the comparative statics. Lastly, a Gamma distribution has higher-moment characteristics (e.g., skewness and excess kurtosis) that allows it to capture salient features of F.[41]

Our simulation exercise is structured as follows. For the CAPIQ data, we first randomly select 10% of strategies and inject mean returns that are generated by the distribution F. For the remaining 90% of strategies, we set their mean returns at the null hypothesis (i.e., zero). Let D_m denote the final data, where m denotes the m-th iteration. D_m can be thought of as the population of strategy returns for the next step of the analysis.

[38] We focus on the CAPIQ data to save computational time. Our simulation exercise is computationally intensive. Each simulation run takes on average one day on a single core. The computing platform we have access to allows us to have 400 cores running at the same time. It therefore takes us one day to complete a specification (i.e., F) with 400 independent simulation runs. It would take us much longer for the data with 18,000 anomalies.

[39] The probability density function for a Gamma distributed variable X with mean μ_0 and standard deviation σ_0 is $f(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}$, where the shape parameter is $k = \frac{\mu_0^2}{\sigma_0^2}$, the scale parameter is $\theta = \frac{\sigma_0^2}{\mu_0}$, and $x > 0$.

[40] In the data, the mean return for the top 10% of strategies ranked in terms of the t-statistic (which is equivalent to the Sharpe ratio in our context) is 6.3%.

[41] For our parameterization of the Gamma distribution (mean μ_0 and standard deviation σ_0), skewness is $\frac{2\sigma_0}{\mu_0}$ and excess kurtosis is $\frac{6\sigma_0^2}{\mu_0^2}$.

We next perturb D_m to generate the in-sample data. In particular, we bootstrap the time periods once while keeping the cross-section intact to generate the in-sample data $D_{m,k}$, where the subscript k represents the round of iteration conditional on m. Note that each in-sample data $D_{m,k}$ is generated conditional on the same population D_m.

For each $D_{m,k}$, we follow a given multiple testing method (e.g., BH) at a pre-specified significance level α to generate the test outcomes. Since we know the true identities of strategies based on the step at the beginning when we create D_m, we are able to calculate the true realized error rates, which we denote as $FDR^a_{m,k}$ ('a' stands for 'actual'). This actual realized error will vary across the different multiple testing methods as well as the different nominal error rates. Implementing our method to $D_{m,k}$ — where we do not know the true strategies, we get the estimated error rates as $FDR^e_{m,k}$ ('e' stands for 'estimated').

We report three statistics in our results. They are the nominal error rate (i.e., δ, which is also the pre-specified significance level), the actual error rate ('Actual'), and our estimated error rate ('Est.'). The actual error rate and our estimated error rate are calculated as:

$$\begin{aligned}\text{Actual} &= \frac{1}{MK}\sum_{m=1}^{M}\sum_{k=1}^{K} FDR^a_{m,k},\\ \text{Est.} &= \frac{1}{MK}\sum_{m=1}^{M}\sum_{k=1}^{K} FDR^e_{m,k}.\end{aligned}$$

For a multiple testing method, the pre-specified significance level (i.e., δ) can be viewed as the stated significance level, which can be different from the true error rate. The difference between δ and 'Actual' therefore captures the bias in error rate committed by the method. This bias can be revealed through our approach if 'Est.' is reasonably close to 'Actual', in particular if 'Est.' is closer to 'Actual' than δ. If our approach provides good approximations to the true error rates across all specifications, then one can use our method to rank existing methods in terms of their performance with respect to the Type I error rate.

Table 5 reports the results when the cross-sectional distribution F is assumed to be a constant. Overall, our method presents a substantial improvement over the usual approach that takes the stated significance level as given. Our results show that 'Est.' is often much closer to 'Actual' than δ.

The performance of our method in approximating the true error rate improves when the mean parameter μ_0 is higher. Note that a higher μ_0 (hence, a higher signal-to-noise ratio in the data) does not necessarily lead to better performance for existing multiple testing methods. For example, at 10% significance level, while BH improves when μ_0 changes from 2.5% to 10% (as the true Type I error rate gets closer to 10%), Storey ($\theta = 0.8$) becomes worse as its true Type I error rate changes from 10.29% to 12.09%. Our approach performs better when μ_0 is higher (greater signal to noise ratio) because it is easier for our first-stage bootstrap to correctly identify the true outperforming strategies, leading to a more accurate calculation of error rates in the second-stage bootstrap. The above two reasons explain the larger improvement of our method over the usual approach when the signal-to-noise ratio is higher in the data.

Our estimation framework also appears to be robust to potential misspecification in p_0. There is some evidence that a correctly specified p_0 (i.e., $p_0 = 10\%$) leads to a better performance on average across all scenarios (i.e., across different methods, significance levels, and specifications of μ_0). However, for a particular specification, this may not be the case. For example, for BH and at 5% level, given a true error rate of 0.0472, a misspecified p_0 of 5% seems to perform somewhat better (Est. = 0.0465) than $p_0 = 10\%$ (Est. = 0.0459). Nonetheless, the performance of our method in approximating the true error rates is consistent across alternative specifications of p_0.

Table IA.1 and IA.2 in Internet Appendix A report the results under alternative specifications of the cross-sectional distribution F. Compared to Table 5, a larger dispersion in the mean return distribution (while keeping the mean constant) seems to generate a higher Type I error rate for BH and BY when μ_0 is low (i.e., $\mu_0 = 2.5\%$). This can be explained as follows. True strategies with a low signal-to-noise ratio (e.g., strategies that have a mean return below 2.5%) generate lower t-statistics in comparison with those in Table 5. This leads BH and BY to lower the cutoff t-statistic to maintain the pre-specified significance level, which in turn makes it easier for null strategies to overcome this threshold, leading to a higher Type I error rate. For these cases, our approach performs even better than in Table 5 in estimating the true error rates. It performs similarly to Table 5 for other specifications.

It is worth noting that while the performance of existing multiple testing methods varies across different specifications, our approach compares favorably with most methods consistently. For example, in Table 5 under $\mu_0 = 2.5\%$, Storey ($\theta = 0.8$) works well under a significance level of 10% as its true Type I error rate (i.e., 10.29%) is close to the desired level. As a result, the stated significance level of $\alpha = 10\%$ is a reasonably good estimate of the true error rate. However, the performance of Storey ($\theta = 0.8$) deteriorates substantially when μ_0 is raised to 5% or 10%, leading to large estimation errors when using the stated significance level to approximate the true error rates. In contrast, our approach performs well across all specifications of μ_0, making it the robust choice when assessing the performance of Storey ($\theta = 0.8$).

Finally, our simulation study focuses on the CAPIQ data, which is the actual data we use in our empirical work in this paper. For researchers who are interested in applying our approach, we also recommend the use of our simulation design to evaluate the performance of our method specifically to the data under consideration.

Table 5: **A Simulation Study on CAPIQ: Mean Return Distribution for True Strategies Is A Constant**

Simulated Type I error rates for CAPIQ when the mean return distribution for true strategies (F) is a constant. The simulation study runs as follows. We fix the fraction of true strategies at 10%. We first randomly identify 10% of strategies as true and assign mean returns to them according to F. Mean returns are set at zero for the remaining 90% of strategies. Let D_m denote the final data ($m = 1, 2, \ldots, M = 400$). Conditional on D_m, we bootstrap the time periods to generate the perturbed in-sample data $D_{m,k}$ ($k = 1, 2, \ldots, K = 100$). For each $D_{m,k}$ and for a given multiple testing method at a pre-specified significance level α, we calculate the true realized error rate (denoted as $FDR^a_{m,k}$). Implementing our approach (with a prior specification of p_0), we obtain the estimated error rate (denoted as $FDR^e_{m,k}$). We report the mean true Type I error rate ('Actual') defined as $\frac{1}{MK}\sum_{m=1}^{M}\sum_{k=1}^{K} FDR^a_{m,k}$ and mean error rate for our estimator ('Est.') defined as $\frac{1}{MK}\sum_{m=1}^{M}\sum_{k=1}^{K} FDR^e_{m,k}$. δ denotes the nominal error rate. Distribution F is a point mass at μ_0, where μ_0 takes the value of 2.5%, 5.0% or 10%. A bold number indicates a better performance of our model, i.e., 'Est.' is closer to 'Actual' compared to 'δ'.

		$\mu_0 = 2.5\%$				$\mu_0 = 5.0\%$				$\mu_0 = 10\%$			
Method	δ	Actual	Est.			Actual	Est.			Actual	Est.		
			p_0=5%	p_0=10%	p_0=15%		p_0=5%	p_0=10%	p_0=15%		p_0=5%	p_0=10%	p_0=15%
BH	1%	0.0075	0.0102	**0.0097**	**0.0091**	0.0104	**0.0111**	**0.0105**	**0.0100**	0.0108	**0.0116**	**0.0109**	**0.0095**
	5%	0.0375	**0.0425**	**0.0410**	**0.0390**	0.0472	**0.0465**	**0.0459**	**0.0445**	0.0461	**0.0480**	**0.0463**	**0.0438**
	10%	0.0718	**0.0794**	**0.0771**	**0.0737**	0.0876	**0.0872**	**0.0845**	**0.0804**	0.0875	**0.0897**	**0.0873**	**0.0827**
BY	1%	0.0016	**0.0025**	**0.0024**	**0.0022**	0.0020	**0.0023**	**0.0021**	**0.0022**	0.0019	**0.0023**	**0.0021**	**0.0020**
	5%	0.0059	**0.0089**	**0.0085**	**0.0080**	0.0080	**0.0081**	**0.0080**	**0.0075**	0.0083	**0.0089**	**0.0083**	**0.0079**
	10%	0.0108	**0.0157**	**0.0150**	**0.0142**	0.0155	**0.0155**	**0.0150**	**0.0142**	0.0154	**0.0163**	**0.0154**	**0.0146**
Storey	1%	0.0127	**0.0138**	**0.0140**	**0.0141**	0.0137	**0.0153**	**0.0150**	**0.0147**	0.0151	**0.0153**	**0.0141**	**0.0117**
($\theta = 0.4$)	5%	0.0525	**0.0548**	**0.0545**	**0.0544**	0.0575	**0.0598**	**0.0596**	**0.0584**	0.0615	**0.0614**	**0.0608**	**0.0599**
	10%	0.0950	**0.0988**	**0.0990**	**0.0984**	0.1049	**0.1080**	**0.1083**	**0.1064**	0.1129	**0.1125**	**0.1110**	**0.1095**
Storey	1%	0.0124	**0.0143**	**0.0141**	**0.0138**	0.0135	**0.0147**	**0.0146**	**0.0143**	0.0147	**0.0148**	**0.0146**	**0.0144**
($\theta = 0.6$)	5%	0.0524	**0.0542**	**0.0544**	**0.0547**	0.0574	**0.0592**	**0.0595**	**0.0588**	0.0616	**0.0598**	**0.0607**	**0.0601**
	10%	0.0981	0.1002	0.1004	0.1007	0.1073	**0.1085**	**0.1099**	**0.1091**	0.1147	**0.1100**	**0.1129**	**0.1121**
Storey	1%	0.0114	**0.0121**	**0.0121**	**0.0120**	0.0127	**0.0135**	**0.0136**	**0.0135**	0.0136	**0.0135**	**0.0136**	**0.0136**
($\theta = 0.8$)	5%	0.0532	**0.0537**	**0.0549**	**0.0552**	0.0576	**0.0581**	**0.0594**	**0.0599**	0.0618	**0.0588**	**0.0609**	**0.0615**
	10%	0.1029	**0.1022**	**0.1050**	**0.1058**	0.1137	**0.1113**	**0.1157**	**0.1174**	0.1209	**0.1132**	**0.1193**	**0.1212**

3.3 Performance Evaluation

3.3.1 Data Description: Mutual Funds

Our mutual fund data is obtained from the Center for Research in Security Prices Mutual Fund database. Since our goal is to take another look at Fama and French (2010), we apply similar screening procedures to Fama and French (2010) to obtain our dataset. In particular, our data starts from January 1984 to mitigate omission bias (Elton, Gruber, and Blake, 2001) and ends in December 2006 to be consistent with Fama and French (2010). We also limit tests to funds that reach five million 2006 dollars in AUM. Once a fund passes this size threshold, its subsequent returns are included in our tests. We also require that a fund has at least eight monthly return observations to be included in our tests.[42] Finally, following Fama and French (2010), we restrict our sample to funds that appear on CRSP at least five years before the end of the sample period (i.e., 2006) to avoid funds that have a short return history. We use the four-factor model in Fama and French (1993) and Carhart (1997) as our benchmark factor model applied to net mutual fund returns.[43]

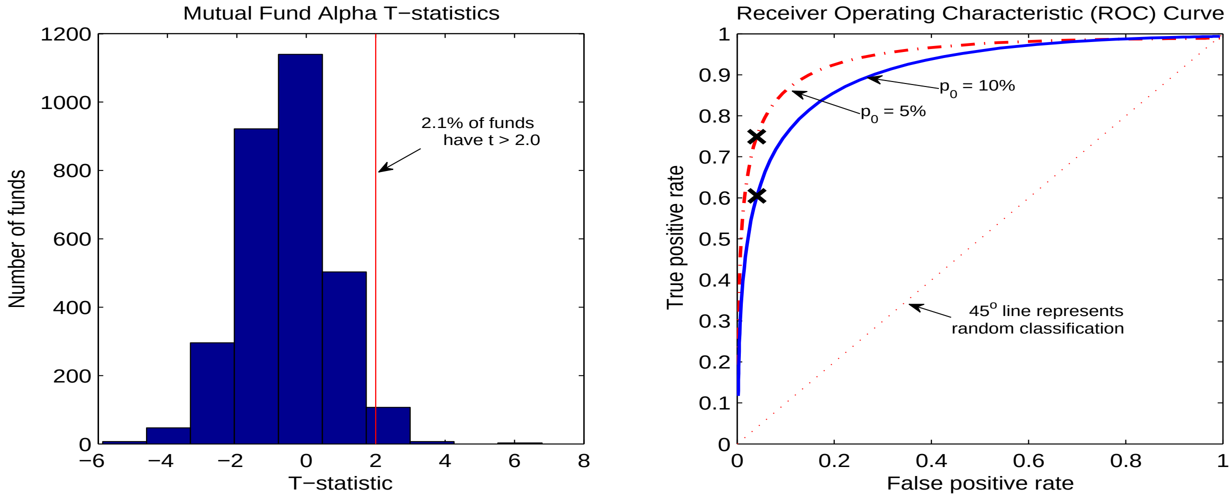
Our mutual fund sample closely follows the one used by Fama and French (2010). For example, there are 3,156 funds in Fama and French's sample that have an initial AUM exceeding 5 million 2006 dollars. In our sample, we have 3,030 funds. Summary statistics on fund returns are reported in Internet Appendix B. For our main results, we focus on the full sample of funds, i.e., all funds that have an initial AUM exceeding 5 million 2006 dollars. We examine alternative samples (i.e., AUM = \$250 million and \$1 billion) in Internet Appendix B.

Figure 4 shows the t-ratio distribution as well as the ROC curve for mutual fund alphas. Compared to the t-ratio distribution for the CAPIQ data in Figure 1, the fraction of funds with large and positive t-ratios is much smaller. This difference is reflected in the ROC curve, where under the same p_0 (i.e., $p_0 = 10\%$), the ROC curve for the CAPIQ data is pushed more towards the northwest direction than that for mutual funds, indicating a higher true positive rate for the same level of false positive rate. The t-ratio distribution for mutual funds is also skewed to the left, which consistent with previous findings that there is more evidence for extreme underperformers than for extreme outperformers. For the purpose of our application, we focus on the right tail of the distribution.

[42] Notice that this requirement applies to funds both in the actual sample and in the bootstrapped samples. We examine alternative cutoffs in the next section.

[43] Both the Fama-French factors and the momentum factor are obtained from Ken French's on-line data library.

Figure 4: **Mutual Fund Performance**



T-statistic distribution and the Receiver Operating Characteristic (ROC) curve for mutual fund alphas. For each fund in our data (1984–2006), we calculate the t-statistic of its alpha corresponding to the Fama-French-Carhart four-factor model. The top figure plots the t-statistic distribution. The bottom figure plots the ROC curves, corresponding to $p_0 = 5\%$ and $p_0 = 10\%$. The crosses on the ROC curves mark the points that correspond to a t-statistic of 2.0. The 'x' notations on the ROC mark the points.

3.3.2 Luck vs. Skill for Mutual Fund Managers

In contrast to our previous applications, we study a different question for the mutual fund data. In particular, we re-examine the question of whether or not there exist any outperforming funds. We focus on the joint test approach used in Fama and French, which treats the mutual fund population as a whole and tests whether the entire mutual fund population has a zero alpha (which is the null hypothesis) or not (i.e., if at least one fund has a positive alpha). Note that the goal of the joint test is different from the goal of multiple testing (as we studied previously), which is to identify the fraction of outperforming funds. We deliberately choose an application that is different from multiple tests to highlight the generality of our framework. It is also our purpose to use our framework to rigorously evaluate the Fama and French approach, which is an important and popular method in the literature.[44] Fama and French (2010)'s joint test suggests very few (if any) funds exhibit skill on a net return basis.

[44] For recent papers that apply the Fama and French approach or its predecessor, Kosowski, Timmermann, Wermers, and White (2006), see Chen and Liang (2007), Jiang, Yao, and Yu (2007), Ayadi and Kryzanowski (2011), D'Agostino, Mcquinn, and Whelan (2012), Cao, Chen, Liang, and Lo (2013), Hau and Lai (2013), Blake, Rossi, Timmermann, Tonks, and Wermers (2013), Busse, Goyal, and Wahal (2014), Harvey and Liu (2017), and Yan and Zheng (2017).

More specifically, we use our double-bootstrap approach to evaluate the Type I and Type II error rates for the single bootstrap approach proposed by Fama and French (2010). As we detailed in Section 2.3, the Type I and Type II error rate in this context refer to the probability of rejecting the null hypothesis of zero performance across all funds when this null is true (Type I error rate) and the probability of not rejecting this null when some funds have the ability to generate a positive alpha (Type II error rate). By varying p_0 (the fraction of funds that generate a positive alpha), we calculate these two error rates for the Fama and French approach.

While p_0 captures the prior belief about the fraction of skilled managers, it does not tell us the magnitude of the average alpha across these managers. We therefore calculate two additional statistics that help gauge the economic significance of the alphas generated by outperforming funds. In particular, "Avg. alpha" and "Avg. t-stat of alpha" calculate the averaged (across simulations) alpha and the t-statistic of alpha for the fraction of p_0 of funds that are assumed to be outperforming.[45,46]

Table 6 presents the simulated Type I and Type II error rates for Fama and French (2010), as well as the average alpha and the t-statistic of alpha for outperforming funds under different values of p_0. We examine four of the extreme percentile statistics used in Fama and French (i.e., 90^{th}, 95^{th}, 98^{th}, and 99^{th} percentiles), as well as two additional extreme percentiles (i.e., 99.9^{th} and 99.5^{th}) and the max statistic. We will later highlight the difference between the test statistics used in Fama and French and the additional test statistics we consider. The use of extreme percentiles in this context is driven by the general idea that extreme percentiles are more sensitive to the existence of outperforming funds, and therefore are potentially more powerful in detecting outperformance. For example, if 1% of funds are outperforming, we should expect to see a statistically significant 99^{th} percentile (as well as the 99.9^{th} and 99.5^{th} percentile) when tested against its distribution under the null hypothesis. In contrast, the other percentiles should be insignificant or at least less significant. In Table 6, a given column (which corresponds to a certain test statistic) calculates the Type I error rates (Panel A) and the Type II error rates (Panel B) associated with different significance levels. If the

[45] We calculate "Avg. alpha" and "Avg. t-stat of alpha" in the following way. Following the four-step procedure of our method presented in Section 2.2, for each bootstrapped sample as generated by Step I, we find the top p_0 of funds and calculate the median alpha and the t-statistic of alpha. We then take the average across J bootstrapped samples (Step III) to calculate "Avg. alpha" and "Avg. t-stat of alpha".

[46] Note that due to the undersampling issue of the Fama and French approach that we identify later, the reported "Avg. alpha" and "Avg. t-stat of alpha" are somewhat higher than the average alpha and average t-statistic of alpha for the actual funds in the bootstrapped sample. However, had smaller alphas been injected into funds in our first-stage bootstrap, the test power for the Fama and French approach would be even lower. Therefore, one can interpret our results as providing the upper bounds on test power for a given average level of alpha under the alternatives.

test statistic performs well, the Type I error rates should be lower than the pre-specified significance levels and the Type II error rates should be close to zero.

Table 6: **Simulated Error Rates for Fama and French (2010)**

Simulated Type I and Type II error rates for the Fama and French (2010) approach. For each p_0 of the fraction of strategies that are believed to be true, we follow our method in Section 2.3 and perturb the adjusted data $M = 1,000$ times to calculate the empirical Type I and Type II error rates. We consider six percentiles as well as the maximum. "Avg." and "Avg. t-stat" report the average (across simulations) of the mean alpha and alpha t-statistic for the fraction p_0 of funds that are assumed to be outperforming. All funds with an initial AUM exceeding \$5 million are included, resulting in 3,030 funds in total in the 1984–2006 period. Panel A reports the Type I error rate, that is, the probability of falsely declaring that some managers have skill when no fund is outperforming. Panel B reports the Type II error rate, that is, the probability of falsely declaring that no manager is skilled when some funds are outperforming.

				Test Statistics (Percentiles)						
p_0 (frac. of true)	Avg. alpha	Avg. t-stat of alpha	α (sig. level)	Max	99.9%	99.5%	99%	98%	95%	90%
Panel A: Type I Error Rate										
0	0	0	1%	0.001	0.000	0.003	0.000	0.002	0.009	0.014
			5%	0.017	0.003	0.004	0.005	0.016	0.029	0.040
			10%	0.033	0.009	0.004	0.007	0.043	0.071	0.079
Panel B: Type II Error Rate										
0.5%	14.60	3.97	1%	0.998	1.000	0.997	0.999	0.997	0.988	0.983
			5%	0.977	0.997	0.997	0.994	0.978	0.965	0.960
			10%	0.953	0.989	0.993	0.980	0.946	0.920	0.916
1%	13.22	3.68	1%	0.997	1.000	0.999	1.000	0.996	0.985	0.981
			5%	0.970	0.996	0.997	0.991	0.971	0.959	0.957
			10%	0.942	0.986	0.990	0.966	0.932	0.912	0.910
2%	10.66	3.30	1%	1.000	0.999	0.998	0.982	0.984	0.983	0.981
			5%	0.976	0.994	0.993	0.869	0.937	0.938	0.948
			10%	0.957	0.986	0.983	0.741	0.852	0.880	0.898
3%	9.75	3.02	1%	0.997	0.997	0.999	0.867	0.914	0.976	0.975
			5%	0.966	0.991	0.992	0.478	0.677	0.896	0.925
			10%	0.940	0.981	0.965	0.277	0.506	0.795	0.849
5%	8.13	2.63	1%	0.998	1.000	0.998	0.802	0.710	0.961	0.969
			5%	0.974	0.995	0.992	0.363	0.336	0.815	0.898
			10%	0.947	0.983	0.957	0.145	0.186	0.691	0.790
10%	6.28	2.10	1%	0.994	0.999	0.999	0.739	0.425	0.860	0.950
			5%	0.974	0.996	0.991	0.223	0.134	0.558	0.804
			10%	0.952	0.984	0.950	0.087	0.050	0.351	0.663
15%	5.19	1.80	1%	0.994	1.000	0.998	0.672	0.283	0.527	0.787
			5%	0.974	0.997	0.992	0.200	0.052	0.206	0.498
			10%	0.947	0.980	0.945	0.051	0.011	0.090	0.332

When $p_0 = 0$ (Panel A of Table 6), we see that most metrics considered meet the pre-specified significance levels (with the exception of the 90^{th} percentile at the 1% level). This means that when the null hypothesis is true (i.e., no fund is outperforming), the chance for us to falsely claim the existence of outperforming funds with a given test statistic is below the pre-specified significance level. This result confirms that the Fama and French approach performs well in terms of the Type I error rate for the mutual fund data. Notice that this is not a trivial finding in that bootstrap-based methods are not guaranteed to meet the pre-specified significance level (Horowitz, 2001).

When $p_0 > 0$ (Panel B of Table 6), we allow some funds to have the ability to generate positive alphas, so a powerful test statistic should be able to detect these outperforming funds and reject the null hypothesis with a high probability. However, this is not the case for the Fama and French approach. The Type II error rates (i.e., the probability of failing to reject the null hypothesis) are very high. For example, when $p_0 = 2\%$ of funds are truly outperforming, the chance for the best performing metric (i.e., the 99^{th}-percentile in this case since it has the lowest Type II error rate among all seven test statistics) to commit the Type II error is 86.9% under 5% significance level. In fact, even when $p_0 = 5\%$ of funds are outperforming, the lowest Type II error rate across different test statistics is still 33.6%.

Our finding of a low test power for the Fama and French method is economically significant. For example, when $p_0 = 2\%$, the average alpha and the average t-statistic of alpha for outperforming funds are 10.66% (per annum) and 3.30, respectively. In other words, even when 2% of funds are truly outperforming and are endowed with on average an annualized alpha of 10.66%, there is still a 86.9% chance (at 5% significance level) for the Fama and French approach to falsely declare a zero alpha for all funds.

Note that while our framework proposes a data-driven approach to inject alphas into funds that are assumed to be truly outperforming, an alternative approach is to inject hypothetical alphas directly. For example, Fama and French (2010) assume that true alphas follow a normal distribution and randomly assign alphas from this distribution to funds. We believe that our approach is potentially advantageous in that we do not need to take a stand on the cross-sectional alpha distribution among outperforming funds. For instance, we do not observe extremely large t-statistics for alphas (e.g., a t-statistic of 10) as large alphas are usually matched with funds with a high level of risk (which is consistent with the intuition that there is a risk and return trade-off). However, the random assignment of normally distributed alphas in Fama and French (2010) may assign large alphas to funds with a small risk, implying abnormally high t-statistics for certain funds that may never be observed in the data. In contrast, our approach injects the alpha for each outperforming fund based on its in-sample alpha estimate. We therefore preserve the data implied cross-sectional alpha distribution for outperforming funds.

Our results also highlight a non-monotonic pattern in Type II error rate as a function of the percentile for the test statistic. When p_0 is relatively large (e.g., $\geq 2\%$), the 98^{th}-percentile or the 99^{th}-percentile seem to perform the best in terms of Type II error rate. On the other hand, test statistics with either the highest percentile (e.g., the maximum and the 99.9^{th}-percentile) or the lowest percentile (i.e., 90^{th}-percentile) perform significantly worse.[47] Our explanation for this result is that while test statistics with a higher percentile are in general more sensitive to the existence of a small fraction of outperforming funds and therefore should be more powerful, test statistics with the highest percentiles may be sensitive to outliers in t-statistics generated by both skilled and unskilled funds.[48]

How important is our result in the context of performance evaluation for mutual funds? The question of whether or not there exist outperforming mutual fund managers is economically important. However, despite its importance, there has been considerable debate about the answer to the question. KTWW independently bootstrap each fund's return and find evidence for outperformance for a significant fraction of funds. Fama and French (2010) modify KTWW by taking cross-sectional dependence into account and find no evidence of outperformance. Aside from the different samples in the two papers, we offer a new perspective that helps reconcile these conflicting findings.[49]

The Fama and French approach lacks power to detect outperforming funds. Even when 2%-5% of funds are truly outperforming and, moreover, these funds are endowed with the large (and positive) returns that the top 2%-5% of funds have in the actual sample as in our

[47] The question of which percentile statistics to use deserves some discussion. In our context, we show all test statistics have low power so this question does not affect the interpretation of our results. In general, more extreme statistics are likely more powerful. For example, if the true p_0 equals 5%, then percentiles higher than 95% should be more likely to be rejected than the other percentiles. On the other hand, more extreme statistics are perhaps less interesting from an economic perspective. For instance, if the max statistic rejects but not the 99.9-th percentile, this seems to suggest that at most 0.1% of funds are outperforming, which is not too different from the overall null hypothesis that no fund is outperforming. Taken all together, one should balance statistical power with economic significance when choosing test statistics.

[48] For example, suppose we use the maximum t-statistic as the test statistic. In the bootstrapped simulations in Fama and French (2010), a fund with a small sample of returns may have an even smaller sample size after bootstrapping, potentially resulting in extreme t-statistics for the bootstrapped samples. While this happens for both the (assumed) skilled and unskilled funds, given the larger fraction of unskilled funds, they are more likely to drive the max statistic, making it powerless in detecting outperforming funds. We examine this issue in depth in the next section by dissecting the Fama and French approach.

[49] Note that a fund is required to have at least 30 observations in KTWW's analysis. As such, the undersampling issue we identify with the Fama-French approach as explained in the next section affects KTWW to a much lesser extent.

simulation design, the Fama and French approach still falsely declares a zero alpha across all funds with a high probability.

Note that the Fama and French approach, by taking cross-sectional information into account, achieves the pre-specified significance level when the null hypothesis is that all funds generate a zero alpha. As such, theoretically, it should make fewer mistakes (i.e., Type I errors) than KTWW when the Fama and French null hypothesis is true. However, the price we have to pay for making fewer Type I errors is a higher Type II error rate. Our results show the Type II error rate for the Fama and French approach is so high that it may mask the existence of a substantial fraction of true outperformers.

In Internet Appendix B, we follow Fama and French (2010) and perform our analysis to alternative groups of funds classified by fund size. In particular, we examine the groups of funds with an initial AUM exceeding \$250 million and \$1 billion, respectively. Table IB.2 and IB.3 report the results. Comparing Table 6, IB.2, and IB.3, there is some evidence for decreasing return to scale in that when p_0 is relatively large (i.e., $p_0 \geq 3\%$), the same p_0 implies a higher mean alpha for a smaller size cutoff, consistent with Chen, Hong, Huang, and Kubik (2004) and Berk and Green (2004). However, in terms of test power, the results in IB.2 and IB.3 are even worse than our main results in Table 6. For example, when $p_0 = 5\%$, the Type II error rates (under 5% significance level) are at least 50.9% (size cutoff = \$250 million) and 81.4% (size cutoff = \$1 billion) across the test statistics we examine. Our finding of a low test power for the Fama-French approach is robust to the particular size group that we study.

Overall, given the emphasis on "stars" for the investment industry,[50] we show that it is, in general, difficult to identify stars. While a stringent statistical threshold has to be used to control for luck by a large number of unskilled managers, it also makes it hard for good managers, even the ones with a stellar performance, to stand out from the crowd.

3.3.3 Dissecting the Fama and French Approach

We probe into the Fama and French (2010) approach to analyze its test power. One key aspect of the Fama and French bootstrapping approach is that if a fund does not exist for the entire sample period, the number of observations for the bootstrapped sample may differ from the number of observations for the actual sample. We find that undersampled funds (i.e., funds with fewer observations in the bootstrapped sample than in the actual sample)

[50] See, e.g., Nanda, Wang, and Zheng (2004), Guercio and Tkac (2008), and Sastry (2013).

with a small number of time-series observations in the bootstrapped sample tend to generate t-statistics that are more extreme than what the actual sample implies. This distorts the cross-sectional t-statistic distribution for the bootstrapped samples, making it difficult for the Fama and French approach to reject the null. We provide details of our analysis in Appendix B.

3.3.4 Are There Outperforming Funds?

Guided by our previous analysis on test power, we now explore ways to improve the performance of the Fama and French approach and revisit the question in Fama and French (2010): do there exist outperforming funds?

Since the undersampling of funds with a relatively short sample period contributes to the low test power of the Fama and French approach, the different methods we explore in this section all boil down to the question of what funds do we drop to alleviate the issue of undersampling. In particular, we explore two ways to drop funds with a small sample size. One way is to focus on the full sample (similar to Fama and French, 2010) but drop funds with a sample size below a certain threshold T, with $T \geq 36$ as guided by our analysis in Figure B.1 and B.2. The other is to focus on different sub-samples and require funds to have complete data over these sub-samples.

Notice that while dropping funds with a short sample improves the performance of the Fama and French approach, it raises an obvious concern about sample selection. To the extent that the average performance among funds with a certain sample size is correlated with sample size, our results in this section may provide a biased evaluation of the existence of outperforming funds. For example, while survivorship may bias our results towards finding significant outperforming funds (Brown et al., 1992, Carhart et al., 2002, and Elton, Gruber, and Blake, 1996), reverse-survivorship bias (Linnainmaa, 2013) or decreasing returns to scale (Chen, Hong, Huang, and Kubik, 2004, Berk and Green, 2004, Harvey and Liu, 2018b) may bias our results in the opposite direction. Because of this selection concern, we limit the interpretation of our results to the particular sample of mutual funds we analyze, acknowledging that we are not trying to answer exactly the same question as in Fama and French (i.e., does any fund in their mutual fund sample outperform?) since we are examining different samples of funds.

For the full sample analysis similar to Fama and French (2010), we explore two specifications for the threshold T: $T = 36$ and $T = 60$. For the sub-sample analysis, we focus on five-year sub-samples and require complete data for each sub-sample.[51]

We first apply our framework to evaluate test power for the new specifications. Table IB.6 and IB.7 in Internet Appendix report the results for the full sample analysis. Table IB.8 reports the average test power across all sub-samples whereas more detailed results on test power for each sub-sample are presented in Table IB.9–IB.14.

When the threshold T is raised to 36 (Table IB.6) and 60 (Table IB.7), the performance of the Fama and French approach in terms of test power is, in general, improved compared to Table 6. Interestingly, when p_0 is relatively small (e.g., $p_0 \leq 2\%$), there is a stark contrast in performance of the more extreme percentiles (i.e., the maximum, the 99.9^{th}-percentile and the 99.5^{th}-percentile). When the threshold number of observations is eight (as in Table 6), the more extreme percentiles perform worse than other test statistics, although all test statistics lack power in detecting outperforming funds. In contrast, when the threshold T is raised to 36 and 60 (as in Table IB.6 and IB.7), the more extreme percentiles perform substantially better than other test statistics and, more importantly, seem to be powerful at detecting outperformers. For example, in Table IB.6 when $p_0 = 0.5\%$, the Type II error rates for the maximum and the 99.9^{th}-percentile are 11.1% and 4.6% (at the 10% significance level), which are much lower than those under other test statistics (e.g., the Type II error rates for the 99^{th}- and 98^{th}-percentile are 76.6% and 82.9%). Overall, our adjustment to the Fama and French approach not only improves test power for the test statistics that are used in Fama and French, but also allows us to uncover the extreme statistics (previously disguised as powerless in the Fama and French framework) that are much more powerful in detecting outperforming funds.

Turning to our results for sub-samples in Table IB.8, there is also a substantial improvement in test power compared to Table 6, although the improvement is not as large as that shown by Table IB.7 (which also requires $T = 60$). Note that we have a much shorter time period over sub-samples than over the full sample, which usually implies a decrease in test power. Despite this decrease in the level of power, we are able to achieve a higher power compared to Table 6 by dropping funds with a short return history.

[51] In particular, we split our data into five five-year sub-samples (i.e., 1984–88, 1989–93, 1994–98, 1999–03, and 2004–08) and an eight-year sub-sample (i.e., 2009–16) for the last part of our data.

We now explore the use of the modified Fama and French approach to test for the existence of outperforming funds. Table 7 presents the full sample results and Table 8 the sub-sample tests.

Focusing on the 1984-2006 sample in Table 7, the p-values for the original Fama-French approach across different test statistics are uniformly higher than those for the adjusted methods. This highlights the lack of power for the original Fama-French approach caused by its missing data bootstrap. Based on the adjusted Fama-French methods, we find evidence of outperforming funds for the max statistic (at 1% level) and the 99.9th-percentile (at 10% level), both for $T \geq 36$ and $T \geq 60$.

It is worth emphasizing that the statistical significance of the α^{th} percentile (as the test statistic) may not be solely related to the performance of funds in the top (100-α) percent. As a result, it is incorrect to use the significance of a given test statistic to infer the fraction of outperforming funds. Our previous application to the $18,000+$ anomalies illustrated a similar point. For example, the significance of the max statistic does not imply that there is only one fund that is outperforming. Based on Table IB.7 (i.e., the cutoff T is set at 60), the fact that both the maximum and the 99.9^{th}-percentile are significant at the 10% level (as in Table 7) while other test statistics are not may indicate that $p_0 = 0.5\%$. This is because, at $p_0 = 0.5\%$, our simulation results show that the probability for the maximum and the 99.9^{th}-percentile to correctly reject the null hypothesis are 95.6% ($= 1 - 0.044$) and 98.2% ($= 1 - 0.018$), respectively, which are much higher than the probabilities implied by other test statistics (with the highest one being $50.8\% = 1 - 0.492$ for the 99.5^{th}-percentile).

While our full-sample results essentially assume the constancy of alpha for funds, there may be low-frequency fluctuations in fund alphas that are caused by time-varying market conditions or decreasing returns to scale.[52] We therefore turn to our results on subsamples in Table 8. Using the adjusted Fama-French approach (requiring longer fund histories) and consistent with existing papers that document the time variation in the average performance across funds, we find strong evidence for outperforming funds in the early part of the sample (in particular, the 1984–1993 sample). But there is still some evidence for outperforming funds for the later part of the sample (i.e., the 1999–2016 sample). In contrast, with the original Fama-French approach, there is essentially no evidence for outperforming funds across all subsamples.

[52] See, e.g., Chen, Hong, Huang, and Kubik (2004), Berk and Green (2004), Avramov, Kosowski, Naik, and Teo (2011), and Harvey and Liu (2018b).

Our analysis so far highlights the difference between the Fama and French approach and other multiple testing methods studied in the previous section, which try to identify the fraction of outperforming funds. While the fraction of outperforming funds (as modeled by p_0 in our framework) influences test power of the Fama and French approach, in general one cannot infer this fraction from the Fama and French test statistics. Romano and Wolf (2005) have a similar discussion of this difference. For other methods (including alternative multiple testing approaches) that can be used to estimate the fraction of outperforming funds, see Barras, Scaillet, and Wermers (2010), Ferson and Chen (2017), Harvey and Liu (2018b), and Andrikogiannopoulou and Papakonstantinous (2016). Although these methods can potentially provide a more detailed description of the cross-sectional distribution of alphas, they are not necessarily as powerful as the Fama-French approach (especially the adjusted Fama-French approach) in detecting the existence of outperforming funds when controlling for cross-sectional dependence in tests. Our focus is on the performance of the Fama and French approach itself. We leave the examination of test power for these alternative methods using our framework to future research.

Overall, our results in this section provide an example of how to use our framework to evaluate the Fama and French approach, which is a joint test that is different from the multiple testing methods we studied in the previous section. We show this method has low test power when applied to the mutual fund data. A further analysis identifies the undersampling of funds with a small number of observations producing unrealistically high t-statistics and greatly decreasing the power to identify outperformers. We modify the Fama and French approach accordingly by dropping funds with a short return history over either the full sample or sub-samples. We revisit the problem in Fama and French (2010) and find some evidence for the existence of outperforming funds. That said, consistent with the long literature in mutual fund evaluation, there is only modest evidence of fund outperformance.

Table 7: **Tests for the Existence of Outperforming Funds: Full Sample**

Tests for the existence of outperforming funds over the full sample. For a given sample period, we drop funds with the number of observations below a certain threshold level and use the Fama and French (2010) approach to test the joint hypothesis of a zero alpha across all funds. For test statistics, we consider six percentiles as well as the maximum. All funds with an initial AUM exceeding $5 million are included. We consider both the Fama and French sample period (i.e., 1984–2006) and the full sample period (i.e., 1984–2016). We present the p-values in the parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

			Test Statistics (for various percentiles)						
Sample Period	# of Funds		Max	99.9%	99.5%	99%	98%	95%	90%
Panel A: Benchmark (Fama and French, 2010): $T \geq 8$									
1984-2006	3030								
(Fama and French)		t-stat	6.816	3.718	2.664	2.387	1.968	1.542	1.087
		p-value	(0.951)	(0.939)	(0.847)	(0.737)	(0.809)	(0.754)	(0.841)
Panel B: $T \geq 36$									
1984-2006	2,668								
(Fama and French)		t-stat	6.816***	3.759*	2.660	2.370	1.971	1.573	1.103
		p-value	(0.003)	(0.097)	(0.489)	(0.509)	(0.643)	(0.617)	(0.760)
1984-2016	3,868								
(Full sample)		t-stat	6.959***	3.487	2.815	2.508	2.010	1.466	1.038
		p-value	(0.004)	(0.289)	(0.309)	(0.337)	(0.629)	(0.809)	(0.884)
Panel C: $T \geq 60$									
1984-2006	2,387								
(Fama and French)		t-stat	6.816***	3.824*	2.699	2.411	2.006	1.621	1.134
		p-value	(0.001)	(0.088)	(0.407)	(0.428)	(0.577)	(0.537)	(0.702)
1984-2016	3,393								
(Full sample)		t-stat	6.959***	3.541	2.897	2.546	2.089	1.529	1.075
		p-value	(0.000)	(0.182)	(0.198)	(0.265)	(0.472)	(0.690)	(0.808)

Table 8: **Tests for the Existence of Outperforming Funds: Sub-samples**

Tests for the existence of outperforming funds over sub-samples. For a given sample period, we only keep funds with complete return data over this period and use the Fama and French (2010) approach to test the joint hypothesis of a zero alpha across all funds. For test statistics, we consider six percentiles as well as the maximum. All funds with an initial AUM exceeding $5 million are included. We present the p-values in the parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

				Test Statistics (for various percentiles)						
Sub-sample		# of Funds		Max	99.9%	99.5%	99%	98%	95%	90%
1984-88	$T \geq 8$	455								
			t-stat	4.061	4.061	3.709	3.389	3.030	2.376	1.916
			p-value	(0.764)	(0.764)	(0.537)	(0.365)	(0.209)	(0.129)	(0.111)
	$T = 60$	238								
			t-stat	4.061**	4.061**	4.001**	3.464**	2.881*	2.375*	1.896*
			p-value	(0.048)	(0.048)	(0.022)	(0.032)	(0.060)	(0.054)	(0.058)
1989-93	$T \geq 8$	1,109								
			t-stat	3.887	3.755	3.183	2.820	2.459	1.924	1.489
			p-value	(0.979)	(0.973)	(0.835)	(0.755)	(0.661)	(0.545)	(0.456)
	$T = 60$	352								
			t-stat	3.887*	3.887*	3.152*	3.011*	2.680*	2.179*	1.704
			p-value	(0.083)	(0.083)	(0.089)	(0.087)	(0.084)	(0.093)	(0.116)
1994-98	$T \geq 8$	1,857								
			t-stat	3.482	3.339	2.675	2.127	1.877	1.437	0.919
			p-value	(0.992)	(0.976)	(0.906)	(0.937)	(0.905)	(0.893)	(0.972)
	$T = 60$	848								
			t-stat	3.237	3.057	2.104	1.913	1.596	1.182	0.692
			p-value	(0.497)	(0.549)	(0.859)	(0.844)	(0.899)	(0.940)	(0.995)
1999-03	$T \geq 8$	2,822								
			t-stat	5.791	3.797	3.043	2.726	2.256	1.729	1.365
			p-value	(0.924)	(0.899)	(0.633)	(0.480)	(0.494)	(0.488)	(0.434)
	$T = 60$	1,511								
			t-stat	3.533	3.508	3.048*	2.756	2.472	1.874	1.508
			p-value	(0.346)	(0.198)	(0.087)	(0.146)	(0.145)	(0.238)	(0.242)
2004-08	$T \geq 8$	3,084								
			t-stat	4.561	3.667	2.969	2.591	2.218	1.727	1.237
			p-value	(0.981)	(0.954)	(0.815)	(0.733)	(0.660)	(0.580)	(0.641)
	$T = 60$	1,722								
			t-stat	4.295*	3.594	2.964	2.685	2.262	1.770	1.303
			p-value	(0.095)	(0.228)	(0.231)	(0.235)	(0.316)	(0.371)	(0.464)
2009-16	$T \geq 8$	2,608								
			t-stat	4.192	3.348	2.338	1.981	1.782	1.271	0.807
			p-value	(0.928)	(0.875)	(0.915)	(0.934)	(0.876)	(0.930)	(0.978)
	$T \geq 60$	1,642								
			t-stat	4.192*	3.341	2.363	1.987	1.813	1.385	0.883
			p-value	(0.074)	(0.240)	(0.614)	(0.764)	(0.685)	(0.741)	(0.903)

4 Conclusion

Two types of mistakes (i.e., false discoveries and missed discoveries) vex empirical research in financial economics. Both are exacerbated in the context of multiple tests. We provide a data-driven approach that estimates the frequencies of both errors for multiple tests. It also allows us to flexibly estimate functions of the two frequencies such as a weighted average of the false discovery frequency and the missed discovery frequency, with the weight determined by the corresponding cost of the type of mistake.

While current research on multiple testing focuses on controlling the Type I error rate, we show that it is also important to consider the Type II error rate. For the selection of investment strategies, a weighted average of the Type I error rate and the Type II error rate is likely more consistent with the objective function of the investor. For the selection of mutual fund managers, current methods, which ignore the Type II error rate, may lead us to miss outperforming managers. Instead of relying on existing multiple-testing adjustments, our approach allows for the provision of user-specific significance thresholds for a particular set of data. Alternatively, if the research wants to use a traditional multiple testing adjustment, our method is able to determine which one is best suited for the particular application.

With the advent of big data and the advances in computing technologies, it becomes increasingly important to correct the biases associated with multiple tests and data mining. We take advantage of current computing technologies and develop a simulation-based framework to make such corrections. We expect our method to be useful for future research in financial economics and other fields.

References

Andrikogiannopoulou, A., and F. Papakonstantinous. 2016. Estimating mutual fund skill: A new approch. *Working Paper.*

Andrikogiannopoulou, A., and F. Papakonstantinous. 2018. Reassessing false discoveries in mutual fund performance: Skill, luck, or lack of power? *Journal of Finance, Forthcoming.*

Avramov, D., R. Kosowski, N. Y. Naik, and M. Teo. 2011. Hedge funds, managerial skill, and macroeconomic variables. *Journal of Financial Economics 99, 672-692.*

Ayadi, M. A., and L. Kryzanowski. 2011. Fixed-income fund performance: Role of luck and ability in tail membership. *Journal of Empirical Finance 18, 379–392.*

Bajgrowicz, P., and O. Scaillet. 2012. Technical trading revisited: False discoveries, persistence tests, and transaction costs. *Journal of Financial Economics 106, 473–491.*

Barras, L. 2018. A large-scale approach for evaluating asset pricing models. *Journal of Financial Economics, Forthcoming.*

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.*

Barras, L., O. Scaillet, and R. Wermers. 2018. Reassessing false discoveries in mutual fund performance: Skill, luck, or lack of power? *Working Paper.*

Beneish, M. D. 1997. Detecting GAAP violation: Implications for assessing earnings management among firms with extreme financial performance. *Journal of Accounting and Public Policy, Fall 1997, 271–309.*

Beneish, M. D. 1999. The detection of earnings manipulation. *Financial Analysts' Journal 1999, 24–36.*

Benjamini, Y., and Y. Hochberg. 1995. Controlling for the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B 57, 289–300.*

Benjamini, Y., and D. Yekutieli. 2001. The control of the false discovery rate in multiple testing under dependency. *Annuals of Statistics 29, 1165–1188.*

Berk, J. B., and R. C. Green. 2004. Mutual fund flows and performance in rational markets. *Journal of Political Economy 112, 1269–1295.*

Blake, D., A. G. Rossi, A. Timmermann, I. Tonks, and R. Wermers. 2013. Decentralized Investment Management: Evidence from the Pension Fund Industry. *Journal of Finance 68, 1133–1178.*

Busse, J. A., A. Goyal, S. Wahal. 2014. Investing in a global world. *Review of Finance 18, 561–590.*

Carhart, M. M. 1997. On persistence in mutual fund performance. *Journal of Finance 52, 57–82.*

Cao, C., Y. Chen, B. Liang, and A. W. Lo. 2013. Can hedge funds time market liquidity? *Journal of Financial Economics 109, 493–516.*

Chen, J., H. Hong, M. Huang, and J. D. Kubik. 2004. Does fund size erode mutual fund performance? The role of liquidity and organization. *American Economic Review 94, 1276–1302.*

Chen, Y., and B. Liang. 2007. Do market timing hedge funds time the market? *Journal of Financial and Quantitative Analysis 42, 827–856.*

Chordia, T., A. Goyal, and A. Saretto. 2018. P-hacking: Evidence from two million trading strategies. *Working Paper.*

Christiansen, C., N. S. Grønborg, and O. L. Nielsen. 2019. Mutual fund selection for realistically short samples. *Working Paper.*

D'Agostino, A., K. McQuinn, and K. Whelan. 2012. Are some forecasters really better than others? *Journal of Money, Credit, and Banking 44, 715–732.*

De Long, J. B., and K. Lang. 1992. Are all economic hypotheses false? *Journal of Political Economy 100, 1257–1272.*

DeGroot, M. 1975. Probability and Statistics. Addison-Wesley, Reading, MA.

DeGroot, M., and M. J. Schervish. 2011. Probability and Statistics, 4th edition. Pearson Education Limited.

Elton, E. J., M. J. Gruber, and C. R. Blake. 2001. A first look at the accuracy of the CRSP mutual fund database and a comparison of the CRSP and Morningstar mutual fund database. *Journal of Finance 56, 2415–2430.*

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., and Y. Chen. 2017. How many good and bad fund managers are there, really? *Working Paper.*

Genovese, C., and L. Wasserman. 2002. Operating charateristics and extensions of the false discovery rate procedure. *J. R. Stat. Soc. Ser. B Stat. Methodol. 64, 499–517.*

Giglio, S., Y. Liao, and D. Xiu. 2018. Thousands of alpha tests. *Working Paper.*

Guercio, D. D., and P. A. Tkac. 2008. Star power: The effect of Morningstar ratings on mutual fund flow. *Journal of Financial and Quantitative Analysis 43, 907–936.*

Hau, H., and S. Lai. Real effects of stock underpricing. *Journal of Financial Economics 108, 392–408.*

Harvey, C. R. 2017. Presidential address: The scientific outlook in financial economics. *Journal of Finance 72, 1399–1440.*

Harvey, C. R., and Y. Liu. 2013. Multiple testing in economics. *Working Paper.* Available at https://ssrn.com/abstract=2358214.

Harvey, C. R., Y. Liu, and H. Zhu. 2016. ... and the cross-section of expected returns. *Review of Financial Studies 29, 5–72.*

Harvey, C. R., and Y. Liu. 2017. Luck vs. skill and factor selection. in *The Fama Portfolio, 250–260*, John Cochrane and Tobias J. Moskowitz, ed., Chicago: University of Chicago Press.

Harvey, C. R., and Y. Liu. 2018a. Lucky Factors. *Working Paper.* Available at https://ssrn.com/abstract=2528780.

Harvey, C. R., and Y. Liu. 2018b. Detecting Repeatable Performance. *Review of Financial Studies 31, 2499–2552.*

Harvey, C. R., Y. Liu, N. Polson, and J. Xu. 2019. Revisiting Semi-Strong Market Efficiency. *Working Notes.*

Hou, K., C. Xue, and L. Zhang. 2018. Replicating anomalies. *Forthcoming, Review of Financial Studies.*

Horowitz, J. L. 2001. The bootstrap. *Handbook of Econometrics 5, Chapter 52, 3159–3228.*

Ioannidis, J. P. A. 2005. Why most published research findings are false? *PLoS Medicine 2 (8): e124.*

Ioannidis, J. P. A., and C. H. Doucouliagos. 2013. What's to know about the credibility of empirical economics. *Journal of Economic Surveys 27, 997–1004.*

Ioannidis, J. P. A., T. D. Stanley, H. Doucouliagos. 2017. The power of bias in economics research. *Economic Journal 127, 236–265.*

Jiang, G. J., T. Yao, and T. Yu. 2007. Do mutual funds time the market? Evidence from portfolio holdings. *Journal of Financial Economics 86, 724–758.*

Kandel, S., and R. F. Stambaugh. 1996. On the predictability of stock returns: An asset-allocation perspective. *Journal of Finance 51, 385–424.*

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.*

Leamer, E. E. 1983. Let's take the con out of econometrics. *American Economic Review 73, 31–43.*

Lehmann, E. L., and J. P. Romano. 2005. Generalizations of the familywise error rate. *Annals of Statistics 33, 1138–1154.*

Nanda, V., Z. J. Wang, and L. Zheng. 2004. Family values and the star phenomenon: Strategies of mutual fund families. *Review of Financial Studies 17, 667–698.*

Powers, D. M. W. 2011. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies 2: 37–63.*

Romano, J. P., A. M. Shaikh, and M. Wolf. 2008. Formalized data snooping based on generalized error rates. *Econometric Theory 24, 404–447.*

Romano, J. P., and A. M. Shaikh. 2006. Stepup procedures for control of generalizations of the familywise error rate. *Annals of Statistics 34, 1850–1873.*

Romano, J. P., and M. Wolf. 2005. Stepwise multiple testing as formalized data snooping. *Econometrica 73, 1237–1282.*

Romano, J. P., and M. Wolf. 2007. Control of generalized error rates in multiple testing. *Annals of Statistics 35, 1378–1408.*

Sarkar, S. K. 2006. False discovery and false nondiscovery rates in single-step multiple testing procedures. *Annuals of Statistics 34, 394–415.*

Sastry, R. 2013. The cross-section of investing skill. *Working Paper.*

Scott, J. G., and J. O. Berger. 2006. An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference 136, 2144–2162.*

Storey, J. D., 2002. A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B 64, 479–498.*

Storey, J. D. 2003. The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics 31, 2013–2035.*

Van Rijsbergen, C. J. 1979. Information retrieval (2nd ed). London: Butterworths.

Yan, X., and L. Zheng. 2017. Fundamental analysis and the cross-section of stocks returns: A data-mining approach. *Review of Financial Studies 30, 1382–1423.*

Ziliak, S. T., and D. N. McCloskey. 2004. Size matters: The standard error of regressions in the *American Economic Review. Journal of Socio-Economics 33, 527–546.*

A Implementing Romano, Shaikh, and Wolf (2008)

Romano, Shaikh, and Wolf (2008) suggest a bootstrapping method to take the dependence structure in the data into account to derive the statistical cutoff. Similar to the implementation in Romano, Shaikh, and Wolf (2008), who set the number of bootstrapped iterations at $B = 500$, we set $B = 1,000$ for the bootstrap procedure.

Romano, Shaikh, and Wolf (2008) is also computationally challenging in that we have to run $B \times O(M^2)$ regression models to derive the t-statistic threshold for a given sample, where M is the total number of tests. This makes it difficult for us to implement the test to the 18,000 anomalies. We therefore randomly sample $N = 100$ times from the 18,113 anomalies, each time drawing 500 anomalies (which is similar to the size of the CAPIQ data). We then average across these random samples to evaluate the performance of Romano, Shaikh, and Wolf (2008).

Table A.1: **Simulated Error Rates for Romano, Shaikh, and Wolf (2008): CAPIQ and 18,000 Anomalies**

Simulated Type I and Type II error rates for CAPIQ and the 18,000 anomalies. For CAPIQ, for each p_0 of the fraction of strategies that are believed to be true, we follow our method in Section 2 and set $I = 100$ (for each i, we bootstrap to obtain the ranking of strategies and set the top p_0 as true) and $J = 1,00$ (conditional on i, for each j, we bootstrap the time periods) to run in total 10,000 ($= 100 \times 1,00$) bootstrapped simulations to calculate the empirical Type I and Type II error rates for Romano, Shaikh, and Wolf (2008). For the 18,113 anomalies, we first randomly sample 500 strategies $N = 100$ times. We then calculate the empirical Type I and Type II error rates for each random sample, and average across these random samples to obtain the averaged Type I and Type II error rates. A bold number indicates the highest (and oversized) Type I error rate among all methods considered in Table 2 and Table 3.

		Type I		Type II	
p_0	α	CAPIQ	18,000 anomalies*	CAPIQ	18,000 anomalies
(frac. of true)	(sig. level)				
2%	1%	0.017	**0.021**	0.002	0.008
	5%	0.045	**0.084**	0.001	0.006
	10%	**0.142**	**0.146**	0.001	0.005
5%	1%	0.008	**0.020**	0.005	0.028
	5%	0.054	**0.080**	0.002	0.019
	10%	**0.119**	**0.146**	0.001	0.016
10%	1%	0.008	0.016	0.019	0.070
	5%	0.060	**0.074**	0.007	0.050
	10%	**0.125**	**0.129**	0.004	0.049
20%	1%	0.006	**0.021**	0.078	0.168
	5%	0.055	0.057	0.039	0.137
	10%	0.114	**0.120**	0.026	0.111
30%	1%	0.004	**0.018**	0.179	0.265
	5%	0.047	0.060	0.115	0.244
	10%	0.093	0.113	0.091	0.201

B Dissecting the Fama and French Approach

The Fama and French (2010) approach has the clear advantage of incorporating cross-sectional information. But why does it have no power? We follow a two-step procedure to probe into the Fama and French approach. In the first step, we examine the impact of Fama and French's bootstrapping method on the distribution of t-statistic at the fund level. In the second step, we analyze test power by artificially increasing the signal to noise ratio for the mutual fund data.

One important difference between the Fama and French bootstrapping approach and the traditional bootstrap is that if a fund does not exist for the entire sample period, the number of observations for the bootstrapped sample may differ from the number of observations for the actual sample, which may lead to a difference in the distribution of t-statistics for the bootstrapped samples and the distribution of t-statistics if bootstrapping is only performed over existing periods for the fund (i.e., the traditional approach). Fama and French (2010) are aware of this difference and they claim it is not a serious issue for their approach.[53] Their argument is that in a given simulation run, although funds that are oversampled (undersampled) have more (less) degrees of freedom and thus thinner (fatter) extreme tails than the distributions for the actual returns of the funds, oversampling of some funds should roughly offset undersampling of others, creating a cross-sectional distribution of t-statistics that has similar properties to the one generated with actual fund returns.[54]

One potential issue in Fama and French (2010)'s argument is that while it is true that the number of oversampled funds should approximately equal the number of undersampled funds in a simulation run, the impact on the individual t-statistic distributions (and hence the cross-sectional distribution of t-statistics) could be very different between oversampling and undersampling. In particular, given that a t-distribution with a degree of freedom of D converges to a standard normal distribution when D is large, oversampling should not be as much of a concern as undersampling. For example, for a fund with $T = 24$ actual returns, oversampling the fund's returns (e.g., $T = 36$) is likely not a serious issue as both $T = 24$ and $T = 36$ generate similar distributions for the t-statistic. In contrast, undersampling (e.g., $T = 12$) leads to a distribution with a much fatter tail than a normal distribution, which poses a problem for the Fama and French method. We therefore conjecture that the asymmetric impact on the distribution of t-statistic between oversampling and undersampling contributes to the poor performance of the Fama and French approach.

To test our hypothesis, we examine the bootstrapped distribution of t-statistic for several selected funds. In particular, for a given T, we randomly select a fund with approximately T monthly observations.[55] Focusing on this fund, we first generate the corresponding zero-alpha fund by subtracting its in-sample alpha estimate from its returns (following the Fama

[53] As mentioned earlier, Fama and French (2010) require that a fund needs to have at least eight monthly observations (both for the actual sample and the bootstrapped sample) to be able to calculate its t-statistic.

[54] See the third paragraph in Fama and French (2010, p. 1925).

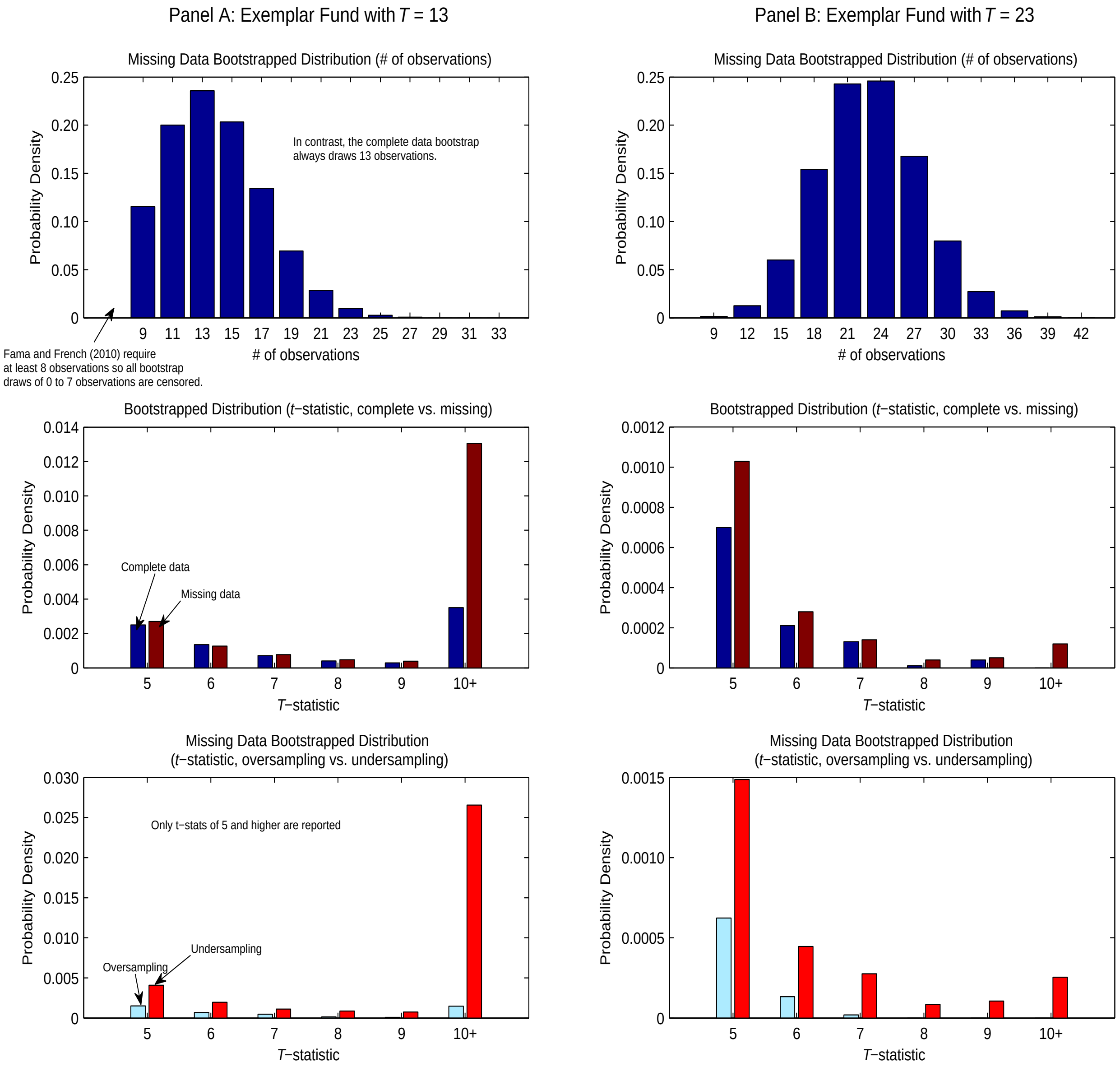
[55] Our results are not specific to the funds that we select. We also perform our analysis by randomly selecting 10 funds that have a sample size close to T and calculate the averaged bootstrapped distribution. It displays a similar pattern to the one for the selected fund that we choose to focus on.

and French approach) and then produce three sets of distributions by bootstrapping one million times. In the first set, we generate the distribution for the number of observations in the bootstrapped samples by following the Fama and French approach. In the second set, we compare the bootstrapped distribution of t-statistic between the traditional approach which we will refer to as the "complete data" bootstrap that only resamples the actual fund returns and the Fama and French method which we will call the "missing data" bootstrap that resamples all time periods, including those for which the fund has missing observations. In the last set, we focus on the Fama and French approach by decomposing its bootstrapped distribution of t-statistic into two separate distributions, one conditional on the number of observations drawn no fewer than T (i.e., oversampling) and the other conditional on undersampling.

Figure B.1 reports the results for two funds with $T \leq 24$ and Figure B.2 for two funds with $24 < T \leq 60$. Panel A of Figure B.1 shows the bootstrapped distributions for a fund with $T = 13$, i.e., roughly one-year of data. The top graph (i.e., bootstrapped distribution for the number of observations) peaks at 13 and is truncated at 8, which is minimum number set by Fama and French. There is a large amount of variation in the bootstrapped number of observations, ranging from 8 to around 33. Given the truncation at eight, the distribution is skewed to the right, implying a higher chance for oversampling than undersampling. However, as we argued before, this does not mean that oversampling will have a larger impact on the bootstrapped distribution than undersampling since the distortion in the distribution of t-statistic (relative to the distribution based on actual fund returns) could be disportionately larger with undersampling than with oversampling.

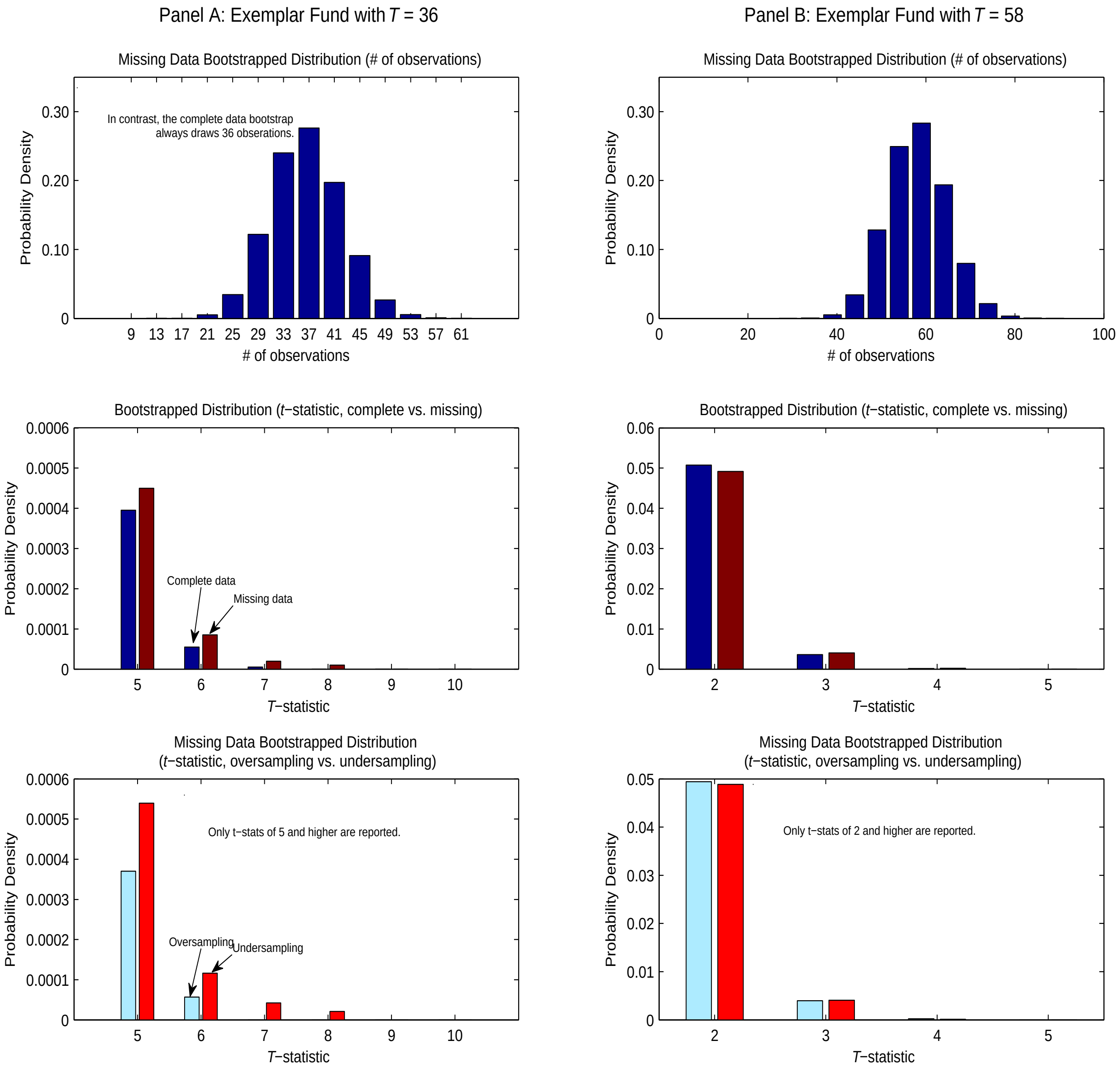
The middle graph in Figure B.1's Panel A (i.e., the complete data vs. missing data comparison in the bootstrapped distribution of t-statistic) shows how the missing data approach distorts the distribution of t-statistics. We focus on large realizations (i.e., t-statistics ≥ 5) of the t-statistic as they are more relevant to the Fama and French approach, which examines the right tail of the cross-sectional distribution of t-statistics. We also winsorize the distribution at 10 to better summarize information in the right tail as the distribution of t-statistic is rather dispersed when t-statistic is larger than 10. While the two distributions are similar for moderately large t-statistics (i.e., t-statistic between 5 and 9), the distribution generated by the Fama and French approach (i.e., missing data distribution) implies a much higher probability for t-statistics ≥ 10 than the complete data distribution.

Figure B.1: **Bootstrapped Distributions for Selected Mutual Funds** ($T \leq 24$)



Bootstrapped distributions for selected mutual funds ($T \leq 24$). We compare the bootstrapped distributions corresponding to the "complete data" bootstrap (traditional approach) and "missing data" bootstrap (Fama and French approach). For each bootstrapping approach, we resample one million times. For each panel, we plot the bootstrapped distribution for the number of observations for the missing data bootstrap in the top figure, the distributions for the bootstrapped t-statistics for both approaches in the middle figure, and the conditional distributions for the bootstrapped t-statistics corresponding to oversampling (i.e., bootstrap sample $\geq T$) and undersampling (i.e., bootstrap sample $< T$) for the missing data bootstrap in the bottom figure. For the top figure, the number of observations is truncated at 8 based on Fama and French (2010). For the middle and bottom figures, t-statistics with a value of five and above are reported and truncated at 10.

Figure B.2: **Bootstrapped Distributions for Selected Mutual Funds ($T > 24$)**



Bootstrapped distributions for selected mutual funds ($T > 24$). We compare the bootstrapped distributions corresponding to the "complete data" bootstrap (traditional approach) and "missing data" bootstrap (Fama and French approach). For each bootstrapping approach, we resample one million times. For each panel, we plot the bootstrapped distribution for the number of observations for the missing data bootstrap in the top figure, the distributions for the bootstrapped t-statistics for both approaches in the middle figure, and the conditional distributions for the bootstrapped t-statistics corresponding to oversampling (i.e., bootstrap sample $\geq T$) and undersampling (i.e., bootstrap sample $< T$) for the missing data bootstrap in the bottom figure. For the top figure, the number of observations is truncated at 8 based on Fama and French (2010). For the middle and bottom figures, t-statistics with a value of five and above are reported and truncated at 10 for panel A, and t-statistics with a value of two and above are reported and truncated at 5 for panel B.

The bottom graph in Figure B.1's Panel A shows the oversampling vs. undersampling decomposition of the Fama and French distribution shown in the middle graph. In particular, conditional on undersampling, the probabilities for generating a large t-statistic is uniformly larger than under oversampling. Importantly, focusing on the group with t-statistics ≥ 10, the combined evidence in the bottom graph and the middle graph shows that the disproportionately larger probability of generating a t-statistic ≥ 10 with undersampling more than offsets the smaller probability with oversampling, resulting in the overall much higher probability for t-statistics ≥ 10 shown in the middle graph.

How are the large t-statistics generated? First of all, in our simulation procedure, we follow Fama and French (2010) and discard simulation runs for which the fund has less than eight monthly returns. Fama and French claim that this ensures a sufficiently large degree of freedom to prevent the occurrence of extreme t-statistics. However, a large degree of freedom is not enough to moderate the simulated t-statistic distribution because bootstrapping with replacement implies that the number of independent (i.e., unique) observations may be small. For example, for a fund with $T = 13$ for the actual sample, it may have $T = 13$ months of returns in a simulation run but only six unique observations. Such a sample with a small number of unique observations, when projected onto five regressors (i.e., a constant and four independent variables), tends to generate an extreme t-statistic for the intercept.[56]

The tilt towards more extreme t-statistics for the bootstrapped distribution implied by the Fama and French approach is not limited to the case with a T around 12. Panel B of Figure B.1 and Panel A in Figure B.2 show the case for $T = 23$ and $T = 36$, respectively. Even for these funds with a relatively large sample size, there are still apparent distortions in the bootstrapped distribution of t-statistics. Eventually, when T gets even larger (Panel B of Figure B.2 shows a case with $T = 58$), this distortion seems dampened, making the Fama and French bootstrapped distribution a better approximation to the distribution based on actual returns.[57]

Our analysis so far focuses on the implication of the Fama and French approach at the fund level. For a given fund, the Fama and French bootstrapped distribution of t-statistic is more fat-tailed compared to the distribution for actual returns for funds with a relatively short sample period (e.g., $T \leq 36$). This fact helps explain the low power of the Fama and French approach. In the Fama and French approach where we resample the time periods to generate the cross-sectional distribution of t-statistics for zero-alpha funds, since extreme t-statistics occur more frequently for funds with a relatively short sample period for this approach compared to the traditional bootstrap (i.e., bootstrapping with complete data) and there are a non-negligible fraction of these funds in the data, the cross-sectional distribution of t-statistics for Fama and French also features more extreme tails than the traditional

[56] In fact, for funds with a small T (e.g., $T = 13$), a significant fraction of simulation runs generate missing values (i.e., values that are too large to store) because the number of independent observations is too small. We discard these values when plotting Figure B.1 and B.2.

[57] Note that when T is large as in Panel B of Figure B.2, we do not observe any t-statistic realization that is larger than 5. We therefore set the upper bound at 5 to plot the distributions of t-statistic.

bootstrap.[58] These extreme tails are an exaggeration of what the tails of the cross-sectional distribution should be if we had complete data for each fund. These exaggerated tails make it difficult for the tails of the actual cross-sectional distribution of t-statistics to surpass, leading to the low rejection rate (i.e., low test power) even when some funds are actually outperforming.

Another fact that likely exacerbates the low power issue for the Fama and French approach is that extreme t-statistics for the bootstrapped cross-sectional distribution of t-statistics are more likely generated by funds with a shorter sample period than those with a longer time period (even if we focus on complete data bootstrap), simply because a t-statistic with a smaller degree of freedom is more fat-tailed. Indeed, our examples in Figure B.1 and B.2 show that although funds with $T \leq 36$ achieve a t-statistic of 8 or above with a non-negligible probability (even under complete data bootstrap), the fund with a T around 60 never achieves a t-statistic exceeding 5 in our simulations. However, funds with a shorter sample period are exactly those that are affected the most by the Fama and French resampling approach. Hence, the fact that the mutual fund data features a substantial fraction of funds with a relatively short sample period contributes to the low power problem of the Fama and French approach.[59]

While we have identified undersampling of funds with a relatively short sample period for the Fama and French approach as a contributor to its low power, there are potentially other contributors. Similar to studies that examine test power, we next investigate how sensitive test power is to the signal-to-noise ratio in the mutual fund data.

In particular, we obtain the alpha estimates for all funds for the actual mutual fund data and multiply them by a factor of F to obtain new alpha estimates. We take these new alpha estimates as the in-sample alpha estimates for funds and repeat our analysis. Note that multiplying the alpha estimate by a scaling factor of F is equivalent to multiplying the in-sample t-statistic (i.e., signal-to-noise ratio) by a factor of F. By varying F, we evaluate how test power varies when the signal-to-noise ratio varies in the data. Table IB.4 and IB.5 in Internet Appendix B report the results for $F = 1.25$ and 1.5. When $F = 1.25$ (i.e., signal-to-noise ratio is 25% higher than that for the actual data), the Type II error rate for the Fama-French approach is substantially reduced. For example, at 5% significance level and assuming $p_0 = 5\%$, the Type II error rate for the 98^{th}-percentile is reduced from 33.6% (Table 6) to 12.7% (Table IB.4). When $F = 1.5$, the Type II error rate for the same test statistic is further reduced to 2.3% (Table IB.5).

Our results show that, not surprisingly, the signal-to-noise ratio in estimating fund alphas has an important impact on test power for the Fama-French approach. While the bootstrap procedure in Fama and French adjusts for multiple tests and successfully controls the Type I error rate at a pre-specified level, the low signal-to-noise ratio in the data makes it difficult for their procedure to detect outperforming funds.

[58] Our results apply to both the left and the right tails of the cross-sectional distribution of t-statistics. Given our focus on testing outperforming funds, we limit our interpretation to the right tail.

[59] The fraction of funds that have a sample size no greater than T is 2.0% ($T = 12$), 12.7% ($T = 36$), and 22.3% ($T = 60$) for the data we use.

We have thus far identified two drivers that contribute to the low test power of the Fama and French approach: the undersampling of funds with a relatively short sample period and the low signal-to-noise ratio in the mutual fund data. Unfortunately, there is little that can be done to fix the signal to noise problem. However, undersampling of funds with a short history can perhaps be mitigated to improve the performance of the Fama and French approach.

False (and Missed) Discoveries in Financial Economics
Online Appendix

Campbell R. Harvey

Duke University, Durham, NC 27708 USA
National Bureau of Economic Research, Cambridge, MA 02138 USA

Yan Liu[*]

Texas A&M University, College Station, TX 77843 USA

Current version: August 10, 2019

[*] Current Version: August 10, 2019. First posted on SSRN: November 19, 2017. Send correspondence to: Yan Liu, Mays Business School, Texas A&M University, College Station, TX 77843. Phone: +1 919.428.1118, E-mail: y-liu@mays.tamu.edu. The full version of the paper is available at https://ssrn.com/abstract=3073799.

A Additional Results for the Simulation Study

Table IA.1: **A Simulation Study on CAPIQ: Mean Return Distribution for True Strategies = A Gamma Distribution with A Standard Deviation of 2.5%**

Simulated Type I error rates for CAPIQ when the mean return distribution for true strategies (F) follows a Gamma distribution. The simulation study runs as follows. We fix the fraction of true strategies at 10%. We first randomly identify 10% of strategies as true and assign mean returns to them according to F. Mean returns are set at zero for the remaining 90% of strategies. Let D_m denote the final data ($m = 1, 2, \ldots, M = 400$). Conditional on D_m, we bootstrap the time periods to generate the perturbed in-sample data $D_{m,k}$ ($k = 1, 2, \ldots, K = 100$). For each $D_{m,k}$ and for a given multiple testing method at a pre-specified significance level α, we calculate the true realized error rate (denoted as $FDR^a_{m,k}$). Implementing our approach (with a prior specification of p_0), we obtain the estimated error rate (denoted as $FDR^e_{m,k}$). We report the mean true Type I error rate ('Actual') defined as $\frac{1}{MK}\sum_{m=1}^{M}\sum_{k=1}^{K} FDR^a_{m,k}$ and mean error rate for our estimator ('Est.') defined as $\frac{1}{MK}\sum_{m=1}^{M}\sum_{k=1}^{K} FDR^e_{m,k}$. δ denotes the nominal error rate. Distribution F is a Gamma distribution with mean μ_0 ($\mu_0 = 2.5\%$, 5.0% or 10%) and standard deviation $\sigma_0 = 2.5\%$. A bold number indicates a better performance of our model, i.e., 'Est.' is closer to 'Actual' compared to 'δ'.

		$\mu_0 = 2.5\%$				$\mu_0 = 5.0\%$				$\mu_0 = 10\%$			
Method	δ	Actual	Est.			Actual	Est.			Actual	Est.		
			p_0=5%	p_0=10%	p_0=15%		p_0=5%	p_0=10%	p_0=15%		p_0=5%	p_0=10%	p_0=15%
BH	1%	0.0112	**0.0115**	**0.0109**	**0.0104**	0.0103	0.0109	**0.0105**	**0.0102**	0.0104	0.0110	**0.0108**	**0.0102**
	5%	0.0443	**0.0453**	**0.0435**	**0.0413**	0.0447	**0.0465**	**0.0446**	**0.0422**	0.0460	**0.0476**	**0.0458**	**0.0433**
	10%	0.0815	**0.0835**	**0.0805**	**0.0767**	0.0840	**0.0867**	**0.0836**	**0.0796**	0.0873	**0.0891**	**0.0867**	**0.0822**
BHY	1%	0.0021	**0.0025**	**0.0024**	**0.0022**	0.0016	**0.0024**	**0.0022**	**0.0021**	0.0020	**0.0022**	**0.0020**	**0.0019**
	5%	0.0085	**0.0089**	**0.0085**	**0.0080**	0.0079	**0.0087**	**0.0083**	**0.0078**	0.0081	**0.0088**	**0.0083**	**0.0078**
	10%	0.0153	**0.0160**	**0.0152**	**0.0144**	0.0142	**0.0159**	**0.0151**	**0.0143**	0.0148	**0.0161**	**0.0152**	**0.0144**
Storey ($\theta = 0.4$)	1%	0.0120	0.0140	**0.0137**	0.0143	0.0157	**0.0151**	**0.0148**	**0.0144**	0.0130	**0.0150**	**0.0148**	**0.0144**
	5%	0.0527	0.0564	0.0557	**0.0552**	0.0606	**0.0589**	**0.0585**	**0.0572**	0.0592	**0.0602**	**0.0604**	**0.0591**
	10%	0.0961	0.1005	**0.0998**	**0.0995**	0.1103	**0.1068**	**0.1068**	**0.1047**	0.1115	**0.1092**	**0.1108**	**0.1089**
Storey ($\theta = 0.6$)	1%	0.0117	0.0135	**0.0133**	0.0138	0.0154	**0.0145**	**0.0143**	**0.0141**	0.0127	**0.0145**	**0.0143**	**0.0141**
	5%	0.0532	0.0567	**0.0560**	**0.0556**	0.0608	**0.0584**	**0.0586**	**0.0578**	0.0591	**0.0596**	**0.0602**	**0.0596**
	10%	0.0974	0.1011	0.1003	**0.1000**	0.1118	**0.1074**	**0.1084**	**0.1074**	0.1131	**0.1095**	**0.1123**	**0.1114**
Storey ($\theta = 0.8$)	1%	0.0111	0.0129	0.0126	0.0128	0.0144	**0.0133**	**0.0133**	**0.0132**	0.0120	**0.0133**	**0.0133**	**0.0133**
	5%	0.0521	0.0543	**0.0540**	**0.0541**	0.0605	**0.0574**	**0.0587**	**0.0581**	0.0593	**0.0585**	**0.0604**	**0.0609**
	10%	0.1014	0.1034	0.1040	0.1040	0.1185	**0.1105**	**0.1229**	**0.1156**	0.1199	**0.1127**	**0.1185**	**0.1202**

Table IA.2: **A Simulation Study on CAPIQ: Mean Return Distribution for True Strategies = A Gamma Distribution with A Standard Deviation of 5.0%**

Simulated Type I error rates for CAPIQ when the mean return distribution for true strategies (F) follows a Gamma distribution. The simulation study runs as follows. We fix the fraction of true strategies at 10%. We first randomly identify 10% of strategies as true and assign mean returns to them according to F. Mean returns are set at zero for the remaining 90% of strategies. Let D_m denote the final data ($m = 1, 2, \ldots, M = 400$). Conditional on D_m, we bootstrap the time periods to generate the perturbed in-sample data $D_{m,k}$ ($k = 1, 2, \ldots, K = 100$). For each $D_{m,k}$ and for a given multiple testing method at a pre-specified significance level α, we calculate the true realized error rate (denoted as $FDR^a_{m,k}$). Implementing our approach (with a prior specification of p_0), we obtain the estimated error rate (denoted as $FDR^e_{m,k}$). We report the mean true Type I error rate ('Actual') defined as $\frac{1}{MK}\sum_{m=1}^{M}\sum_{k=1}^{K} FDR^a_{m,k}$ and mean error rate for our estimator ('Est.') defined as $\frac{1}{MK}\sum_{m=1}^{M}\sum_{k=1}^{K} FDR^e_{m,k}$. δ denotes the nominal error rate. Distribution F is a Gamma distribution with mean μ_0 ($\mu_0 = 2.5\%$, 5.0% or 10%) and standard deviation $\sigma_0 = 5.0\%$. A bold number indicates a better performance of our model, i.e., 'Est.' is closer to 'Actual' compared to 'δ'.

		$\mu_0 = 2.5\%$				$\mu_0 = 5.0\%$				$\mu_0 = 10\%$			
Method	δ	Actual	Est.			Actual	Est.			Actual	Est.		
			p_0=5%	p_0=10%	p_0=15%		p_0=5%	p_0=10%	p_0=15%		p_0=5%	p_0=10%	p_0=15%
BH	1%	0.0096	0.0103	**0.0098**	**0.0099**	0.0104	0.0111	**0.0108**	**0.0102**	0.0105	0.0111	**0.0107**	**0.0101**
	5%	0.0395	**0.0454**	**0.0434**	**0.0412**	0.0455	**0.0470**	**0.0449**	**0.0425**	0.0463	**0.0477**	**0.0458**	**0.0433**
	10%	0.0750	**0.0836**	**0.0804**	**0.0765**	0.0853	**0.0873**	**0.0840**	**0.0798**	0.0875	**0.0891**	**0.0864**	**0.0820**
BHY	1%	0.0019	**0.0024**	**0.0023**	**0.0022**	0.0018	**0.0022**	**0.0021**	**0.0020**	0.0018	**0.0022**	**0.0020**	**0.0019**
	5%	0.0070	**0.0088**	**0.0084**	**0.0080**	0.0081	**0.0088**	**0.0084**	**0.0079**	0.0080	**0.0088**	**0.0082**	**0.0078**
	10%	0.0133	**0.0158**	**0.0150**	**0.0143**	0.0150	**0.0161**	**0.0153**	**0.0145**	0.0153	**0.0161**	**0.0153**	**0.0144**
Storey ($\theta = 0.4$)	1%	0.0140	**0.0149**	**0.0146**	**0.0141**	0.0142	**0.0151**	**0.0148**	**0.0144**	0.0142	**0.0151**	**0.0149**	**0.0145**
	5%	0.0541	**0.0571**	**0.0552**	**0.0547**	0.0569	**0.0591**	**0.0584**	**0.0569**	0.0575	**0.0603**	**0.0604**	**0.0591**
	10%	0.0984	0.1008	**0.0996**	**0.0992**	0.1043	**0.1068**	**0.1061**	**0.1038**	0.1074	**0.1097**	**0.1107**	**0.1087**
Storey ($\theta = 0.6$)	1%	0.0137	**0.0144**	**0.0142**	**0.0139**	0.0139	**0.0146**	**0.0144**	**0.0141**	0.0139	**0.0145**	**0.0144**	**0.0142**
	5%	0.0541	**0.0566**	**0.0562**	**0.0552**	0.0573	**0.0585**	**0.0583**	**0.0573**	0.0581	**0.0597**	**0.0602**	**0.0595**
	10%	0.0993	0.1012	0.1001	0.1006	0.1062	**0.1073**	**0.1076**	**0.1063**	0.1095	**0.1101**	**0.1122**	**0.1113**
Storey ($\theta = 0.8$)	1%	0.0127	**0.0132**	**0.0131**	**0.0130**	0.0130	**0.0133**	**0.0133**	**0.0132**	0.0131	**0.0133**	**0.0134**	**0.0134**
	5%	0.0535	**0.0553**	**0.0558**	**0.0557**	0.0577	**0.0574**	**0.0581**	**0.0582**	0.0592	**0.0554**	**0.0606**	**0.0622**
	10%	0.1017	0.1044	0.1034	0.1040	0.1116	**0.1100**	**0.1129**	**0.1139**	0.1170	**0.1132**	**0.1184**	**0.1200**

B Additional Results for Fama and French (2010)

Table IB.1: **Summary Statistics for Mutual Fund Returns**

Summary statistics for mutual fund returns in our samplel. For funds with an initial AUM exceeding a certain threshold (2006 dollars), we calculate fund alphas based on different benchmark factor models and report the mean, standard deviation ("Stdev."), as well as several percentiles of the cross-section of alphas and their *t*-statistics. The sample period is from 1984 to 2006.

Model		Mean	Stdev.	5%	10%	50%	90%	95%
Panel A: AUM $\geq$ 5M								
CAPM	alpha	-0.005	0.068	-0.105	-0.074	-0.005	0.064	0.093
	alpha t-stat	-0.184	1.503	-2.675	-2.075	-0.159	1.693	2.339
3-factor	alpha	-0.014	0.059	-0.099	-0.065	-0.012	0.038	0.060
	alpha t-stat	-0.553	1.326	-2.763	-2.273	-0.548	1.142	1.562
4-factor	alpha	-0.016	0.062	-0.097	-0.065	-0.013	0.033	0.056
	alpha t-stat	-0.561	1.285	-2.691	-2.187	-0.548	1.099	1.539
Panel B: AUM $\geq$ 250M								
CAPM	alpha	-0.003	0.067	-0.091	-0.063	-0.005	0.056	0.082
	alpha t-stat	-0.186	1.412	-2.585	-2.015	-0.153	1.597	2.119
3-factor	alpha	-0.014	0.070	-0.090	-0.059	-0.013	0.032	0.055
	alpha t-stat	-0.612	1.264	-2.759	-2.254	-0.585	1.037	1.447
4-factor	alpha	-0.015	0.076	-0.085	-0.059	-0.014	0.029	0.053
	alpha t-stat	-0.627	1.238	-2.713	-2.255	-0.628	0.986	1.445
Panel C: AUM $\geq$ 1B								
CAPM	alpha	-0.009	0.051	-0.096	-0.069	-0.006	0.043	0.064
	alpha t-stat	-0.310	1.398	-2.805	-2.131	-0.223	1.450	1.933
3-factor	alpha	-0.016	0.054	-0.087	-0.061	-0.015	0.027	0.058
	alpha t-stat	-0.664	1.304	-2.817	-2.278	-0.662	0.983	1.454
4-factor	alpha	-0.015	0.056	-0.085	-0.059	-0.013	0.027	0.055
	alpha t-stat	-0.642	1.309	-2.754	-2.260	-0.617	1.087	1.414

Table IB.2: **Simulated Error Rates for Fama and French (2010): AUM $\geq$ 250M**

Simulated Type I and Type II error rates for the Fama and French (2010) approach. For each p_0 of the fraction of strategies that are believed to be true, we follow our method in Section 2.3 and perturb the adjusted data $M = 1,000$ times to calculate the empirical Type I and Type II error rates. We consider six percentiles as well as the maximum. "Avg." and "Avg. t-stat" report the average (across simulations) of the mean alpha and alpha t-statistic for the fraction p_0 of funds that are assumed to be outperforming. All funds with an initial AUM exceeding \$250 million are included, resulting in 1,631 funds in total. The sample period is from 1984 to 2006. Panel A reports the Type I error rate, that is, the probability of falsely declaring that some managers have skill when no fund is outperforming. Panel B reports the Type II error rate, that is, the probability of falsely declaring that no manager is skilled when some funds are outperforming.

				Test Statistics						
p_0	Avg.	Avg. t-stat	α	Max	99.9%	99.5%	99%	98%	95%	90%
(frac. of true)	alpha	of alpha	(sig. level)							
Panel A: Type I Error Rate										
0	0	0	1%	0.006	0.002	0.002	0.001	0.000	0.004	0.003
			5%	0.038	0.015	0.002	0.005	0.026	0.039	0.046
			10%	0.079	0.050	0.005	0.020	0.056	0.078	0.102
Panel B: Type II Error Rate										
0.5%	16.06	3.80	1%	0.992	0.997	0.997	0.999	0.999	0.993	0.994
			5%	0.956	0.978	0.998	0.990	0.963	0.957	0.950
			10%	0.915	0.948	0.993	0.966	0.926	0.909	0.888
1%	15.98	3.57	1%	0.991	0.994	0.998	0.998	0.998	0.990	0.991
			5%	0.950	0.971	0.998	0.985	0.950	0.951	0.949
			10%	0.911	0.943	0.989	0.950	0.907	0.896	0.879
2%	10.53	3.15	1%	0.991	0.996	0.998	0.986	0.992	0.988	0.987
			5%	0.945	0.976	0.997	0.929	0.938	0.942	0.934
			10%	0.903	0.945	0.977	0.834	0.864	0.866	0.865
3%	9.68	2.96	1%	0.994	0.995	0.996	0.955	0.933	0.971	0.977
			5%	0.950	0.973	0.992	0.758	0.773	0.882	0.911
			10%	0.914	0.942	0.971	0.562	0.611	0.787	0.831
5%	7.76	2.58	1%	0.982	0.996	0.997	0.938	0.787	0.957	0.975
			5%	0.935	0.960	0.998	0.662	0.509	0.792	0.879
			10%	0.881	0.931	0.963	0.441	0.357	0.694	0.778
10%	5.75	2.05	1%	0.988	0.996	0.999	0.905	0.615	0.852	0.946
			5%	0.943	0.969	0.993	0.604	0.311	0.628	0.795
			10%	0.901	0.931	0.959	0.352	0.155	0.447	0.686
15%	4.69	1.72	1%	0.986	0.991	0.998	0.906	0.483	0.587	0.825
			5%	0.932	0.969	0.994	0.538	0.166	0.259	0.561
			10%	0.891	0.925	0.965	0.269	0.049	0.142	0.394

Table IB.3: **Simulated Error Rates for Fama and French (2010): AUM $\geq$ 1B**

Simulated Type I and Type II error rates for the Fama and French (2010) approach. For each p_0 of the fraction of strategies that are believed to be true, we follow our method in Section 2.3 and perturb the adjusted data $M = 1,000$ times to calculate the empirical Type I and Type II error rates. We consider six percentiles as well as the maximum. "Avg." and "Avg. t-stat" report the average (across simulations) of the mean alpha and alpha t-statistic for the fraction p_0 of funds that are assumed to be outperforming. All funds with an initial AUM exceeding \$1 B are included, resulting in 823 funds in total. The sample period is from 1984 to 2006. Panel A reports the Type I error rate, that is, the probability of falsely declaring that some managers have skill when no fund is outperforming. Panel B reports the Type II error rate, that is, the probability of falsely declaring that no manager is skilled when some funds are outperforming.

				Test Statistics						
p_0	Avg.	Avg. t-stat	α	Max	99.9%	99.5%	99%	98%	95%	90%
(frac. of true)	alpha	of alpha	(sig. level)							
Panel A: Type I Error Rate										
0	0	0	1%	0.009	0.007	0.012	0.006	0.000	0.004	0.009
			5%	0.039	0.033	0.030	0.017	0.001	0.024	0.034
			10%	0.070	0.070	0.057	0.026	0.014	0.040	0.066
Panel B: Type II Error Rate										
0.5%	14.45	3.56	1%	0.990	0.991	0.988	0.994	1.000	0.996	0.992
			5%	0.957	0.962	0.966	0.980	0.997	0.976	0.965
			10%	0.921	0.923	0.937	0.964	0.980	0.952	0.928
1%	12.14	3.13	1%	0.988	0.988	0.988	0.993	0.999	0.996	0.993
			5%	0.954	0.958	0.964	0.973	0.996	0.976	0.965
			10%	0.914	0.917	0.934	0.956	0.976	0.946	0.923
2%	11.47	3.13	1%	0.985	0.990	0.984	0.994	0.998	0.988	0.987
			5%	0.944	0.955	0.963	0.961	0.981	0.977	0.961
			10%	0.912	0.906	0.932	0.947	0.942	0.921	0.902
3%	9.06	3.02	1%	0.982	0.985	0.981	0.990	0.990	0.984	0.985
			5%	0.925	0.936	0.963	0.966	0.911	0.946	0.955
			10%	0.902	0.901	0.928	0.946	0.809	0.878	0.868
5%	7.60	2.62	1%	0.973	0.978	0.979	0.992	0.976	0.976	0.975
			5%	0.934	0.944	0.955	0.961	0.814	0.890	0.928
			10%	0.895	0.896	0.928	0.931	0.664	0.781	0.833
10%	5.66	2.14	1%	0.982	0.985	0.985	0.995	0.956	0.929	0.965
			5%	0.943	0.948	0.955	0.959	0.709	0.698	0.859
			10%	0.913	0.901	0.927	0.929	0.450	0.503	0.714
15%	4.56	1.73	1%	0.986	0.987	0.977	0.991	0.905	0.648	0.862
			5%	0.946	0.955	0.966	0.961	0.597	0.318	0.592
			10%	0.913	0.911	0.936	0.915	0.348	0.177	0.422

Table IB.4: **Simulated Error Rates for Fama and French (2010): Injected Alpha = 1.25× Original Alpha**

Simulated Type I and Type II error rates for the Fama and French (2010) approach. We first obtain the in-sample alpha estimates for all funds in our sample and then multiply them by $F = 2.0$. For each p_0 of the fraction of strategies that are believed to be true, we follow our method in Section 2.3 and perturb the adjusted data $M = 1,000$ times to calculate the empirical Type I and Type II error rates. We consider six percentiles as well as the maximum. "Avg." and "Avg. t-stat" report the average (across simulations) of the mean alpha and alpha t-statistic for the fraction p_0 of funds that are assumed to be outperforming. All funds with an initial AUM exceeding $5 million are included, resulting in 3,030 funds in total. The sample period is from 1984 to 2006. Panel A reports the Type I error rate, that is, the probability of falsely declaring that some managers have skill when no fund is outperforming. Panel B reports the Type II error rate, that is, the probability of falsely declaring that no manager is skilled when some funds are outperforming.

				Test Statistics						
p_0 (frac. of true)	Avg. alpha	Avg. t-stat of alpha	α (sig. level)	Max	99.9%	99.5%	99%	98%	95%	90%
Panel A: Type I Error Rate										
0	0	0	1%	0.005	0.002	0.000	0.000	0.000	0.010	0.011
			5%	0.024	0.017	0.002	0.001	0.017	0.028	0.044
			10%	0.049	0.036	0.002	0.003	0.037	0.073	0.093
Panel B: Type II Error Rate										
0.5%	18.21	4.95	1%	0.994	0.998	1.000	1.000	0.999	0.987	0.987
			5%	0.971	0.979	0.993	0.989	0.977	0.964	0.953
			10%	0.945	0.961	0.982	0.965	0.941	0.920	0.898
1%	16.54	4.61	1%	0.993	0.999	0.999	0.999	0.998	0.985	0.987
			5%	0.966	0.976	0.987	0.979	0.971	0.955	0.950
			10%	0.935	0.960	0.968	0.933	0.921	0.911	0.887
2%	13.30	4.09	1%	0.993	0.998	0.999	0.933	0.986	0.985	0.982
			5%	0.970	0.983	0.988	0.751	0.911	0.940	0.940
			10%	0.944	0.963	0.935	0.544	0.814	0.865	0.855
3%	12.17	3.77	1%	0.994	0.999	1.000	0.674	0.798	0.971	0.977
			5%	0.967	0.984	0.977	0.216	0.461	0.880	0.911
			10%	0.936	0.962	0.892	0.078	0.261	0.744	0.787
5%	10.16	3.29	1%	0.993	0.994	0.998	0.575	0.442	0.938	0.966
			5%	0.964	0.980	0.974	0.131	0.127	0.740	0.865
			10%	0.937	0.959	0.900	0.036	0.050	0.559	0.729
10%	7.86	2.63	1%	0.992	0.999	1.000	0.472	0.165	0.716	0.928
			5%	0.968	0.981	0.978	0.070	0.016	0.360	0.708
			10%	0.931	0.959	0.888	0.009	0.003	0.179	0.581
15%	6.49	2.26	1%	0.990	0.996	0.999	0.401	0.070	0.239	0.684
			5%	0.964	0.977	0.967	0.049	0.006	0.031	0.331
			10%	0.933	0.955	0.875	0.009	0.004	0.010	0.147

Table IB.5: **Simulated Error Rates for Fama and French (2010): Injected Alpha = 1.5× Original Alpha**

Simulated Type I and Type II error rates for the Fama and French (2010) approach. We first obtain the in-sample alpha estimates for all funds in our sample and then multiply them by $F = 1.5$. For each p_0 of the fraction of strategies that are believed to be true, we follow our method in Section 2.3 and perturb the adjusted data $M = 1,000$ times to calculate the empirical Type I and Type II error rates. We consider six percentiles as well as the maximum. "Avg." and "Avg. t-stat" report the average (across simulations) of the mean alpha and alpha t-statistic for the fraction p_0 of funds that are assumed to be outperforming. All funds with an initial AUM exceeding $5 million are included, resulting in 3,030 funds in total. The sample period is from 1984 to 2006. Panel A reports the Type I error rate, that is, the probability of falsely declaring that some managers have skill when no fund is outperforming. Panel B reports the Type II error rate, that is, the probability of falsely declaring that no manager is skilled when some funds are outperforming.

				Test Statistics						
p_0 (frac. of true)	Avg. alpha	Avg. t-stat of alpha	α (sig. level)	Max	99.9%	99.5%	99%	98%	95%	90%
Panel A: Type I Error Rate										
0	0	0	1%	0.002	0.002	0.001	0.000	0.000	0.002	0.007
			5%	0.023	0.016	0.002	0.000	0.008	0.034	0.057
			10%	0.043	0.026	0.005	0.001	0.046	0.092	0.110
Panel B: Type II Error Rate										
0.5%	21.70	5.88	1%	0.996	0.996	0.996	1.000	1.000	0.995	0.991
			5%	0.966	0.983	0.993	0.988	0.982	0.955	0.942
			10%	0.946	0.970	0.981	0.952	0.927	0.894	0.883
1%	19.82	5.53	1%	0.992	0.996	0.995	1.000	0.999	0.992	0.988
			5%	0.958	0.979	0.990	0.979	0.969	0.943	0.939
			10%	0.937	0.965	0.969	0.906	0.899	0.882	0.876
2%	15.98	4.94	1%	0.994	0.999	0.999	0.903	0.990	0.988	0.985
			5%	0.962	0.976	0.986	0.592	0.889	0.921	0.922
			10%	0.936	0.967	0.909	0.351	0.769	0.839	0.845
3%	14.63	4.53	1%	0.994	0.995	0.996	0.448	0.652	0.970	0.978
			5%	0.971	0.982	0.973	0.076	0.268	0.853	0.885
			10%	0.943	0.968	0.868	0.017	0.150	0.708	0.797
5%	12.20	3.96	1%	0.992	0.991	0.996	0.339	0.149	0.915	0.961
			5%	0.970	0.977	0.980	0.033	0.023	0.677	0.846
			10%	0.943	0.969	0.869	0.008	0.007	0.514	0.719
10%	9.40	3.13	1%	0.998	0.995	0.996	0.289	0.021	0.561	0.906
			5%	0.964	0.974	0.977	0.024	0.002	0.208	0.692
			10%	0.938	0.967	0.848	0.005	0.001	0.080	0.557
15%	7.77	2.69	1%	0.990	0.995	0.996	0.266	0.002	0.056	0.541
			5%	0.949	0.977	0.976	0.010	0.000	0.002	0.202
			10%	0.928	0.966	0.844	0.002	0.000	0.001	0.091

Table IB.6: **Simulated Error Rates for Fama and French (2010):** $T \geq 36$

Simulated Type I and Type II error rates for the Fama and French (2010) approach. For each p_0 of the fraction of strategies that are believed to be true, we follow our method in Section 2.3 and perturb the adjusted data $M = 1,000$ times to calculate the empirical Type I and Type II error rates. We consider six percentiles as well as the maximum. "Avg." and "Avg. t-stat" report the average (across simulations) of the mean alpha and alpha t-statistic for the fraction p_0 of funds that are assumed to be outperforming. All funds with an initial AUM exceeding $5 million are included, resulting in 3,030 funds in total. The sample period is from 1984 to 2006. Panel A reports the Type I error rate, that is, the probability of falsely declaring that some managers have skill when no fund is outperforming. Panel B reports the Type II error rate, that is, the probability of falsely declaring that no manager is skilled when some funds are outperforming.

				Test Statistics						
p_0	Avg.	Avg. t-stat	α	Max	99.9%	99.5%	99%	98%	95%	90%
(frac. of true)	alpha	of alpha	(sig. level)							
Panel A: Type I Error Rate										
0	0	0	1%	0.001	0.002	0.002	0.011	0.013	0.017	0.018
			5%	0.017	0.020	0.041	0.045	0.059	0.073	0.071
			10%	0.052	0.065	0.083	0.111	0.110	0.122	0.127
Panel B: Type II Error Rate										
0.5%	12.74	4.01	1%	0.675	0.342	0.918	0.979	0.982	0.979	0.981
			5%	0.255	0.112	0.682	0.877	0.902	0.909	0.917
			10%	0.111	0.046	0.506	0.766	0.829	0.847	0.853
1%	10.61	3.53	1%	0.646	0.277	0.509	0.895	0.964	0.975	0.981
			5%	0.243	0.059	0.220	0.635	0.852	0.890	0.907
			10%	0.099	0.014	0.108	0.443	0.751	0.818	0.832
2%	9.28	3.13	1%	0.669	0.225	0.321	0.533	0.881	0.967	0.965
			5%	0.257	0.044	0.068	0.210	0.616	0.853	0.881
			10%	0.088	0.011	0.020	0.089	0.431	0.731	0.799
3%	8.89	2.92	1%	0.674	0.178	0.236	0.365	0.633	0.921	0.958
			5%	0.196	0.023	0.027	0.085	0.324	0.751	0.836
			10%	0.084	0.003	0.011	0.020	0.177	0.621	0.734
5%	7.66	2.59	1%	0.669	0.179	0.184	0.232	0.402	0.835	0.925
			5%	0.234	0.022	0.028	0.046	0.128	0.535	0.759
			10%	0.073	0.003	0.008	0.019	0.044	0.336	0.627
10%	5.85	2.04	1%	0.668	0.160	0.122	0.181	0.244	0.528	0.782
			5%	0.219	0.015	0.009	0.007	0.045	0.195	0.499
			10%	0.078	0.007	0.001	0.000	0.007	0.089	0.340
15%	4.78	1.71	1%	0.666	0.153	0.142	0.167	0.215	0.417	0.641
			5%	0.207	0.020	0.012	0.026	0.047	0.145	0.320
			10%	0.074	0.003	0.003	0.000	0.011	0.052	0.208

Table IB.7: **Simulated Error Rates for Fama and French (2010):** $T \geq 60$

Simulated Type I and Type II error rates for the Fama and French (2010) approach. For each p_0 of the fraction of strategies that are believed to be true, we follow our method in Section 2.3 and perturb the adjusted data $M = 1,000$ times to calculate the empirical Type I and Type II error rates. We consider six percentiles as well as the maximum. "Avg." and "Avg. t-stat" report the average (across simulations) of the mean alpha and alpha t-statistic for the fraction p_0 of funds that are assumed to be outperforming. All funds with an initial AUM exceeding \$5 million are included, resulting in 3,030 funds in total. The sample period is from 1984 to 2006. Panel A reports the Type I error rate, that is, the probability of falsely declaring that some managers have skill when no fund is outperforming. Panel B reports the Type II error rate, that is, the probability of falsely declaring that no manager is skilled when some funds are outperforming.

				Test Statistics						
p_0 (frac. of true)	Avg. alpha	Avg. t-stat of alpha	α (sig. level)	Max	99.9%	99.5%	99%	98%	95%	90%
Panel A: Type I Error Rate										
0	0	0	1%	0.001	0.006	0.001	0.003	0.005	0.006	0.007
			5%	0.033	0.043	0.038	0.034	0.040	0.041	0.040
			10%	0.069	0.089	0.092	0.091	0.085	0.094	0.099
Panel B: Type II Error Rate										
0.5%	12.32	4.00	1%	0.376	0.243	0.898	0.988	0.990	0.992	0.987
			5%	0.117	0.048	0.668	0.879	0.928	0.950	0.956
			10%	0.044	0.018	0.492	0.797	0.857	0.880	0.884
1%	11.02	3.62	1%	0.330	0.159	0.513	0.882	0.981	0.989	0.983
			5%	0.095	0.020	0.209	0.672	0.882	0.926	0.940
			10%	0.033	0.006	0.144	0.488	0.765	0.842	0.867
2%	10.15	3.28	1%	0.307	0.129	0.279	0.500	0.872	0.970	0.982
			5%	0.101	0.004	0.072	0.207	0.627	0.870	0.909
			10%	0.028	0.001	0.027	0.094	0.426	0.753	0.818
3%	8.52	2.97	1%	0.355	0.109	0.200	0.339	0.647	0.953	0.968
			5%	0.110	0.012	0.033	0.093	0.320	0.786	0.885
			10%	0.038	0.002	0.019	0.033	0.152	0.644	0.747
5%	7.92	2.68	1%	0.331	0.087	0.130	0.234	0.416	0.869	0.958
			5%	0.113	0.008	0.007	0.038	0.133	0.576	0.785
			10%	0.034	0.004	0.001	0.010	0.046	0.362	0.631
10%	6.08	2.16	1%	0.332	0.102	0.122	0.186	0.253	0.517	0.828
			5%	0.097	0.008	0.010	0.024	0.058	0.218	0.502
			10%	0.035	0.001	0.004	0.008	0.026	0.109	0.333
15%	4.96	1.82	1%	0.313	0.110	0.102	0.158	0.217	0.413	0.656
			5%	0.107	0.015	0.014	0.024	0.043	0.147	0.343
			10%	0.038	0.006	0.005	0.004	0.007	0.066	0.189

Table IB.8: **Simulated Error Rates for Fama and French (2010): Subsamples**

Simulated Type I and Type II error rates across subsamples the Fama and French (2010) approach. We split the full sample into six subsamples (1984–88, 1989–93, 1994–98, 1999–03, 2004–08, and 2009–16). For each subsample and for each p_0 of the fraction of strategies that are believed to be true, we follow our method in Section 2.3 and perturb the adjusted data $M = 1,000$ times to calculate the empirical Type I and Type II error rates. We report the average (across subsamples) Type I and Type II error rates. We consider six percentiles as well as the maximum. "Avg." and "Avg. t-stat" report the average (both across simulations and subsamples) of the mean alpha and alpha t-statistic for the fraction p_0 of funds that are assumed to be outperforming. All funds with an initial AUM exceeding $5 million are included, resulting in 3,030 funds in total for the full sample. Panel A reports the Type I error rate, that is, the probability of falsely declaring that some managers have skill when no fund is outperforming. Panel B reports the Type II error rate, that is, the probability of falsely declaring that no manager is skilled when some funds are outperforming.

				Test Statistics						
p_0 (frac. of true)	Avg. alpha	Avg. t-stat of alpha	α (sig. level)	Max	99.9%	99.5%	99%	98%	95%	90%
Panel A: Type I Error Rate										
0	0	0	1%	0.004	0.003	0.003	0.005	0.004	0.006	0.006
			5%	0.023	0.024	0.027	0.031	0.032	0.032	0.037
			10%	0.057	0.057	0.065	0.068	0.075	0.078	0.081
Panel B: Type II Error Rate										
0.5%	10.47	4.32	1%	0.680	0.650	0.942	0.990	0.992	0.993	0.993
			5%	0.384	0.378	0.781	0.945	0.954	0.960	0.959
			10%	0.248	0.241	0.637	0.862	0.893	0.906	0.907
1%	9.78	3.93	1%	0.618	0.569	0.720	0.950	0.988	0.991	0.991
			5%	0.303	0.263	0.435	0.811	0.932	0.950	0.955
			10%	0.170	0.155	0.289	0.662	0.854	0.891	0.899
2%	8.94	3.50	1%	0.564	0.479	0.543	0.701	0.939	0.986	0.987
			5%	0.243	0.202	0.253	0.406	0.769	0.921	0.941
			10%	0.129	0.107	0.141	0.257	0.597	0.838	0.875
3%	8.42	3.24	1%	0.549	0.460	0.492	0.595	0.790	0.979	0.985
			5%	0.233	0.177	0.204	0.288	0.517	0.886	0.925
			10%	0.119	0.090	0.109	0.169	0.336	0.783	0.845
5%	7.61	2.88	1%	0.534	0.445	0.432	0.509	0.627	0.924	0.975
			5%	0.213	0.155	0.157	0.210	0.332	0.725	0.885
			10%	0.104	0.082	0.073	0.106	0.191	0.547	0.777
10%	6.37	2.38	1%	0.508	0.414	0.382	0.431	0.515	0.691	0.906
			5%	0.184	0.134	0.122	0.150	0.204	0.372	0.673
			10%	0.079	0.055	0.048	0.066	0.103	0.221	0.488
15%	5.55	2.05	1%	0.510	0.430	0.375	0.418	0.483	0.606	0.765
			5%	0.204	0.143	0.115	0.135	0.178	0.305	0.469
			10%	0.089	0.062	0.045	0.056	0.086	0.164	0.297
20%	4.90	1.80	1%	0.516	0.415	0.364	0.399	0.462	0.568	0.687
			5%	0.195	0.142	0.116	0.139	0.178	0.266	0.388
			10%	0.091	0.063	0.047	0.062	0.085	0.148	0.243

Table IB.9: **Simulated Error Rates for Fama and French (2010): 1984–1988**

Simulated Type I and Type II error rates across subsamples the Fama and French (2010) approach. We split the full sample into six subsamples (1984–88, 1989–93, 1994–98, 1999–03, 2004–08, and 2009–16). For each subsample and for each p_0 of the fraction of strategies that are believed to be true, we follow our method in Section 2.3 and perturb the adjusted data $M = 1,000$ times to calculate the empirical Type I and Type II error rates. We report the average (across subsamples) Type I and Type II error rates. We consider six percentiles as well as the maximum. "Avg." and "Avg. t-stat" report the average (both across simulations and subsamples) of the mean alpha and alpha t-statistic for the fraction p_0 of funds that are assumed to be outperforming. All funds with an initial AUM exceeding $5 million are included, resulting in 3,030 funds in total for the full sample. Panel A reports the Type I error rate, that is, the probability of falsely declaring that some managers have skill when no fund is outperforming. Panel B reports the Type II error rate, that is, the probability of falsely declaring that no manager is skilled when some funds are outperforming.

				Test Statistics						
p_0 (frac. of true)	Avg. alpha	Avg. t-stat of alpha	α (sig. level)	Max	99.9%	99.5%	99%	98%	95%	90%
Panel A: Type I Error Rate										
0	0	0	1%	0.009	0.007	0.005	0.010	0.008	0.010	0.010
			5%	0.035	0.034	0.035	0.043	0.043	0.044	0.046
			10%	0.068	0.068	0.070	0.079	0.091	0.082	0.094
Panel B: Type II Error Rate										
0.5%	8.29	5.13	1%	0.650	0.650	0.932	0.981	0.983	0.988	0.989
			5%	0.407	0.407	0.761	0.942	0.941	0.953	0.948
			10%	0.274	0.274	0.604	0.876	0.876	0.896	0.891
1%	8.40	4.82	1%	0.523	0.522	0.602	0.972	0.976	0.984	0.988
			5%	0.268	0.267	0.343	0.831	0.920	0.945	0.943
			10%	0.157	0.156	0.238	0.702	0.840	0.875	0.880
2%	8.28	4.15	1%	0.375	0.374	0.362	0.486	0.831	0.978	0.975
			5%	0.137	0.138	0.120	0.214	0.581	0.889	0.927
			10%	0.066	0.067	0.061	0.112	0.389	0.792	0.846
3%	8.08	3.87	1%	0.353	0.352	0.311	0.412	0.621	0.966	0.970
			5%	0.109	0.108	0.088	0.119	0.327	0.854	0.913
			10%	0.046	0.046	0.047	0.066	0.172	0.742	0.824
5%	7.65	3.46	1%	0.362	0.361	0.289	0.305	0.440	0.853	0.960
			5%	0.102	0.102	0.091	0.107	0.156	0.599	0.855
			10%	0.058	0.058	0.037	0.042	0.082	0.410	0.737
10%	6.85	2.92	1%	0.285	0.284	0.204	0.214	0.261	0.431	0.816
			5%	0.054	0.053	0.043	0.058	0.063	0.145	0.506
			10%	0.018	0.018	0.005	0.017	0.024	0.065	0.306
15%	6.06	2.56	1%	0.310	0.310	0.220	0.235	0.242	0.343	0.538
			5%	0.063	0.061	0.035	0.035	0.042	0.083	0.202
			10%	0.022	0.023	0.012	0.013	0.015	0.034	0.081
20%	5.51	2.27	1%	0.284	0.284	0.168	0.173	0.196	0.263	0.376
			5%	0.054	0.053	0.034	0.031	0.035	0.052	0.101
			10%	0.027	0.027	0.009	0.013	0.011	0.015	0.037

Table IB.10: **Simulated Error Rates for Fama and French (2010): 1989–1993**

Simulated Type I and Type II error rates across subsamples the Fama and French (2010) approach. We split the full sample into six subsamples (1984–88, 1989–93, 1994–98, 1999–03, 2004–08, and 2009–16). For each subsample and for each p_0 of the fraction of strategies that are believed to be true, we follow our method in Section 2.3 and perturb the adjusted data $M = 1,000$ times to calculate the empirical Type I and Type II error rates. We report the average (across subsamples) Type I and Type II error rates. We consider six percentiles as well as the maximum. "Avg." and "Avg. t-stat" report the average (both across simulations and subsamples) of the mean alpha and alpha t-statistic for the fraction p_0 of funds that are assumed to be outperforming. All funds with an initial AUM exceeding $5 million are included, resulting in 3,030 funds in total for the full sample. Panel A reports the Type I error rate, that is, the probability of falsely declaring that some managers have skill when no fund is outperforming. Panel B reports the Type II error rate, that is, the probability of falsely declaring that no manager is skilled when some funds are outperforming.

				Test Statistics						
p_0 (frac. of true)	Avg. alpha	Avg. t-stat of alpha	α (sig. level)	Max	99.9%	99.5%	99%	98%	95%	90%
Panel A: Type I Error Rate										
0	0	0	1%	0.008	0.007	0.003	0.007	0.004	0.008	0.005
			5%	0.031	0.033	0.027	0.040	0.045	0.041	0.044
			10%	0.064	0.064	0.069	0.066	0.073	0.075	0.079
Panel B: Type II Error Rate										
0.5%	12.27	4.60	1%	0.658	0.658	0.848	0.985	0.990	0.990	0.994
			5%	0.398	0.399	0.647	0.932	0.941	0.947	0.952
			10%	0.253	0.252	0.489	0.855	0.916	0.915	0.907
1%	11.24	4.16	1%	0.559	0.558	0.593	0.817	0.974	0.985	0.987
			5%	0.260	0.260	0.336	0.625	0.909	0.933	0.945
			10%	0.155	0.156	0.226	0.486	0.837	0.895	0.896
2%	10.54	3.90	1%	0.519	0.518	0.459	0.581	0.909	0.977	0.989
			5%	0.218	0.219	0.214	0.299	0.691	0.919	0.937
			10%	0.113	0.112	0.137	0.187	0.527	0.853	0.879
3%	10.02	3.61	1%	0.437	0.438	0.426	0.475	0.632	0.965	0.982
			5%	0.176	0.175	0.141	0.217	0.378	0.864	0.921
			10%	0.076	0.075	0.055	0.097	0.210	0.766	0.839
5%	8.97	3.24	1%	0.435	0.433	0.320	0.380	0.464	0.871	0.959
			5%	0.139	0.139	0.112	0.134	0.209	0.628	0.874
			10%	0.070	0.071	0.039	0.051	0.098	0.452	0.766
10%	7.72	2.75	1%	0.430	0.430	0.309	0.315	0.355	0.531	0.862
			5%	0.125	0.124	0.082	0.087	0.092	0.224	0.566
			10%	0.039	0.039	0.029	0.024	0.037	0.109	0.382
15%	6.89	2.44	1%	0.438	0.439	0.311	0.324	0.360	0.439	0.660
			5%	0.135	0.137	0.082	0.075	0.087	0.168	0.327
			10%	0.042	0.041	0.015	0.025	0.032	0.064	0.165
20%	6.16	2.17	1%	0.432	0.432	0.264	0.257	0.298	0.380	0.527
			5%	0.116	0.116	0.064	0.059	0.062	0.109	0.201
			10%	0.034	0.034	0.016	0.026	0.026	0.040	0.091

Table IB.11: **Simulated Error Rates for Fama and French (2010): 1994–1998**

Simulated Type I and Type II error rates across subsamples the Fama and French (2010) approach. We split the full sample into six subsamples (1984–88, 1989–93, 1994–98, 1999–03, 2004–08, and 2009–16). For each subsample and for each p_0 of the fraction of strategies that are believed to be true, we follow our method in Section 2.3 and perturb the adjusted data $M = 1,000$ times to calculate the empirical Type I and Type II error rates. We report the average (across subsamples) Type I and Type II error rates. We consider six percentiles as well as the maximum. "Avg." and "Avg. t-stat" report the average (both across simulations and subsamples) of the mean alpha and alpha t-statistic for the fraction p_0 of funds that are assumed to be outperforming. All funds with an initial AUM exceeding \$5 million are included, resulting in 3,030 funds in total for the full sample. Panel A reports the Type I error rate, that is, the probability of falsely declaring that some managers have skill when no fund is outperforming. Panel B reports the Type II error rate, that is, the probability of falsely declaring that no manager is skilled when some funds are outperforming.

				Test Statistics						
p_0 (frac. of true)	Avg. alpha	Avg. t-stat of alpha	α (sig. level)	Max	99.9%	99.5%	99%	98%	95%	90%
Panel A: Type I Error Rate										
0	0	0	1%	0.000	0.001	0.002	0.002	0.003	0.003	0.002
			5%	0.023	0.018	0.019	0.017	0.016	0.009	0.017
			10%	0.053	0.051	0.062	0.053	0.059	0.070	0.059
Panel B: Type II Error Rate										
0.5%	10.54	3.73	1%	0.231	0.238	0.015	0.003	0.002	0.001	0.002
			5%	0.372	0.372	0.107	0.142	0.127	0.025	0.022
			10%	0.533	0.541	0.222	0.130	0.091	0.079	0.069
1%	9.26	3.23	1%	0.859	0.846	0.903	0.993	0.997	0.998	0.996
			5%	0.586	0.543	0.700	0.891	0.956	0.968	0.969
			10%	0.393	0.372	0.534	0.751	0.889	0.910	0.924
2%	7.81	2.83	1%	0.812	0.774	0.792	0.890	0.987	0.998	0.994
			5%	0.528	0.495	0.523	0.683	0.865	0.942	0.960
			10%	0.327	0.304	0.353	0.507	0.741	0.855	0.899
3%	7.20	2.59	1%	0.819	0.765	0.773	0.846	0.947	0.991	0.994
			5%	0.492	0.446	0.460	0.546	0.737	0.921	0.953
			10%	0.320	0.280	0.292	0.406	0.575	0.809	0.869
5%	6.20	2.22	1%	0.801	0.745	0.739	0.813	0.873	0.979	0.990
			5%	0.473	0.422	0.385	0.476	0.628	0.838	0.921
			10%	0.290	0.257	0.224	0.298	0.427	0.699	0.804
10%	4.68	1.70	1%	0.776	0.717	0.693	0.740	0.798	0.928	0.971
			5%	0.449	0.385	0.328	0.375	0.476	0.652	0.811
			10%	0.234	0.190	0.158	0.206	0.282	0.460	0.630
15%	3.79	1.38	1%	0.792	0.763	0.675	0.736	0.797	0.898	0.949
			5%	0.485	0.426	0.302	0.344	0.408	0.619	0.735
			10%	0.270	0.218	0.159	0.179	0.241	0.278	0.536
20%	3.04	1.13	1%	0.798	0.746	0.697	0.718	0.774	0.864	0.927
			5%	0.449	0.396	0.318	0.367	0.447	0.556	0.670
			10%	0.259	0.221	0.157	0.201	0.250	0.363	0.480

Table IB.12: **Simulated Error Rates for Fama and French (2010): 1999–2003**

Simulated Type I and Type II error rates across subsamples the Fama and French (2010) approach. We split the full sample into six subsamples (1984–88, 1989–93, 1994–98, 1999–03, 2004–08, and 2009–16). For each subsample and for each p_0 of the fraction of strategies that are believed to be true, we follow our method in Section 2.3 and perturb the adjusted data $M = 1,000$ times to calculate the empirical Type I and Type II error rates. We report the average (across subsamples) Type I and Type II error rates. We consider six percentiles as well as the maximum. "Avg." and "Avg. t-stat" report the average (both across simulations and subsamples) of the mean alpha and alpha t-statistic for the fraction p_0 of funds that are assumed to be outperforming. All funds with an initial AUM exceeding $5 million are included, resulting in 3,030 funds in total for the full sample. Panel A reports the Type I error rate, that is, the probability of falsely declaring that some managers have skill when no fund is outperforming. Panel B reports the Type II error rate, that is, the probability of falsely declaring that no manager is skilled when some funds are outperforming.

				Test Statistics						
p_0 (frac. of true)	Avg. alpha	Avg. t-stat of alpha	α (sig. level)	Max	99.9%	99.5%	99%	98%	95%	90%
Panel A: Type I Error Rate										
0	0	0	1%	0.003	0.000	0.000	0.000	0.002	0.002	0.003
			5%	0.027	0.017	0.025	0.027	0.023	0.031	0.030
			10%	0.067	0.069	0.067	0.079	0.077	0.090	0.091
Panel B: Type II Error Rate										
0.5%	17.21	4.29	1%	0.585	0.540	0.949	0.996	0.997	0.998	0.997
			5%	0.216	0.249	0.729	0.935	0.955	0.960	0.966
			10%	0.116	0.127	0.573	0.833	0.874	0.902	0.893
1%	16.21	3.98	1%	0.535	0.463	0.716	0.970	0.994	0.996	0.997
			5%	0.168	0.140	0.341	0.810	0.936	0.957	0.963
			10%	0.057	0.066	0.193	0.635	0.826	0.878	0.889
2%	14.69	3.62	1%	0.503	0.372	0.475	0.699	0.966	0.993	0.994
			5%	0.101	0.075	0.159	0.343	0.812	0.928	0.944
			10%	0.043	0.026	0.055	0.174	0.606	0.826	0.865
3%	13.38	3.34	1%	0.506	0.362	0.407	0.546	0.790	0.988	0.995
			5%	0.130	0.067	0.113	0.184	0.465	0.887	0.916
			10%	0.031	0.019	0.043	0.086	0.272	0.772	0.832
5%	12.11	3.00	1%	0.444	0.320	0.353	0.421	0.560	0.936	0.985
			5%	0.101	0.068	0.070	0.129	0.252	0.683	0.880
			10%	0.032	0.016	0.029	0.046	0.132	0.494	0.772
10%	10.19	2.54	1%	0.461	0.304	0.281	0.333	0.439	0.653	0.899
			5%	0.094	0.057	0.054	0.072	0.113	0.273	0.642
			10%	0.017	0.010	0.018	0.030	0.052	0.138	0.458
15%	9.04	2.25	1%	0.432	0.308	0.239	0.279	0.359	0.511	0.727
			5%	0.093	0.042	0.045	0.058	0.087	0.208	0.390
			10%	0.022	0.014	0.009	0.010	0.034	0.080	0.239
20%	8.23	2.04	1%	0.444	0.311	0.274	0.319	0.376	0.501	0.654
			5%	0.104	0.051	0.052	0.053	0.072	0.157	0.303
			10%	0.033	0.017	0.011	0.011	0.018	0.054	0.149

Table IB.13: **Simulated Error Rates for Fama and French (2010): 2004–2008**

Simulated Type I and Type II error rates across subsamples the Fama and French (2010) approach. We split the full sample into six subsamples (1984–88, 1989–93, 1994–98, 1999–03, 2004–08, and 2009–16). For each subsample and for each p_0 of the fraction of strategies that are believed to be true, we follow our method in Section 2.3 and perturb the adjusted data $M = 1,000$ times to calculate the empirical Type I and Type II error rates. We report the average (across subsamples) Type I and Type II error rates. We consider six percentiles as well as the maximum. "Avg." and "Avg. t-stat" report the average (both across simulations and subsamples) of the mean alpha and alpha t-statistic for the fraction p_0 of funds that are assumed to be outperforming. All funds with an initial AUM exceeding \$5 million are included, resulting in 3,030 funds in total for the full sample. Panel A reports the Type I error rate, that is, the probability of falsely declaring that some managers have skill when no fund is outperforming. Panel B reports the Type II error rate, that is, the probability of falsely declaring that no manager is skilled when some funds are outperforming.

				Test Statistics						
p_0	Avg.	Avg. t-stat	α	Max	99.9%	99.5%	99%	98%	95%	90%
(frac. of true)	alpha	of alpha	(sig. level)							
Panel A: Type I Error Rate										
0	0	0	1%	0.003	0.000	0.009	0.008	0.008	0.010	0.008
			5%	0.008	0.020	0.034	0.036	0.037	0.051	0.057
			10%	0.041	0.041	0.067	0.077	0.082	0.091	0.091
Panel B: Type II Error Rate										
0.5%	8.03	4.34	1%	0.743	0.646	0.956	0.989	0.993	0.987	0.988
			5%	0.403	0.314	0.795	0.949	0.948	0.946	0.936
			10%	0.240	0.190	0.653	0.861	0.893	0.898	0.899
1%	7.81	3.94	1%	0.702	0.583	0.779	0.971	0.992	0.988	0.984
			5%	0.326	0.199	0.456	0.850	0.926	0.931	0.939
			10%	0.157	0.091	0.261	0.709	0.862	0.887	0.889
2%	7.28	3.54	1%	0.687	0.489	0.616	0.789	0.963	0.980	0.981
			5%	0.307	0.180	0.270	0.456	0.825	0.910	0.921
			10%	0.163	0.089	0.135	0.286	0.677	0.846	0.881
3%	7.21	3.29	1%	0.671	0.480	0.533	0.642	0.858	0.976	0.981
			5%	0.301	0.166	0.233	0.350	0.556	0.885	0.914
			10%	0.159	0.077	0.120	0.177	0.380	0.803	0.848
5%	6.65	2.92	1%	0.675	0.482	0.502	0.579	0.688	0.944	0.968
			5%	0.297	0.122	0.157	0.222	0.383	0.780	0.890
			10%	0.112	0.057	0.058	0.103	0.199	0.604	0.789
10%	5.44	2.41	1%	0.627	0.423	0.424	0.486	0.601	0.739	0.930
			5%	0.244	0.122	0.133	0.162	0.240	0.426	0.741
			10%	0.119	0.053	0.044	0.066	0.112	0.245	0.576
15%	4.77	2.09	1%	0.623	0.452	0.446	0.487	0.576	0.676	0.819
			5%	0.279	0.119	0.117	0.150	0.204	0.357	0.553
			10%	0.116	0.044	0.044	0.060	0.089	0.191	0.366
20%	4.14	1.84	1%	0.636	0.395	0.406	0.456	0.542	0.630	0.756
			5%	0.290	0.154	0.121	0.158	0.216	0.335	0.483
			10%	0.137	0.050	0.049	0.057	0.105	0.182	0.330

Table IB.14: **Simulated Error Rates for Fama and French (2010): 2009–2016**

Simulated Type I and Type II error rates across subsamples the Fama and French (2010) approach. We split the full sample into six subsamples (1984–88, 1989–93, 1994–98, 1999–03, 2004–08, and 2009–16). For each subsample and for each p_0 of the fraction of strategies that are believed to be true, we follow our method in Section 2.3 and perturb the adjusted data $M = 1,000$ times to calculate the empirical Type I and Type II error rates. We report the average (across subsamples) Type I and Type II error rates. We consider six percentiles as well as the maximum. "Avg." and "Avg. t-stat" report the average (both across simulations and subsamples) of the mean alpha and alpha t-statistic for the fraction p_0 of funds that are assumed to be outperforming. All funds with an initial AUM exceeding $5 million are included, resulting in 3,030 funds in total for the full sample. Panel A reports the Type I error rate, that is, the probability of falsely declaring that some managers have skill when no fund is outperforming. Panel B reports the Type II error rate, that is, the probability of falsely declaring that no manager is skilled when some funds are outperforming.

				Test Statistics						
p_0 (frac. of true)	Avg. alpha	Avg. t-stat of alpha	α (sig. level)	Max	99.9%	99.5%	99%	98%	95%	90%
Panel A: Type I Error Rate										
0	0	0	1%	0.001	0.000	0.000	0.000	0.000	0.007	0.007
			5%	0.011	0.024	0.024	0.025	0.029	0.017	0.024
			10%	0.053	0.049	0.054	0.056	0.066	0.063	0.070
Panel B: Type II Error Rate										
0.5%	10.47	4.32	1%	0.574	0.545	0.985	0.994	0.994	0.994	0.995
			5%	0.253	0.262	0.865	0.955	0.962	0.979	0.977
			10%	0.136	0.140	0.726	0.875	0.903	0.917	0.922
1%	9.78	3.93	1%	0.534	0.443	0.730	0.979	0.994	0.993	0.993
			5%	0.207	0.169	0.433	0.857	0.942	0.966	0.972
			10%	0.099	0.086	0.284	0.687	0.872	0.899	0.912
2%	8.94	3.50	1%	0.492	0.348	0.551	0.760	0.974	0.991	0.990
			5%	0.164	0.106	0.233	0.439	0.841	0.937	0.958
			10%	0.067	0.049	0.103	0.280	0.639	0.856	0.883
3%	8.42	3.24	1%	0.511	0.361	0.503	0.650	0.893	0.987	0.991
			5%	0.190	0.096	0.191	0.312	0.639	0.901	0.936
			10%	0.082	0.044	0.098	0.183	0.407	0.810	0.859
5%	7.61	2.88	1%	0.483	0.327	0.387	0.556	0.738	0.961	0.989
			5%	0.165	0.076	0.127	0.193	0.369	0.821	0.889
			10%	0.065	0.031	0.051	0.094	0.208	0.625	0.796
10%	6.37	2.38	1%	0.474	0.323	0.379	0.494	0.635	0.867	0.959
			5%	0.138	0.060	0.095	0.143	0.240	0.510	0.770
			10%	0.049	0.021	0.036	0.052	0.109	0.309	0.577
15%	5.55	2.05	1%	0.466	0.309	0.360	0.446	0.564	0.768	0.902
			5%	0.167	0.075	0.107	0.150	0.238	0.399	0.608
			10%	0.061	0.029	0.029	0.048	0.108	0.235	0.397
20%	4.90	1.80	1%	0.505	0.322	0.377	0.471	0.585	0.768	0.884
			5%	0.159	0.080	0.104	0.161	0.238	0.384	0.565
			10%	0.060	0.030	0.043	0.060	0.101	0.236	0.371

C FAQ

- *Why not use the empirical correlation instead of bootstrap?*

  There are a large number of pair-wise correlations to estimate, which raises concern over estimation uncertainty and how this uncertainty would impact the calibration of Type I and Type II error rates. In addition, correlation only captures the second cross-moment between return strategies. Other moments may have an impact on the results.

- *Suppose you simply focus on the top 5% of funds based on t-statistics. Alternatively, suppose we perturb the ranking in t-statistics by bootstrapping, say, one million times. Won't the bootstrap simply identify the top 5% that we start with?*

  Suppose there are 100 strategies and we rank them and label them accordingly. For the original data, the top 5% are, say, strategy 100, 99, 98, 97, 96, and 95. In bootstrapped samples, suppose the chance for us to have the same strategies as the top 5% is 30%. So it is true that 30% of the time in our simulations we will have the same top 5% of funds. But for the remaining 70% of the time, the identities for the top 5% will be different and will likely impact the Type I and Type II error rate calculation. Notice that for a large cross-section of funds (say 10,000), it is almost impossible to have the same top 5% in the bootstrapped samples as in the original sample. What particular 5% we identify as true will have an impact on the error rate calculations. Compare the scenario where the signal-to-noise ratio is low so even the weakest strategy will have a non-negligible chance to be included in the top 5% to the scenario where the ratio is high so it is almost always the case that the same 5% are included. Our perturbed ranking approach takes the difference between these two scenarios into account.

- *For a fund with 13 observations over 400 time periods, isn't it more likely to draw zero or just one observation in the bootstrapped simulations?*

  Suppose we do independent bootstrapping. The probability of drawing zero observations is 0.0000018 ($= (\frac{400-13}{400})^{400}$), which is actually quite low. The reason is that although the probability of getting a missing observation is high for each draw (i.e., a probability of $0.97 = \frac{400-13}{400}$), it becomes much less likely to draw all missing observations among 400 trials. In fact, as we show in the paper, the number of bootstrapped observations is roughly symmetric around 13, which is the expected number of draws.

- *How are extreme t-statistics generated in the bootstrapped simulations for a fund with a small sample?*

  Suppose the fund has $T = 13$ observations. In the bootstrapped simulations, there is a non-zero probability for us to draw five distinct observations for this fund. If this were the case, since we use the Fama-French-Carhart four-factor model as the benchmark (so there are five independent variables, including the intercept), the linear regression model would imply a perfect fit, which in turn generates a t-statistic that is undefined (i.e., infinity). This example, although unrealistic since T is required to be no less than eight for the bootstrapped simulations, illustrates the idea of how extreme t-statistics can be generated in the bootstrapped simulations for funds with a small sample.