How Many Good and Bad Funds Are there, Really?

Wayne Ferson and Yong Chen*

This version: April 18, 2018

Abstract

Building on the work of Barras, Scaillet and Wermers (BSW, 2010), we propose a modified approach to inferring performance for a cross-section of investment funds. Our model assumes that funds belong to groups of different abnormal performance or alpha. Using the structure of the probability model, we simultaneously estimate the alpha locations and the fractions of funds for each group, taking multiple testing into account. Our approach allows for tests with imperfect power that may falsely classify good funds as bad, and vice versa. Examining both mutual funds and hedge funds, we find smaller fractions of zero-alpha funds and more funds with abnormal performance, compared with the BSW approach. We also use the model as prior information about the cross-section of funds to evaluate and predict fund performance.

* Ferson is the Ivadelle and Theodore Johnson Chair of Banking and Finance and a Research Associate of the National Bureau of Economic Research, Marshall School of Business, University of Southern California, 3670 Trousdale Parkway Suite 308, Los Angeles, CA. 90089-0804, ph: (213) 740-5615, ferson@marshall.usc.edu, www-rcf.usc.edu/~ferson/. Chen is Associate Professor of Finance at Mays Business School, Texas A&M University, ph: (979) 845-3870, ychen@mays.tamu.edu. We are grateful to Laurent Barras, Stéphane Chrétien, Sanjiv Das, Chris Hrdlicka, Richard Roll, Luke Taylor, Kenneth Singleton, Haibei Zhao, and to participants at the 2013 Financial Research Association (FRA) early ideas session, the 2014 FRA Conference, the 2014 Morgan Stanley Quantitative Equity Research Conference, the 2015 Southwestern Finance meetings, the 2015 Northern Finance Association meetings, the 2015 China International Conference in Finance, the 2015 Society of Financial Studies Cavalcade, and workshops at Arizona State University, Baylor University, California Institute of Technology, Louisiana State University, Rice University, Santa Clara University, Texas A&M University, the Wharton School, University of Melbourne, University of Miami, University of Southern California, and University of Washington for feedback.

1. Introduction

A big problem for studies that empirically examine a cross-section of investment funds is separating true performance from luck. This is inherently a problem of classifying or grouping the funds. We contribute to the literature by further developing an approach for grouping fund alphas, motivated by Barras, Scaillet and Wermers (BSW, 2010). BSW evaluate fund performance accounting for multiple comparisons in the cross-section of mutual funds. To illustrate the multiple comparisons problem, consider a large cross-section where 4,000 funds are evaluated and a 10% test size is used. We may expect 400 funds to record abnormal performance even if all of the funds have zero alphas and no future performance is expected. BSW describe a model in which the population of funds consists of three subpopulations. A fraction of funds, $\pi_{0,}$ have zero alphas, while a fraction π_g of "good" funds have positive alphas and a fraction π_b are "bad" funds with negative alphas.

BSW estimate the fractions for mutual funds by adjusting for false discovery rates (FDR). The FDR is the fraction of "discoveries," or rejections of the null hypothesis of zero alpha when it is correct (Storey, 2002). Their estimate of π_g is the fraction of funds where the null hypothesis that alpha is zero is rejected in favor of a good fund, minus the expected FDR among the zero-alpha funds. We refer to this approach in our paper as the "classical" FDR method. BSW's main estimates for mutual funds are π_0 = 75% and π_g = 1%, using data up to 2006. A number of subsequent studies have applied their approach to mutual funds and hedge funds.[1]

In this paper, by modifying the classical FDR method, we propose an approach to estimate the fractions of funds in different alpha groups based on the structure of the probability model. Our model assumes that a fund's performance is drawn from a mixture of distributions, where there are fixed expected alpha values (to be estimated) and certain fractions of the funds belong to each group. Unlike the classical FDR method that does not use the locations of the fund groups, we use the location information as part of the model.[2] Using simulations, we estimate the alpha locations of the fund groups and the fractions of the funds in each group simultaneously, while accounting for multiple comparisons.

[1] Cuthbertson et al. (2012) apply the BSW approach to UK mutual funds. Criton and Scaillet (2014) and Ardia and Boudt (2018) apply the method to hedge funds and Dewaele et al. (2011) apply it to funds of hedge funds. Romano et al. (2008) also present a small hedge fund example. Bajgrowicz and Scaillet (2012) apply false discovery methods to a large sample of technical trading rules.

[2] BSW do compute the locations of the nonzero alpha funds and explore the locations through the tails of the t-ratio distributions, but they do not simultaneously estimate the locations and the fractions of funds in each group. In an example, they estimate the alpha location of the good funds by finding the noncentrality parameter of a noncentral t-distribution that matches the expected fractions of funds rejected (see their Internet Appendix).

Our approach accounts for imperfect test power and the possibility of confusion where tests with imperfect power can confuse a truly bad fund (alpha<0) for a good fund (alpha>0) or vice-versa. Hence, our estimates of the fractions of funds in each group adjust for both false discovery rates (i.e., type I error) and low test power (i.e., type II error), and we show how the benefits of this adjustment depends on the power of the tests as well as the fractions of funds. Our approach reduces to the BSW estimator when the power of the tests is 100%. However, when the test power is low (as in the tests for hedge funds), our approach can deliver more accurate inferences.

In addition, we show that the structure of our model helps to better separate skill from luck in the cross-section of fund performance. We estimate the probability that a given fund has a positive alpha, using information from the entire cross-section of funds. For example, if the model says that almost no funds have positive alphas, the chances that a particular fund with a positive alpha estimate was just lucky, are much higher than if the model says many funds have positive alphas.[3] In simulations with known fractions of zero and positive-alpha funds, our method detects the positive-alpha funds more reliably than the classical FDR method. In this sense, our method has improved "power" to detect good funds.

We apply our approach to both active US equity mutual funds during January 1984–December 2011 and hedge funds during January 1994–March 2012.[4] For mutual funds, during our sample period the classical estimator suggests that about 72% of the mutual funds have zero alphas. Our model implies that about 51% of the mutual funds have zero alphas, about 7% are bad (alphas of -0.03% per month) and the remaining are "ugly" (alphas of -0.20% per month).

In our sample of hedge funds, the classical estimate of π_0 is 76%. In contrast, we estimate that very few of the hedge funds' alphas are zero. The best models imply that more than 50% of the hedge funds are good, with alphas centered around 0.25% per month, and most of the others are bad, with alphas centered around -0.11% per month.

[3] Formally, we compute the posterior probability of a positive alpha using Bayes rule, where the cross-section of funds, as characterized through the probability model, represents the prior. Previous studies that use Bayesian methods for fund performance measurement and fund selection include Brown (1979), Baks, Metrick, and Wachter (2001), Pástor and Stambaugh (2002), Jones and Shanken (2005), Avramov and Wermers (2006), among others. Our application is different from the earlier studies because our prior reflects multiple skill distributions for subpopulations of different skill levels.

[4] Following BSW, we measure net-of-fee fund alpha based on benchmark factor returns so that we can compare the inferences between the two approaches. Our approach can be applied to other fund performance measures, such as holding based measures (Daniel, Grinblatt, Titman, and Wermers, 1997), stochastic discount factor alpha (Farnsworth et al., 2002), measure of value added (Berk and van Binsbergen, 2015), and gross alpha (Pástor, Stambaugh, and Taylor, 2015).

The basic probability model assumes that funds are drawn from one of three distributions centered at different alphas. We find that models with three subpopulations fit the cross-section of funds' alpha t-ratios better than models in which there is only a single group, or in which there are two groups. In principle, the approach can be extended for many different groups. However, in the three-alpha model joint estimation reveals that there are linear combinations of the nonzero alpha values that produce a similar fit for the data, indicating that a three-group model is likely all that is needed.

We estimate the models on 60-month rolling windows to examine trends in the parameters over time, and we use rolling formation periods to assess the information in the model about future fund returns. In our hedge fund sample, the difference between the Fung and Hsieh (2004) alphas of the good and bad fund groups in the first year after portfolio formation is about 7% per year, with a t-ratio of 2.3. The rolling window parameters show worsening performance over time for the good mutual funds and hedge funds, while the alphas of the bad funds are relatively stable over time. Much of the potential investment value for mutual funds comes from avoiding bad funds, whereas for hedge funds there is value in finding good funds.

While the classical FDR method is essentially non-parametric with no assumption about the distribution of fund performance, our approach uses the structure of the probability model. It is worth noting the tradeoff between these two approaches. The classical method is flexible about the performance distribution, but it may have low power and underestimate π_g and π_b. Our parametric approach improves the power when the probability model is reasonable, but its accuracy depends on the probability model being correctly specified. In Section 2, we explain the economic reasoning behind the probability model used in our approach.

The approach here also generalizes studies such as Kowsowski, Timmerman, White, and Wermers (2006) and Fama and French (2010), who bootstrap the cross-section of mutual fund alphas. In those studies, all of the inferences are conducted under the null hypothesis of zero alphas, so there is only one group of funds. The analysis is directed at the hypothesis that all funds have zero alphas, accounting for the multiple hypothesis tests. The current approach also accounts for multiple hypothesis tests, but allows that some of the funds have nonzero alphas.

Recently, Chen, Cliff and Zhao (2017) consider a parametric model in which there are groups of hedge funds, with each group's alpha drawn form a normal distribution with a different mean and standard deviation. The unconditional distribution in their model is a mixture of normals. They use a modified EM algorithm to estimate the locations of the groups and the fractions of hedge funds in each group. Subsequent work by Harvey and Liu (2018) develops a full-blown maximum likelihood approach to estimate the parameters. Chen, Cliff and Zhao (2017) use a larger sample of hedge funds that include more strategies than our sample, and they find a larger number of groups best fits their data—four groups of hedge funds while we find three groups. Unlike our approach, these two studies do not explicitly examine the effect of imperfect test power and confusion on the inferences.

This paper is organized as follows. Section 2 describes the model and its estimation. Section 3 describes the data. Section 4 describes our simulation methods and presents a series of simulation experiments to evaluate the models. Section 5 presents the empirical results. Section 6 discusses robustness checks and Section 7 concludes. The Appendix describes our approach to the standard errors. An Internet Appendix provides technical details and ancillary results.

2. The Model

The model assumes that the mutual funds or the hedge funds are members of one of three subpopulations. This is appealing, as grouping is common when evaluating choices on the basis of quality that is hard to measure. For example, Morningstar rates mutual funds into "star" groups. Security analysts issue buy, sell and hold recommendations. Academic journals are routinely categorized into groups, as are firms' and nations' credit worthiness, restaurants, hotels, etc.

For investment funds, a structure of three groups associated with zero, negative and positive alphas seems natural. That some funds should have zero alphas is predicted by Berk and Green (2004). Under decreasing returns to scale, new money should flow to positive-ability managers until the performance left for competitive investors is zero. There are also many reasons to think that some funds may have either positive or negative alphas. Frictions (e.g., taxes, imperfect information, agency costs, and cognitive errors) may keep investors from quickly pulling their money out of bad funds. It is also natural to hope for funds with positive alphas, and these funds could be available because costs slow investors' actions to bid them away. A number of empirical studies find evidence that subsets of funds may have positive alphas. Jones and Mo (2016) identify more than 20 fund characteristics that previous studies associate with alphas.

The stylized structure of the probability model is illustrated in Figure 1 for hypothetical densities. There are three subpopulations of funds whose alphas are centered around the values [α_b<0, 0, α_g >0]. A fraction of the funds belongs to each subpopulation. The fractions, which sum to 1.0, are [π_b, π_0, π_g]. The unconditional distribution of funds' alphas is a mixture of the three distributions.

2.1. Estimation by Simulation

The set of unknown parameters is $\theta = [\pi_g, \pi_b, \alpha_g, \alpha_b, \beta_g, \beta_b, \delta_g, \delta_b]$, where the last four are the power and confusion parameters of the tests. The parameters are estimated in three stages as follows. In the first stage, for given values of $[\alpha_b, \alpha_g]$ we use three simulations to estimate the parameters $[\beta_b, \beta_g, \delta_b, \delta_g]$. We set the size of the tests, $(\gamma/2)$, say to 10% in each tail.

The first simulation of the cross-section of fund alphas, imposing the null hypothesis that all of the alphas are zero, produces two critical values for the t-statistics, t_g and t_b. The critical value t_g is the t-ratio above which 10% of the simulated t-statistics lie when the null of zero alphas is true. The critical value t_b is the value below which 10% of the simulated t-statistics lie under the null hypothesis of zero alphas.

The second simulation imposes the alternative hypothesis that funds are good; that is, the alphas are centered at the value $\alpha_g>0$. The fraction of the simulated t-ratios above t_g is the power of the test for good funds, β_g. The fraction of the simulated t-ratios below t_b is an empirical estimate of the probability of rejecting the null in favor of finding a bad fund when the fund is actually good. This is the confusion parameter, δ_b.[5]

The third simulation adopts the alternative hypothesis that funds are bad; that is, the alphas in the simulation are centered at the value $\alpha_b<0$. The fraction of the simulated t-ratios below t_b is the power of the test to find bad funds, β_b. The fraction of the simulated t-ratios above t_g is the confusion parameter, δ_g.

Our approach says that a good or bad fund has a single value of alpha. However, it should be robust to a model where the good and bad alphas are random and centered around the values (α_g, α_b).[6]

The second stage of our procedure combines the simulation estimates with results from the cross-section of funds in the actual data as follows. Let F_b and F_g be the fractions of rejections of the null hypothesis in the actual data using the simulation-generated critical values, t_b and t_g. We model:

[5] Formally, δ_b is one corner of the 3 x 3 probabilistic confusion matrix characterizing the tests. See Das (2013, p.148) for a discussion.

[6] Suppose, for example, that in the simulation we drew for each good fund a random true alpha, α_{pTRUE}, equal to α_g plus mean zero independent noise. In order to match the sample mean and variance of the fund's return in the simulation to that in the data, we would reduce the variance of $\{r_{pt} - \alpha_{pTRUE}\}$, by the amount of the variance of the noise in α_{pTRUE} around α_g. The results for the cross-section should come out essentially the same.

$$E(F_g) = P(\text{reject at } t_g|H_0)\,\pi_0 + P(\text{reject at } t_g|\text{Bad})\,\pi_b + P(\text{reject at } t_g|\text{Good})\,\pi_g. \quad (1)$$

$$= (\gamma/2)\,\pi_0 + \delta_g\,\pi_b + \beta_g\,\pi_g, \text{ and similarly:}$$

$$E(F_b) = P(\text{reject at } t_b|H_0)\,\pi_0 + P(\text{reject at } t_b|\text{Bad})\,\pi_b + P(\text{reject at } t_b|\text{Good})\,\pi_g. \quad (2)$$

$$= (\gamma/2)\,\pi_0 + \beta_b\,\pi_b + \delta_b\,\pi_g.$$

Equations (1) and (2) present two equations in the two unknowns, π_b and π_g, (since $\pi_0 = 1 - \pi_b - \pi_g$) which we can solve for given values of [β_g, β_b, δ_g, δ_b, F_g, F_b]. The solution to this problem is found numerically, by minimizing the squared errors of equations (1) and (2) subject to the Kuhn-Tucker conditions for the constraints that $\pi_b \geq 0$, $\pi_g \geq 0$ and $\pi_b + \pi_g \leq 1$. We estimate $E(F_g)$ and $E(F_b)$ by the fractions rejected in the actual data and we calibrate the parameters [β_g, β_b, δ_g, δ_b] from the simulations.

We assume that with enough simulation trials, we can nail down the β and δ parameters with zero error. The impact of this assumption is addressed in the Internet Appendix, where we conclude that variation in these parameters across the simulation trials has a trivial impact on the results.

The estimates of the π's that result from the second stage are conditioned on the values of [α_g, α_b]. The third stage of our approach is to search for the best-fitting values of the alphas. The search proceeds as follows. Each choice for the alpha values generates estimates $\pi(\alpha)$ for the fractions. At these parameter values, the model implies a mixture of distributions for the cross-section of fund returns. We simulate data from the implied mixture of distributions, and we search over the choice of alpha values and the resulting $\pi(\alpha)$ estimates (repeating the first two stages at each point in the search grid) until the simulated mixture distribution generates a cross-section of estimated fund alpha t-ratios that best matches the cross-sectional distribution of alpha t-ratios estimated in the actual data. We determine the best match using the familiar Pearson χ^2 statistic as the criterion:

$$\text{Pearson } \chi^2 = \Sigma_i\,(O_i - M_i)^2/O_i, \quad (3)$$

where the sum is over K cells, O_i is the frequency of t-statistics for alpha that appear in cell *i* in the original data, and M_i is the frequency of t-statistics that appear in cell *i* using the model, where the null hypothesis is that the model frequencies match those of the original data.

The Pearson statistic requires choosing the cell sizes. We choose K=100 cells, with the cell boundaries set so that an approximately equal number of t-ratios in the original data appear in each cell (i.e., $O_i \approx N/100$). The Pearson statistic may also be affected by the fact that the alpha t-ratios are estimates, so that the estimation error creates an errors-in-variables problem. In a robustness section we address these issues by examining other goodness-of-fit measures, and find that our results are robust to alternative measures.

In summary, the fractions of actual funds discovered to be good or bad at the simulation-generated critical values determine the fractions of funds in each group according to the equations (1) and (2), given the locations of the groups. This is a modification of the classical FDR framework, accounting for the power and confusion parameters. Unlike the FDR estimator that is nonparametric, our approach is based on parametric simulations. We use three simulations and one overall goodness-of-fit measure to identify the four main parameters of our model, $[\pi_g, \pi_b, \alpha_g, \alpha_b]$, taking the power and confusion parameters of the tests as given.

2.2. Using the Model

We use the information from the cross-section of funds and the estimated parameters of the probability model to draw inferences about individual funds. Given a fund's alpha estimate α_p, we use Bayes rule to compute:

$$P(\alpha > 0 \mid \alpha_p) = f(\alpha_p \mid \alpha>0)\, \pi_g / f(\alpha_p), \tag{4}$$

$$f(\alpha_p) = f(\alpha_p \mid \alpha=0)\, \pi_0 + f(\alpha_p \mid \alpha>0)\, \pi_g + f(\alpha_p \mid \alpha<0)\, \pi_b. \tag{5}$$

The densities $f(\alpha_p|\alpha=0)$, $f(\alpha_p|\alpha>0)$ and $f(\alpha_p|\alpha<0)$ are the empirical conditional densities of alpha estimates from the three subpopulations of funds. These are estimated by simulation and fit using a standard kernel density estimator for f(.) as described in the Internet Appendix. We actually use the t-ratios of the alphas instead of the alpha estimates in our simulations, because the t-ratio is a pivotal statistic.

The inference for a given fund reflects the estimated fractions of funds in each subpopulation. For example, as the prior probability, π_g, that there are good funds approaches zero, the posterior probability that the particular fund has a positive alpha, given its point estimate α_p, approaches zero. This captures the idea that if there are not many good funds, a particular fund with a positive alpha estimate is likely to have been lucky. The inference also reflects the locations of the groups through the likelihood that the fund's alpha estimate could be drawn from the subpopulation of good funds. If the likelihood $f(\alpha_p|\alpha>0)$ is small, the fund is less likely to have a positive alpha than if $f(\alpha_p|\alpha>0)$ is large. This uses information about the position of the fund's alpha estimate relative to the other funds in a category.

2.3. Relation to the Classical FDR Method

The classical false discovery rate estimator makes the assumption that the fraction of funds not rejected in the data is the fraction of zero alpha funds, multiplied by the probability that the test will not reject when the null is true. The estimator $\pi_{g,C}$ adjusts the fraction of funds found to be good by the expected fraction of false discoveries among the zero-alpha funds. The idea is that if the size of the test is large enough, as suggested by Storey (2002) and BSW, then all the good funds will have alpha t-statistics larger than the small critical value that results.

The Internet Appendix shows that our estimators of the π fractions, derived by solving equations (1) and (2), reduce to the classical estimators from Storey (2002) and BSW when $\beta_g = \beta_b = 1$ and $\delta_g = \delta_b = 0$. The classical FDR estimators of the fractions are:

$$\pi_{0,C} = [1- (F_b + F_g)]/(1- \gamma); \;\; \pi_{g,C} = F_g - (\gamma /2)\, \pi_{0,C}. \tag{6}$$

The assumption that $\beta_g = \beta_b = 1$ says that the tests have 100% power, an assumption that would bias the estimates toward finding too many zero-alpha funds in the presence of imperfect test power. Storey (2002) motivates $\beta = 1$ as a "conservative" choice, justified by choosing the size of the tests to be large enough.[7] This bias, however, can be important when the power of the tests is low. Low-power tests have long been seen as a problem in the fund performance literature. The potential bias of the classical estimator can be measured by combining Equations (1), (2) and (6) as follows.

[7] BSW, following Storey (2002), search using simulations for the size of the test, γ, that solves $\text{Min}_\gamma\, E\{[\pi_0(\gamma) - \text{Min}_\gamma\, \pi_0(\gamma)]^2\}$, where $\pi_0(\gamma)$ is the estimate that results from Equation (6) when γ determines the fractions rejected. This step reduces the estimate of π_0 and its bias. For this reason, we refer to estimates based on Equation (6) but without this additional estimation step, as the "classical" FDR estimator. BSW find that using fixed values of $\gamma/2$ near 0.25-0.30 without the additional minimization produces similar results to estimates that minimize the mean squared errors. We use these sizes in the classical estimators below, without the mean square error minimization. In our paper, the value γ is also used in Equation (6) as the threshold (denoted as λ in Storey (2002) and BSW) to estimate $\pi_{0,C}$, as BSW (2010, p. 189) suggest a similarly large value (such as 0.5 or 0.6) for the threshold.

$$
\begin{aligned}
E(\pi_{0,C}) &= [1 - E(F_b) - E(F_g)]/(1- \gamma) \qquad (7)\\
&= \{1 - [(\gamma/2)\ \pi_0 + \delta_g\, \pi_b + \beta_g\, \pi_g] - [(\gamma/2)\ \pi_0 + \beta_b\, \pi_b + \delta_b\, \pi_g]\}/(1- \gamma)\\
&= [\pi_0\, (1- \gamma) + \pi_g\, (1- \beta_g - \delta_b) + \pi_b\, (1- \beta_b - \delta_g)]/(1- \gamma)\\
&= \pi_0 + [\pi_g\, (1- \beta_g - \delta_b) + \pi_b\, (1- \beta_b - \delta_g)]/(1- \gamma).
\end{aligned}
$$

The second term in the last line of Equation (7) captures the bias under the structure of the probability model. If we assume $\beta_g = \beta_b = \beta$ and $\delta_g = \delta_b = \delta$ for ease of illustration, then the bias becomes $(1- \pi_0)(1- \beta - \delta)/(1- \gamma)$. The bias is likely to be small if the true fraction of zero-alpha funds is large or the power of the tests is close to 100%. Intuitively, the accuracy of the classical estimator depends on large π_0 or large β. However, in simulations and empirical analysis to actual funds, we find that the bias can be remarkable when the fraction of zero-alpha funds is small and the test power is low, as in the case of hedge funds. Our approach delivers more accurate inferences accounting for imperfect test power.

More specifically, our analysis in the Internet Appendix reveals two offsetting biases in the classical estimator. Assuming perfect power (i.e., setting $\beta_g = \beta_b =1$) biases π_0 upwards, while assuming that the tests will never confuse a good and bad fund (i.e., setting $\delta_g = \delta_b = 0$) biases π_0 downwards. The result of our simulations shows that the upward bias from assuming perfect power dominates and the classical estimator finds too many zero alpha funds.

The $\pi_{0,C}$ estimator does not rely on the location of the good and bad funds, $[\alpha_g, \alpha_b]$, in the sense that it only considers the null distributions. Thus, it is likely to be robust to the structure of the alpha distributions. For example, if the true distribution is multimodal but we assume too-small a number of groups, our results can be biased due to model misspecification. The choice between the two approaches trades off the robustness and smaller sampling variability of the classical method with the smaller bias and greater power of our approach using more of the structure of the probability model.

Our estimators use the structure of the probability model, including β's and δ's, which are functions of the alpha locations under the alternative hypotheses. Like BSW, our estimator of π_g adjusts the fraction of funds discovered to be good for false discoveries among zero-alpha funds. Moreover, allowing for imperfect test power, we control for cases where the tests are confused, and "very lucky" funds with negative alphas are found to have significant positive performance.[8]

[8] The Internet Appendix derives the false discovery rates from our model, relates them to previous studies and applies the methods in a trading strategy.

In summary, there is a tradeoff between the two approaches in inferring the fractions of funds in each alpha group. The classical FDR method is a good choice if the fraction of zero alpha funds is known to be large and the power of the tests is high, or if the distribution of fund performance is a complex mixtures of distributions with different moments.[9] On the other hand, our estimator is ideal if the structure of the probability model (e.g., a three-alpha distribution) is justified by economic theory, in which case our parametric approach improves power and provides more accurate inferences.

3. The Data

We study mutual fund returns measured after expense ratios and funds' trading costs for January 1984–December 2011 from the Center for Research in Security Prices Mutual Fund database, focusing on active US equity funds. We subject the mutual fund sample to a number of screens to mitigate omission bias (Elton, Gruber, and Blake, 2001) and incubation bias (Evans, 2010). We exclude observations prior to the reported year of fund organization, and we exclude funds that do not report a year of organization or which have initial total net assets (TNA) below $10 million or less than 80% of their holdings in stock in their otherwise first eligible year to enter our data set. Funds that subsequently fall below $10 million in assets under management are allowed to remain, in order to avoid a look-ahead bias. We combine multiple share classes for a fund, focusing on the TNA-weighted aggregate share class.[10] These screens leave us with a sample of 3,619 mutual funds with at least 8 months of returns data.

Our hedge fund data are from Lipper TASS. We study only funds that report monthly net-of-fee US dollar returns, starting in January 1994. We focus on US equity oriented funds, including only those categorized as either dedicated short bias, event driven, equity market neutral, fund-of-funds or long/short equity hedge. We require that a fund have more than $10 million in assets under management as of the first date the fund would otherwise be eligible to be included in our analysis. To mitigate backfill bias, we remove the first 24 months of returns and returns before the dates when funds were first entered into the database, and funds with missing values in the field for the add date. Our

[9] As another useful feature of the classical method, the FDR theory provides asymptotic results that can be used for statistical inference (e.g., Genovese and Wasserman, 2002, 2004).

[10] We identify and remove index funds both by Lipper objective codes (SP, SPSP) and by searching the funds' names with key word "index." Our funds include those with Policy code (1962-1990) CS, Wiesenberger OBJ (1962-1993) codes G, G-I, G-I-S, G-S, G-S-I, I, IFL, I-S, I-G, I-G-S, I-S-G, S-G, S-G-I, S-I-G, GCI, IEQ, LTG, MCG, SCG, Strategic Insight OBJ code (1993-1998) AGG, GMC, GRI, GRO, ING, SCG, Lipper OBJ/Class code (1998-present), CA, EI, G, GI, MC, MR, SG, EIEI, ELCC, LCCE, LCGE, LCVE, LSE, MCCE, MCGE, MCVE, MLCE, MLGE, MLVE, SESE, SCCE, SCGE, SCVE and S.

sample includes 3,620 hedge funds over January 1994–March 2012.

Table 1 presents summary statistics of the mutual fund and hedge fund data. We require at least 8 monthly observations of return to include a fund in this table. The mean hedge fund return (0.37% per month) is smaller than the average mutual fund return (0.62%), but the longer sample for the mutual funds includes the high-return 1984–1993 period. The range of average returns across funds is much greater in the hedge fund sample, especially in the negative-return, left tail. A larger fraction of the hedge funds lose money for their investors, and the losses have been larger than in the mutual funds.

The two right hand columns of Panel A of Table 1 summarize the Fama-French (1996) three-factor alphas and their heteroskedasticity-consistent t-ratios for the mutual funds. For the hedge funds in Panel B, we use the Fung and Hsieh (2001, 2004) seven factors. The Internet Appendix shows the results for hedge funds when the Fama and French factors are used.[11] We use standard alphas, similar to BSW, in all of our analyses.

The median alpha for the hedge funds is positive, while for the mutual funds it is slightly negative. The tails of the cross-sectional alpha distributions extend to larger values for the hedge funds. For example, the upper 5% tail value for the t-ratio of the alphas in the hedge fund sample is 4.18 (the alpha is 1.25% per month), while for the mutual funds it is only 1.47 (the alpha is 0.27%). In the left tails the two types of funds also present different alpha distributions, with a thicker lower tail for the alphas and t-ratios in the hedge fund sample. One of our goals is to see how these impressions of performance hold up when we consider multiple hypothesis testing, and use bootstrapped samples to capture the correlations and departures from normality that are present in the data.

Table 1 shows that the sample volatility of the median hedge fund return (2.62% per month) is smaller than for the median mutual fund (5.34%). The range of volatilities across the hedge funds is greater, with more mass in the lower tail. For example, between the 10% and 90% fractiles of hedge funds the volatility range is 1.2%–6.7%, while for the mutual funds it is 4.2%–7.0%. Getmansky, Lo and Makarov (2004) study the effect of return smoothing on the standard deviations of hedge fund returns and show that smoothed returns reduce the standard deviations and induce positive

[11] The hedge fund alphas are slightly smaller on average and the cross-sectional distribution of the alphas shows thinner tails when the Fama and French factors are used. The Fung and Hsieh seven factors include the excess stock market return and a "small minus big" stock return similar to the Fama and French factors, except constructed using the S&P 500, and the difference between the S&P500 and the Russell 2000 index. In addition, they include three "trend-following" factors constructed from index option returns; one each for bonds, currencies and commodities. Finally, there are two yield changes; one for ten-year US Treasury bonds and one for the spread between Baa and ten-year Treasury yields.

autocorrelation in the returns. The autocorrelations of the returns are slightly higher for the hedge funds, consistent with more return smoothing in the hedge funds. The median autocorrelation for the hedge funds 0.16, compared with 0.12 for the mutual funds, and some of the hedge funds have substantially higher autocorrelations. The 10% right tail for the autocorrelations is 0.50 for the hedge funds, versus only 0.24 for the mutual funds. Asness, Krail, and Liew (2001) show that return smoothing can lead to upwardly biased estimates of hedge fund alphas. We consider the effect of return smoothing in the robustness section and in the Internet Appendix, and conclude that smoothing is not likely to be material for our results.

4. Simulation Exercises

This section describes the simulation method in more detail and presents simulation exercises with two main goals. The first goal is to evaluate small sample biases in the estimators and their standard errors and to inform our choices for the values of some of the parameters, such as the size of the tests ($\gamma/2$), for our empirical analyses. The second goal is to evaluate the "power," through the discovery rates, of our approach compared with classical methods.

4.1. Simulation Details

The first stage of our procedure follows Fama and French (2010), simulating under the null hypothesis that alpha is zero. We draw randomly with replacement from the rows of $\{r_{pt} - \alpha_p, f_t\}_t$, where α_p is the vector of funds' alpha estimates in the actual data, r_{pt} is the funds' excess returns vector and f_t is a vector of the factor excess returns. This allows for correlation between the residuals of the factor models used to estimate alphas. (Alternative approaches are considered in the robustness section.) All returns are in excess of the one month Treasury bill return. This imposes the null that the "true" alphas are zero in the simulation. When we simulate under the assumption that the true alphas are not zero for a given fraction of the population, π, we select $N\pi$ funds at random, where N is the total number of funds in the sample, and add the relevant value of alpha to their returns, net of the estimated alpha. (An alternative approach is considered in the robustness section.) Each trial of the three simulations delivers an estimate of the β and δ parameters. We use and report the average of these across 1,000 simulation trials.

Following Fama and French (2010), we use an 8-month survival screen for mutual funds (and

12 months for hedge funds). We impose the selection criterion only after a fund is drawn for an artificial sample. This raises the issue of a potential inconsistency in the bootstrap, as the missing values will be distributed randomly through “time” in the artificial sample, while they tend to occur in blocks in the original data. We consider an alternative approach to address this issue in a robustness section.

While we describe the results in terms of the alpha values, all of the simulations are conducted using the t-ratios for the alphas as the test statistic, where the standard errors are the White (1980) heteroskedasticity consistent standard errors. We use the t-ratio because it is a pivotal statistic, which should improve the properties of the bootstrap compared with simulating the alphas themselves. An overview of the bootstrap is provided by Efron and Tibshirani (1993).

4.2. Finite Sample Properties

To evaluate the finite sample properties of the estimators, we conduct simulations of the simulation method. In each of 1,000 draws, artificial data are generated from a mixture of three fund distributions, determined by given values of the π fractions and alphas. A given draw is generated by resampling months at random from the hedge fund data, where the relevant fractions of the funds have the associated alphas added to each of their returns, after the sample estimates of their alphas have been subtracted. We use the hedge fund data because, as Table 1 suggests, the departures from normality are likely to be greater for hedge funds than for mutual funds, providing a tougher test of the finite sample performance.

For each draw of artificial data from the mixture with known parameters, we run the estimation by simulation for a given value of the size of the tests, $\gamma/2$. The π fractions are estimated in each of these trials from the three simulations as described above, each using 1,000 artificial samples generated by resampling from the one draw from the mixture distribution, treating that draw the same way we treat the original sample when conducting the estimation by simulation. The α-parameters are held fixed for these experiments.

Table 2 presents the results of simulating the simulations. The π fractions are set to $\pi_0 = 0.10$, $\pi_g = 0.60$ and $\pi_b = 0.30$, and the bad and good alphas are set to -0.138% and 0.262% per month. We report experiments in the Internet Appendix where we set all the π fractions to 1/3, and where we set the π fractions to the banner values reported by BSW: $\pi_0 = 0.75$, $\pi_g = 0.01$ and $\pi_b = 0.24$. These experiments generate broadly similar results.

The Avg. estimates in Table 2 are the averages over the 1,000 draws from the mixture distribution. These capture what we expect to find when applying a given estimator in the simulated economy. The *empirical standard deviations* are the standard deviations of the parameter estimates, taken across the 1,000 simulation draws. This is the variability in the estimators that the reported standard errors should capture. The Root MSE's are the square roots of the averages over the 1,000 draws, of the squared difference between an estimated and true parameter value.[12] The four panels of the table use different choices for ($\gamma/2$).

Table 2 shows that, under the mixture distribution, the classical estimator of π_0 can be severely biased in favor of finding too many zero-alpha funds, and the estimators of the fractions of good and bad funds biased toward zero. When 10% of the funds have zero alphas, the classical estimates are 75-96%, depending on the size of the tests. The bias is smaller at the larger test sizes, as suggested by BSW. Like the classical estimator our approach finds too many zero alpha funds and too few good and bad funds. But our point estimates are much less biased than the classical estimators.

Our point estimates are typically within one empirical standard deviation of the true values of the π fractions at the 5% test size. Our estimator is more accurate at the 10% size and slightly more accurate still at the 20% size, where the expected point estimate is within 0.3-0.6 standard errors of the true parameter value.

The results of Table 2 are remarkably different from Figure 3 in BSW, and from reported simulation and empirical results in the literature for the Storey (2002) estimator. There are several reasons for the differences. While BSW use mutual fund data, we conduct our simulations using hedge fund data that exhibit greater dispersion in alpha, as shown in Table 1. We do not employ the additional minimization step described in footnote 7. This difference will make the bias in the BSW approach smaller for smaller test sizes, but as BSW report the results are not changed much by this step for the larger sized tests reported in Table 2 Panels C and D. Another difference is that BSW use standard normal p-values to estimate the fractions rejected, F_g and F_b, whereas we use the empirical critical values from the bootstrapped null distribution. Using standard normal p-values the tests will be improperly sized. We compare the critical values in our simulations with the standard normal values and find that the empirical critical values are larger. Using empirical critical values in our analysis as opposed to the normal ones, we get smaller F_b and F_g, and thus larger estimate of π_0 in our calculations.

[12] In the event that a parameter estimate is on the boundary of the parameter space (a π fraction is zero or 1.0), we drop the estimated standard error for that simulation trial for the calculations. This choice has only a small effect on the results.

When the test sizes $\gamma/2$ are 0.3 and 0.4, however, these differences are small.

Panel D of Table 2 takes the size of the tests to 30% in each tail. This is the approximate test size that BSW advocate. The classical point estimates are improved but remain substantially biased. This means that, for the hedge fund sample, even when we use a large size of the tests in the classical estimator $\pi_{0,C}$, there are still too many good and bad funds whose alpha t-ratios fall between t_b and t_g in the simulations, a situation that BSW assume to be unlikely for mutual funds. Thus, the simulation results imply that the test power for hedge funds is well below 100% even when the test size is set to a large value.[13]

Table 2 also evaluates the standard errors of the estimators. The reported classical standard errors understate the sampling variability of the estimates for all test sizes. Even when $\gamma/2 = 0.20$ or 0.30 (Panels C and D), where they perform the best, the average classical standard errors range from 20% to 60% of the empirical standard errors.[14] Our standard error estimates are also biased. When the test size is 5% they are far too small for π_0 and too large for π_g and π_b. When the size is 10% the standard errors are reasonably accurate for π_0 but still too large for π_g and π_b by 50-100%.[15]

The empirical standard errors for our estimators in Table 2 get smaller as the size of the tests is increased from 5% to 20%. The average empirical standard error at the 20% size is about 60% of the value at the 5% size. This is the opposite of the pattern in the average reported standard errors, which are larger at the larger test sizes. Given this tradeoff, the 10% test size appears to be the best choice in our method. The reported standard errors for π_0 are fairly accurate, and they are overstated

[13] As a numerical example, suppose the power of the tests $\beta_g = \beta_b = 0.63$ and the confusions $\delta_g = \delta_b = 0.12$, roughly the case with hedge funds when $\gamma/2 = 0.3$ for the chosen alpha values in Panel A of Table 5. Given the true $\pi_0 = 0.1$ in the simulations and based on Equation (7), $E(\pi_{0,C}) = \pi_0 + (1- \pi_0) (1- \beta - \delta)/(1- \gamma) = 0.1 + (0.9\times0.25)/0.4 = 66.3\%$. Here, $\pi_{0,C}$ is severely biased when the fraction of zero alpha funds is small and the test power is low. On the other hand, if the true π_0 is large at say 75% and the test power is as high as 0.98 with no confusion when $\gamma/2 = 0.3$, then the bias of $\pi_{0,C}$ would be $(0.25\times0.02)/0.4 = 1.25\%$, a very small value. The choice of a large γ in the classical estimator is used to deliver high test power, but in the case of hedge funds, the power is still well below 100% even at $\gamma/2 = 0.3$ or 0.4. Our approach allows for imperfect test power. Thus, while the classical FDR method may well suit the mutual fund setting where large γ is associated with high test power as in BSW, our approach has advantages in the setting of hedge funds, in which π_0 is relatively small and large γ does not provide near-perfect power.

[14] As a check, the reported standard error in BSW (Table II on p.197) for π_b is 2.3%. Our simulations of the reported classical standard errors, when $\gamma/2=0.20$, average 2.2%. The simulations in BSW do not account for the variance of the factors, while our simulations do capture the factor variances. According to Fama and French (2010), not accounting for the factor variances understates the sampling variability of the estimates. Consequently, understated variability would inflate the test power in the simulations. In addition, while BSW study mutual funds, our simulations here use hedge funds that exhibit greater dispersion than mutual funds (see Table 1).

[15] We conduct experiments where we set the correlation of the tests across funds to zero, as assumed by BSW, and we find that the standard errors are then an order of magnitude too small.

for π_g and $\pi_{b,}$ and thus conservative.[16] The reported standard errors are close to the RMSEs when the test sizes are 10%.

The simulations show that the classical estimators display lower sampling variability than our estimators. This makes sense, given their relative simplicity. However, the classical estimators concentrate around biased values. For example, when the size is 20%, the classical estimators RMSEs are larger than for our estimators by 150-300%.

When the size of the tests is 30% in Panel D of Table 2, the average point estimates, RMSEs and empirical standard errors of our estimators are similar to those in Panel C, but the average reported standard errors are more overstated. The reported standard errors still average only 30-60% of the empirical standard errors. The RMSEs are slightly improved, but still larger than the RMSEs of our estimators by 160-300%.

Overall, the simulations of the simulation approach lead to several conclusions. First, in finite samples the classical estimator of π_0 overstates the fraction of zero-alpha funds, and understates the fractions of good and bad funds, even at the large test sizes. The classical standard errors are understated in finite samples and the mean squared errors can be quite large. Our approach has a smaller finite sample bias, and performs best overall at the sample sizes used here, when the size of the tests is set to $\gamma/2 = 10\%$. When the size is 10% the point estimates are usually within about one empirical standard error of the true parameter values. Our standard errors are reasonably accurate for π_0 but overstated for π_g and $\pi_{b.}$ They are closer to the RMSEs in these cases.

We conduct some experiments where we expand the number of time-series observations in the simulations to 5,000, in order to see which of the biases are finite sample issues and which are likely inconsistencies. (These experiments are reported in the Internet Appendix.) These large-sample experiments suggest that the classical standard error estimators are consistent. The experiments also suggest that the classical point estimator of π_0 is inconsistent. For example, the average estimated value is about 30% when the true value is 10%, and the expected estimate is 49% when the true value is 1/3. Our estimates are much closer to the true values when T=5,000, suggesting that their biases in Table 2 are finite sample biases.

[16] As previously described, our reported standard errors do not reflect sampling variability in the δ and β parameters, but this variation is captured in the simulations. Formally incorporating this variation in the standard errors would make them larger, but the impact would be small as shown in the Internet Appendix.

4.3. Empirical Power Evaluation

The empirical power of a test is usually measured as the fraction of simulation trials in which the test rejects the null hypothesis in favor of the alternative, when the alternative hypothesis is true. To evaluate "power" in a multiple comparisons setting it is natural to measure the expected *discovery rates*. For example, in a model with zero and positive alphas, the *correct discovery rate* is the fraction of funds that the tests find to have positive alphas, which actually have positive alphas. The *false discovery rate* is the fraction of funds that the method detects as good funds, but which actually have zero alphas. We are interested in both the correct discovery rates and the false discovery rates. The total discovery rate is the sum of the correct and false discovery rates. Of course, we cannot know which funds actually have positive and zero alphas, except in a simulation exercise.

This section evaluates the discovery rates in simulations, comparing our approach with two classical methods. To simplify we use the two-group model where there are only good funds and zero-alpha funds. We vary the true fraction π_g and the good alpha parameter α_g, and keep track of which funds in the simulation actually have good or zero alphas. We record the discovery rates for each method as a function of the parameter values, averaging the number of true and false discoveries across 1,000 simulation trials. Like in the previous section we bootstrap from the hedge fund data, but now we form a mixture of the two known distributions and we run a two-group version of our model on each draw of the mixture, treating each draw as if it was the original data. (The equations for the two-group model are given in the Internet Appendix.)

We compare the discovery rates of three approaches. The first is a naïve application of the simple t-ratio. Setting the size of the one-tailed test to 10%, an empirical critical value of the t-ratio is found by simulating each fund separately under the null hypothesis that its alpha is zero. The funds whose alpha t-ratios exceed their critical values are discovered to have positive alphas. The calculation is naïve in that it takes no account of the multiple comparisons.

The second approach is the classical false discovery rate (FDR) method. Here, the critical value of the t-ratio is adjusted to obtain a desired false discovery rate in the cross-section of funds. We search for a size of the test, γ, so that γ $(\pi_{0,C})/F_g = 0.10$, where $\pi_{0,C}$ is the classical estimator of the fraction of zero-alpha funds and F_g is the fraction of funds where the null hypothesis is rejected. We select FDR=10%, following BSW who find (Table V) their best results at this FDR value. Both $\pi_{0,C}$ and F_g depend on the size of the tests. Simulation under the null hypothesis at the optimal test size determines a critical value for the alpha t-ratio, and all funds with alpha t-ratios in excess of this single critical value are discovered to have positive alphas. Note that since there is no confusion parameter in the two-group model, the classical FDR cannot be biased by confusion in this example.

The third approach is our method as described in Equations (4) and (5). For each draw from the mixture of distributions we implement our simulation method, treating that draw like we do the original data. We use a two-group model with a test size of γ = 10%. Using the estimated model parameters, a fund is discovered to have a positive alpha if the posterior probability given its alpha estimate, $\alpha_{p,}$ satisfies $P(\alpha>0|\alpha_p) > P(\alpha=0|\alpha_p)$, so that a positive alpha is more likely than a zero-alpha.

Table 3 summarizes the results of our analysis of the discovery rates. We present results for two choices of the true alpha values. The first, 0.252% per month, is the value from Table 5 below, estimated on the hedge funds when the test size is 10%. The second value, 0.861% is the cutoff for the upper 10% tail of the alphas in the original hedge fund sample, as shown in Table 1.

The rows in Table 3 vary the true fraction of good funds between 0.0 and 100%. Consider the results in the first ten rows. In the first row, when the true fraction of good funds is zero, there are no correct discoveries to be had. The number of false discoveries by both the classical t-ratio and the FDR method are close to the desired 10%. In row ten, when all of the funds are good, there are no false discoveries to be had. As the fractions of good funds is increased across the rows, the FDR method delivers the smallest number of false discoveries among the three methods, but the classical t-ratio is very close behind. The FDR method is based on the biased classical estimator, which finds too many zero-alpha funds, so it tends to over-correct for false discoveries. For example, when the true fraction of good funds exceeds 50%, the FDR method delivers 5% or fewer false discoveries, even though it is calibrated to a 10% false discovery rate. The FDR method also posts the smallest number of correct discoveries among the methods, topping out at only 32.4% when the true fraction is 100%. The correct discovery rates of the classical t-ratio are slightly better. Interpreting the discovery rates as size and power, this says that both the naïve t-ratio and the FDR method using the classical estimator are undersized: when the desired false discovery rate is 10% the actual false discovery rate is lower than 10%.

Our "CF" method turns in the highest rate of correct discoveries. The estimates of the π_g fractions are considerably more accurate, and the total discovery rates are usually much closer to the actual fractions of good funds than with the other methods. Our method excels in correct detection, especially when the fraction of good funds reaches and exceeds 50% (55% is the value that we estimate using the two-group model for hedge funds in Table 5). The correct discovery rates of the CF method top out at 90.7% when the true fraction is 100%, where the other two methods deliver 33% or less.

The cost of the improved power of our CF method is more false discoveries than either the classical t-ratio or the FDR method. When the true fraction of good funds is in the 30-80% range, our method posts false discovery rates in the 14-20% range. It would be possible in future research, to assign utility costs to the different cases of potential misclassification, and our method could then make a different tradeoff between correct and false discoveries. In this example, once the investor has incurred the costs of engaging in active fund selection, the utility cost of falsely discovering a zero alpha fund to be a good fund is likely to be very low, given that the investor's alternative decision is to invest in a zero alpha fund. The improved ability of our method to correctly detect positive-alpha funds is the important result of this example.

The last ten rows of Table 3 represent a world in which the good alpha is the value that defines the top 10% in our hedge fund sample. The classical estimates of π_g are less biased in this example. This illustrates that while the classical estimator does not refer to the locations of the nonzero alpha funds, its performance under the alternative hypothesis does depend on the locations of the nonzero alpha funds. The FDR method correctly discovers more good funds than it did before. It now outperforms the classical t-test in this regard, and with slightly smaller false discovery rates. Our approach again delivers the best correct discovery rates, matching or beating the other methods for all true fractions of good funds above 40%, and the false discovery rates are also improved. In this exercise our method has false discovery rates below 9% for all values of the true fractions of good funds.

In summary, the simulations show that our approach to discriminating between good funds and zero-alpha funds presents an improvement over previous methods, especially when the fraction of good funds is large. By using the full probability structure of the model, we obtain better power to detect funds with nonzero alphas. We examine below the performance of our methods when there are three groups of funds, in rolling estimation on actual data.

5. Empirical Results

5.1. Mutual Funds

Table 4 presents empirical results for the mutual fund data. Our parameter estimates are compared with the classical FDR estimators. The alphas are estimated using the Fama and French

(1996) three-factor model. (We check the sensitivity of these findings to the factor model in a robustness section.) The fractions of managers in the population with zero, good or bad alphas are estimated using our method with 1,000 simulation trials. The standard errors for the π fractions, shown in parentheses, are the *empirical standard errors* from our bootstrap simulations. These are the standard deviations of the parameter values obtained across the 1,000 trials of the simulations. The standard errors account for the dependence of the tests across funds (see the analyses in the Appendix).

In Panel A of Table 4 we first set the good and bad alphas equal to the values specified by BSW and vary the size of the tests. This confirms in the data the importance of using the right test sizes as suggested by BSW and the simulations of the simulations. The fourth and fifth columns show the π fractions using the classical FDR estimators. In our sample period when the test size is 10% the BSW estimators say π_0 = 81.3% and π_g = -1.7%.[17] As noted by BSW, the estimate of π_0 gets smaller as the size of the tests $\gamma/2$ increases. We find 67.9% when $\gamma/2 = 0.40$. Interpolating, we obtain a value close to BSWs banner estimate of 75% when $\gamma/2$ is about 0.25. This reconfirms the appeal of the larger test sizes, as suggested by BSW and by our simulations, for the classical FDR approach.

As BSW and Storey (2002) argue, the power parameters β increase as the size of the tests increases, ranging from about 50% to just over 90% in Table 4. The confusion parameters δ also increase with the size of the tests, but are 3.9% or less. Our approach delivers smaller estimates for the fractions of zero alpha funds than the classical estimators at each test size. At the preferred 10% test size, our point estimate of π_0 is 65.6% (with a standard error of 10.8%), which is fairly close to the FDR estimate of 67.9% when $\gamma/2 = 0.40$, given the specified good and bad alpha values. At the size $\gamma/2 = 0.40$, the combined values of powers and confusions are close to one, at 95.6% ($\beta_g + \delta_b$) and 93.9% ($\beta_b + \delta_g$), which suggests a small bias with the classical estimator (see Equation (7)).

We find no evidence for any good mutual funds in the population in Panel A, as all of the π_g estimates are 0.0. The inference that there are no good funds is consistent with the conclusions of Fama and French (2010), who simulate the cross-section of alphas for mutual funds under the null hypothesis that the alphas are zero, but do not estimate the π fractions.

Our approach simultaneously considers the alpha-locations as well as the fractions of funds. (In Table A.2 of the Internet Appendix, we show how the alpha locations affect the estimates of π fractions.) Panel B of Table 4 presents the results when the alpha parameters are set equal to the

[17] The negative values arise in the BSW calculation because the fraction rejected F_g is smaller than $(\gamma/2)\pi_0$. Our simultaneous approach, using constrained optimization, avoids negative probabilities.

"optimal" values that best fit the cross-section of the t-ratios for alpha in the actual data, as discussed below. We focus on the preferred 10% test size. (In the Internet Appendix Table A.2, we report results for the 5% test size to show that the results are relatively insensitive to this choice.) In the unconstrained domain case, where the search is free to pick positive or negative alpha values, we estimate that 50.7% of the mutual funds have zero alphas, whereas the classical estimates would suggest about 72% (the results of the classical estimator are reported in the Internet Appendix Table A.2). We estimate that the rest of the mutual funds have negative alphas. These results are interpreted more fully in Section 5.3.

5.2. Hedge Funds

Table 5 repeats the analysis in Table 4 for the hedge fund sample. We use the Fung and Hsieh seven-factor model to compute alphas. (The Internet Appendix Table A.3 contains results using the Fama and French three-factor alphas for hedge funds. The results are similar.) Many of the patterns in Table 5 are similar to the results for the mutual funds. For example, in Panel A the BSW estimates of π_0 decrease in the test size ($\gamma/2$), and the power of the tests increases but remains substantially below 100% even at ($\gamma/2$) = 0.40, where the power parameter is 73%. The power parameter is lower than it is for the mutual fund sample, consistent with the greater dispersion in the hedge fund data. The confusion parameters get larger with the size of the tests like in Table 4, topping out here at just over 15%. Our estimates of π_0 in Panel A of Table 5 indicate smaller fractions of zero-alpha hedge funds and thus, more good and bad hedge funds than the classical estimator at any test size. The classical estimator says that 76% of the funds have zero alphas at the preferred 30% test size. Our estimate at the preferred 10% size is 41%.

The empirical standard errors of the π fractions, shown in Panel A, are larger for the hedge funds than we saw for the mutual funds. They do not increase with the size of the tests like they did for the mutual funds. (The asymptotic standard errors, however, do increase with the test size for the hedge funds; see the Internet Appendix Table A.3.)

Panel B of Table 5 presents the results when the true alpha parameters are set equal to the values that best fit the cross-section of the actual alpha estimates in the data. For these alpha values we estimate that very few hedge funds have zero alphas, whereas the classical estimates suggest 76% (see the Internet Appendix, Table A.2). We estimate that 53.2% of the hedge funds have positive alphas, while the classical estimator suggests about 22%. These estimates are now described and interpreted.

5.3. Joint Estimation

As discussed above, our inferences about the fractions of good and bad funds in the population are sensitive to the assumptions about the alpha locations of the good and bad funds. Our estimates are sensitive because the power of the tests is strongly sensitive to the alpha locations. The confusion parameters also vary with the alpha values, but with a smaller effect.

In this section we search over the choice of the good and bad alpha parameters, and the corresponding estimates of the π fractions, to find those values of the parameters that best fit the distribution of the t-ratios in the actual data, according to the χ^2 statistic in Equation (3). (We consider alternative distance measures in a robustness section.) The best-fitting good and bad alpha parameters minimize the difference between the cross section of fund alpha t-ratios estimated in the actual data, versus the cross section estimated from a mixture of return distributions, formed from the zero, good and bad alpha parameters and the estimated π fractions for each of the three types.

The good and bad alpha parameters, α_g and α_b, are found with a grid search. The search looks from the lower 5% to the upper 95% tail values of the alpha t-ratio estimates in the data, summarized in Table 1, with a grid size of 0.001% for mutual funds and 0.005% for hedge funds. At each point in the grid, the π fractions are estimated using simulation. We start with models in which there are three groups, with zero, good and bad alphas. In the first case, the "unconstrained domain" case, the search does not impose the restriction that the good alpha is positive or the bad alpha is negative. The probability model remains valid without these restrictions, so we let the data speak to what are the best-fitting values.

Figure 2 depicts the results of the grid search for the alpha parameters for hedge funds. The search is able to identify global optima at $\alpha_g = 0.237$, $\alpha_b = -0.098$ when the size of the tests is 5% in each tail. When the size is 10% the values are $\alpha_g = 0.252$, $\alpha_b = -0.108$, as shown in Panel B of Table 5. The Internet Appendix (Table A.4) presents the joint estimation results using different values for the size of the tests, $\gamma/2$.

Figure 2 reveals "valleys" in the criterion surface where linear combinations of the two nonzero alpha values produce a similar fit for the data. The impression is that three groups based on their alphas are plenty to describe the data, and that even fewer groups might suffice. Based on this impression we do not consider models with more than three groups.

The joint estimation for the three-group model applied to mutual funds is summarized in the first row of Panel B of Table 4. We find two negative alphas and some zero alpha funds, but no mutual funds with positive alphas. For the 10% test size we estimate that 50.7% of mutual funds are “good” (which here, means zero alpha), 6.9% are “bad” (meaning, alphas of -0.034% per month) and the remaining 42.4% are “ugly” (meaning, alphas of -0.204% per month). Thus, our estimates of mutual fund performance paint a picture that is similar to but somewhat more pessimistic than the estimates in BSW. Similar to BSW, we find that a large fraction of mutual funds have zero alphas and some funds have strong negative alphas. The bad alpha estimate is similar to the -0.267% per month value suggested by BSW, but our negative “good” alpha estimate is much smaller than the 0.317% good alpha value they suggest.

In the second row of Panel B of Table 4, we summarize the results from repeating the joint estimation for mutual funds, where we constrain the values of the good alpha to be positive and the bad alpha to be negative. The goodness-of-fit measures are larger, indicating a relatively poor fit to the data compared with the unconstrained case.[18] It is interesting that the best-fitting good alphas for the mutual funds are very close to zero: 0.001% per month at the 10% test size. Because the good alphas are so close to zero, the zero-alpha null and the good-alpha alternative distributions are very close to each other. As a result, the power of the tests to find a good alpha and the confusion parameter δ_b are both very close to the size of the tests.

The evidence for mutual funds suggests that the best fitting alphas are either zero or negative, which motivates a simpler model with only two distributions in the population instead of three. The Internet Appendix describes the model when there are two groups. In the hedge fund sample, we let there be one zero and one positive alpha. The results from the two-group models are summarized in the third row of Panel B of Table 4 for mutual funds, and in the second row of Panel B of Table 5 for hedge funds. For mutual funds the nonzero alpha is -0.17% per month and the model says that 52% of the mutual funds have the negative alpha. For hedge funds the nonzero alpha is positive: 0.25% per month, and the estimates say that about 55% of the hedge funds have the positive alpha and 45% have a zero alpha. The larger fractions of zero alphas for hedge funds in the two-group model, compared to the three-group model makes sense, as the two-group model best fits the data by assigning a zero alpha to some of the previously negative-alpha hedge funds. The goodness-of-fit measure, however, shows that the two-group models do not fit the cross-section of funds’ alphas as well as the three-group models.

[18] Asymptotically, the goodness-of-fit statistic is Chi-squared with 99 degrees of freedom and the standard error is about 14. The p-values for all of the statistics and the differences across the models are essentially zero.

Finally, we consider models in which there is only a single value of alpha around which all the funds are centered. The results are summarized in the last row of Panel B of Tables 4 and 5. For mutual funds, the single alpha is estimated to be negative, at -0.21%. For the hedge funds, the single alpha is estimated to be positive, at 0.43%. For mutual funds the goodness-of-fit statistics say that the one-group model fits the data better than the constrained three-group model or the two-group model, but not as well as the unconstrained three-group model. For hedge funds the three-group model provides the best fit.

In summary, the joint estimation results indicate that smaller fractions of funds have zero alphas and larger fractions have nonzero alphas, compared with the evidence using the classical FDR approach. The difference in the results between the two approaches is larger for hedge funds than for mutual funds. As shown in the simulation exercises, the classical FDR approach tends to overestimate π_0 and our approach fares better in correct detection, when the true π_0 is small. Thus, our approach is appealing for inferring performance of funds with disperse performance, like hedge funds.

5.4. Rolling Estimation

We examine the models in 60-month rolling estimation periods. Our goals are two-fold. First, we wish to see how stable the model parameters are over time and to detect any trends. Second, the end of each estimation period serves as a formation period for assigning funds annually into one of the three groups. If there is no information in the model's parameter estimates about future performance, the subsequent performance of the three groups should be the same. If the group of positive-alpha (negative-alpha) funds continues to have abnormal performance, it indicates persistence in performance that may have investment value. The first formation period ends in December of 1998 for hedge funds and in December of 1988 for mutual funds.[19]

Figure 3 summarizes the 60-month rolling estimates of the good and bad alphas for both mutual funds (Panel A) and hedge funds (Panel B). These are jointly estimated like in Panel B of Tables 4 and 5. The bad alphas fluctuate with no obvious trend, but the good alphas show a marked downward trend for mutual funds, and especially for the hedge funds. The smoothness of these graphs, especially for

[19] The cross section includes every fund with at least 8 observations (12 for hedge funds) during the formation period. BSW use five year fund performance records and a 60-month survival screen on the funds. Fama and French (2010) criticize the 60 month survival screen, and we prefer not to impose such a stringent survival screen. If we encounter an alpha estimate larger than 100% per month in any simulation trial, we discard that simulation trial.

the hedge funds, increases our confidence that the alpha estimates are well identified, despite the 60-month window estimation.

For hedge funds the good alpha starts at more than 1% per month and leaves the sample at 0.3%. The ending value is similar to the full sample estimate for the good alpha of 0.25% when the test size is 10%. For mutual funds, both the good and bad alphas are below zero after 2007, consistent with the full sample estimates. It makes sense that the full sample estimates are strongly influenced by the end of the sample period, when there are many more funds in the data. For both kinds of funds the good and the bad alphas get closer together over time.

There are reasons to think that fund performance should be worse and more similar across funds in more recent data. BSW (2010) find evidence of better mutual fund performance in the earlier parts of their sample. Cremers and Petajisto (2009) find a negative trend in funds' active shares over time, and suggest that recent data may be influenced by more "closet indexing" among "active" mutual funds. Kim (2012) finds that the flow-performance relation in mutual funds attenuates after the year 2000, which could be related to the trend toward more similar performance in the cross-section of funds.

Next, funds are assigned each year to one of the three alpha groups on the basis of the model parameters estimated during the formation period. We do this in three different ways. The first way uses the classical false discovery rate method. We find critical t-ratios that control the false discovery rates in the cross-section of funds, accounting for lucky funds with zero alphas that are found to be good funds, and also for the very lucky bad funds that the test confuses with good. The second approach uses a simple group assignment based on the ranked alphas and the estimated proportions of funds in each group. The results of these two approaches are presented in the Internet Appendix. Here we present the results using Bayesian selection to group the funds. Bayesian methods for fund performance evaluation and fund selection have been used in previous studies, such as Brown (1979), Baks, Metrick, and Wachter (2001), Pástor and Stambaugh (2002), Jones and Shanken (2005), and Avramov and Wermers (2006).

The assignment using Bayesian selection follows equations (4) and (5). A fund is assigned to the Good group based on its point estimate of alpha, α_p, if $f(\alpha_p | \alpha>0)\, \pi_g > f(\alpha_p | \alpha<0)\, \pi_b$ and $f(\alpha_p | \alpha>0)\, \pi_g > f(\alpha_p | \alpha=0)\, \pi_0$. A fund is assigned to the Bad group if $f(\alpha_p | \alpha<0)\, \pi_b > f(\alpha_p | \alpha>0)\, \pi_g$ and $f(\alpha_p | \alpha<0)\, \pi_b > f(\alpha_p | \alpha=0)\, \pi_0$. For these calculations the densities f(.|.) are the simulated conditional distributions estimated recursively to the end of the formation period, and evaluated at the rolling alpha estimates using standard kernel density estimation as described in the Internet Appendix.

Equal-weighted portfolios of the selected funds are examined during a holding period. If a fund ceases to exist during the holding period, the portfolio allocates its investment equally among the remaining funds in the group that month. The holding period is a one-year future period; either the first, second, third or fourth year after formation. The 60-month formation period is rolled forward year by year. This gives us a monthly series of holding period returns for each of the first four years after portfolio formation, starting in January 1999 for the hedge funds and in January 1989 for the mutual funds. The holding period returns for the fourth year after portfolio formation start in January 2002 for the hedge funds, and in January 1992 for the mutual funds.

The average returns, their alphas and t-ratios during the holding periods are shown in Table 6. The alphas use the Fama and French factor model in the case of mutual funds, and the Fung and Hsieh factor model in the case of hedge funds. We show results for the selected good funds (Good), the funds in the zero-alpha group (Zero) and the funds in the bad alpha group (Bad). We also show the excess returns of a Good-Bad portfolio (G-B). Of course, the G-B excess return is not obtainable when we cannot short mutual funds or hedge funds. This should be interpreted as the difference between the return obtained by identifying the good funds, compared with choosing bad funds. The differences between the means and alphas of the good and bad groups are not equal to the G-B values, because in many years no bad hedge funds are selected or no good mutual funds are selected, and the G-B series uses only those months where funds exist in both groups.[20]

The second columns of Table 6 show the averages of the numbers of funds in the various portfolios, averaged across the formation years. The smallest group of hedge funds is the bad group. On average, only eight hedge funds are in the bad group, while 378 are in the good group and 245 are in the zero alpha group. For the mutual funds in Panel B, there are many more zero-alpha funds (511) and bad funds (291) than there are good funds (216). Early in the evaluation period there are more good funds, and later in the sample there are more bad funds.

A portfolio of all hedge funds has a positive alpha over the first annual holding period, as does the zero-alpha group, both with t-ratios in excess of three. This reflects the good performance of the hedge funds during our sample period. The bad hedge fund group has a negative alpha, -57 basis points

[20] The estimate of the fraction of bad hedge funds is less than 12% in all of the formation years, and either one or zero hedge funds are selected as bad for the first seven years. There are no good mutual funds selected during six of the last 13 formation years.

per month, and the G-B difference alpha is 59 basis points per month, or about 7% per year, with a t-ratio of 2.3. This compares favorably with the evidence in BSW and our findings using FDR methods to select funds.[21] Thus, there is persistence in the hedge fund performance, detectable by our grouping procedures. In the second and later years after portfolio formation, with one exception the three groups become statistically indistinguishable.

The results for the mutual funds are summarized in Panel B of Table 6. All three groups have negative alphas during the first year, and the bad alpha t-ratio is -2.3. This reflects the poor performance of the mutual funds during our sample period. The G-B difference alpha is 9 basis points per month, with a t-ratio of 1.0. During the second year, the G-B alpha is 14 basis points per month, with a t-ratio of 1.8. With one exception, in the third and fourth years after portfolio formation the three groups become statistically indistinguishable. The evidence for persistence is weaker than it is for the hedge funds.

6. Robustness

This section describes a number of experiments to assess the sensitivity of our results to several issues. These include the pattern of missing values, the level of noise in the simulated fund returns, a possible relation between funds' alphas and active management, alternative factor models and return smoothing.

6.1. The Pattern of Missing Values

There is a potential issue of inconsistency in the bootstrap, as the missing values will be distributed randomly through "time" in the artificial sample, while they tend to occur in blocks in the original data. In fund return data we are much more concerned with cross-sectional dependence and conditional heteroskedasticity, which the simulations do preserve, than we are with serial dependence that can create inconsistency, which is very small in monthly returns. Nevertheless, we conduct an experiment to assess the impact of this issue.

[21] BSW use the FDR method for selecting good mutual funds and find alphas of 1.45% per year or less, with p-values for the alphas of 4% or larger. In the Internet Appendix we select funds using our modification of the FDR application, and find that the G-B alpha difference for mutual funds in our sample is 8 basis points per month, or 0.96% per year during the first year after portfolio formation, with a t-ratio of 1.4. For hedge funds, our G-B alpha is 13 basis points per month, with a t-ratio of 1.8.

We exploit the fact that the beta and alpha estimates when funds are combined are the results of a seemingly-unrelated regression model (SURM), with the same right-hand side variables for each fund. Thus, equation-by-equation OLS produces the same point estimates of the alphas as does the estimation of the full system. We bootstrap artificial data for each fund, *i*, separately, drawing rows at random from the data matrix (f, rf, r_i), which concatenates the factors (f), the risk-free rate (rf) and the returns data for fund *i*, r_i. If we encounter a missing value for a fund, we keep searching randomly over the rows until we find one that is not missing, and we include the value with its associated monthly observation for (f, rf). In this way, we preserve the relation between r_i, the risk-free rate and the vector of factors. When the time-series has been filled out for each fund, we have a simulated sample with no missing values. We then form a "Hole Matrix," H, which is the same size as the original fund sample, with zeros where the original fund data are missing and ones elsewhere. We apply the H matrix to assign missing values in the simulated data for the same months in which they appear in the original data. We estimate the alphas treating this simulated data the same way we treat the original data and the baseline simulation data.

We compare the results of this approach with that of our baseline simulation method in the Internet Appendix and find that the Baseline and Hole-preserving simulations deliver similar statistical properties for funds' residual standard deviations and factor model R-squares. Either method closely reproduces the statistical properties of the original data. The alphas and t-ratios for alpha at various fractiles show that the cross-sectional distributions of alphas and alpha t-ratios that is produced by the two simulation methods are very similar. These results are tabulated in the Internet Appendix.

6.2. The Choice of Goodness-of-fit Criterion

We assess the sensitivity of our joint estimation results to the use of the Pearson Chi-square goodness-of-fit measure. The Pearson measure has the disadvantage that the number of cells must be chosen. Because the alpha t-ratios are estimates, they have estimation error which can affect the measure. In the Internet Appendix we consider two alternative goodness-of-fit measures.

The first alternative measure is the two-sample, Kolmogorov-Smirnov distance:

$$D_{KS} = \mathrm{Sup}_x \mid F_1(x) - F_2(x) \mid, \tag{7}$$

where $F_1(.)$ and $F_2(.)$ are the two empirical cumulative distribution functions (CDFs) of alpha t-ratios;

one from the data and one from the model. This measure looks at the maximum vertical distance between the CDFs. The second alternative measure is the Cramer-von Mises distance,

$$D_{CvM} = E_x \{[F_1(x) - F_2(x)]^2\}, \tag{8}$$

which looks at the mean squared distance.

To implement the alternative measures we combine the observations of the alpha t-ratios from the original data and from a model, rank the values and calculate the two CDFs at each of the discrete points. The results, reported in the Internet Appendix, are similar to those using the original measures.

6.3. Are the Alphas Correlated with Active Management?

Several studies suggest that more active funds have larger alphas (e.g., Cremers and Petajisto, 2009; Titman and Tiu, 2011; Amihud and Goyenko, 2013; Ferson and Mo, 2016). In particular, funds with lower factor model regression R-squares are found to have larger alphas. We examine the correlations between the factor model R-squares and the estimated alphas and find a correlation of -0.015 in the mutual fund sample and -0.111 in the hedge fund sample. The mixtures of distributions simulated above do not accommodate this relation.

We modify the simulations to allow a relation in the cross-section between alpha and active management, measured by the R-squares in the factor model regressions that deliver the alphas. We sort the funds by their factor model R-squares, and group them into three groups with the group sizes determined by the π-fractions at any point in the simulations, and assign the good alpha first to the low R-square group and the bad alpha first to the high R-square group. Thus, this approach builds in a relation between the alpha and active management, measured by the factor model R-squares.

The results are summarized in the Internet Appendix, Table A.4, Panel D. We find that this modification improves the goodness-of-fit statistic. The best-fitting alpha parameters are further out in the tails. In the left tail, the bad alpha moves about ¼ of a percent to the left. The estimates of the good alpha and the fraction of good hedge funds are similar to those in the original design. Fewer fractions of the hedge funds are estimated to have the more pessimistic bad alpha, and the fraction of zero alpha hedge funds increases to 30%-44%, which is a significant positive fraction when the test size is 10%. The BSW estimates are similar to those in the original design. The results suggest that incorporating the relatively poor performance of the high R-square hedge funds improves the fit of the model. This result may not be surprising, but it does suggest that future work on estimation by simulation might profit from building in associations between other fund characteristics and the performance groups.

6.4. Alternative Alphas

While the Fama and French (1996) three-factor model is less controversial for fund performance evaluation than for asset pricing, it is still worth asking if the results are sensitive to the use of different models for alpha. We examine two alternatives for mutual funds: one with fewer factors and one with more factors. The first is the Capital Asset Pricing Model (Sharpe, 1964), with a single market factor and the second is the Carhart (1997) model, which adds a momentum factor. For the hedge fund sample we use the multifactor model of Fung and Hsieh (2001, 2004) in the main tables and try the Fama and French three factor model as a robustness check. Results using the alternative alphas are similar, and some are reported in the Internet Appendix.

6.5. Return Smoothing

Return smoothing tends to reduce the standard errors of fund returns, can increase alphas by lowering the estimated betas, and may be important for hedge funds (e.g., Asness et al., 2001). We address return smoothing by replacing the estimates of the betas and alphas with Scholes-Williams (1977) estimates. Here we include the current and lagged values of the factors in the performance regressions and the beta is the sum of the coefficients on the contemporaneous and lagged factor. The first thing to check is the impact on the t-ratios for alpha in the original data. The results are presented in the Internet Appendix. The effect of using the Scholes-Williams betas on the alpha t-ratio distributions is small, so we do not further investigate the issue.

7. Conclusions

We build on the approach to mutual fund classification of Barras, Scaillet and Wermers (BSW, 2010). We simultaneously estimate the fractions of good, zero-alpha and bad funds in the population along with the alpha locations of the three groups. We modify the False Discovery Rate framework of BSW to allow for imperfect test power and confusion, where a test indicates that a good fund is bad, or vice versa. We show how to use the model as prior information about the cross-section of funds to predict fund performance.

We apply our approach to a sample of US active equity mutual funds and a sample of TASS hedge funds. Large fractions of hedge funds are estimated to have either positive or negative alphas. For mutual funds, a model with only zero and negative alphas best fits the data. Both mutual funds and hedge funds present a trend toward decreasing performance over time in the high-alpha group, while the performance of the low-alpha group shows no trends.

We study the finite sample performance of the estimators through a parametric bootstrap simulation of the simulations. We show both analytically and through the simulations that the classical FDR approach finds too many zero-alpha funds and thus too few good and bad funds when the true fraction of zero-alpha funds is small. Our approach offers improved power in the sense of better detection rates for funds with abnormal performance.

Our simulation-based empirical standard errors indicate that the confidence intervals around the fractions of good and bad funds are wide. In an example using hedge funds, the classical FDR method implies a two-standard error confidence band of (11.0%, 15.8%) for the fraction of zero alpha funds. However, adjusting for finite sample biases in the standard errors, the confidence band is (21.1%, 99.1%). Despite the low precision, we can say with statistical confidence using our estimators, that there are positive and negative alpha hedge funds in our sample. The mutual funds are a different story, where there is no evidence of positive alphas and strong evidence for negative alphas.

Our results motivate future research. For example, one of our robustness checks suggests that more precise inferences might be available by associating fund performance groups with other fund characteristics. Chen, Cliff and Zhao (2017) present some analyses along these lines, and further investigation of this idea seems warranted. Another tack is to find more precise performance measures. We illustrate our approach with the alphas from standard factor models, but our approach can be applied to other fund performance measures, such as holding based measures (e.g., Daniel et al., 1997), stochastic discount factor alpha (Farnsworth et al., 2002; Ferson, Henry, and Kisgen, 2006), measure of value added (Berk and van Binsbergen, 2015), and gross alpha (Pástor, Stambaugh, and Taylor, 2015). Each of these measures has been shown to have their own appeals in measuring fund performance.

Appendix: Standard Errors

Solving equations (1) and (2) for the π fractions we obtain the estimators:

$$\pi_b = B\,(F_g - \gamma/2) + C\,(F_b - \gamma/2) \qquad (A.1)$$

$$\pi_g = D\,(F_g - \gamma/2) + E\,(F_b - \gamma/2),$$

where the constants B, C, D and E depend only on γ, the β's, and the δ coefficients.[22] We assume that by simulating with a large enough number of trials, we can accurately identify the β and the δ parameters as constants. When the power and confusion parameters are equal, the coefficients in (A.1) imply division by zero, and the π fractions are not identified.

Using (A.1) we compute the variances of the π fractions:

$$\mathrm{Var}(\pi_b) = B^2\,\mathrm{Var}(F_g) + C^2\,\mathrm{Var}(F_b) + 2BC\,\mathrm{Cov}(F_g, F_b),$$

$$\mathrm{Var}(\pi_g) = D^2\,\mathrm{Var}(F_g) + E^2\,\mathrm{Var}(F_b) + 2DE\,\mathrm{Cov}(F_g, F_b). \qquad (A.2)$$

The variance of the π_0 estimator is found from $\mathrm{Var}(1-\pi_b -\pi_g) = \mathrm{Var}(\pi_b) + \mathrm{Var}(\pi_g) + 2\mathrm{Cov}(\pi_b, \pi_g)$, where the covariance term is evaluated by plugging in the expressions in (A.1).

The standard errors depend on $\mathrm{Cov}(F_g,F_b)$, $\mathrm{Var}(F_b)$ and $\mathrm{Var}(F_g)$. Consider that the fractions F_g and F_b are the result of Bernoulli trials. Let x_i be a random variable, which under the null hypothesis that alpha is zero, takes the value 1 if test *i* rejects the null (with probability $\gamma/2$) and 0 otherwise (with probability $1-\gamma/2$). Then under the null, $E(x_i) =(\gamma/2)=E(x_i^2)$ and $\mathrm{Var}(x_i) = (\gamma/2)(1- \gamma/2)$, and we have:

$$\mathrm{Var}(F_g) = \mathrm{Var}(F_b) = \mathrm{Var}((1/N)\Sigma_i\, x_i)= (\gamma/2)(1- \gamma/2)(1/N)[1+(N-1)\rho]. \qquad (A.3)$$

when there are N funds tested, and $\rho = [N(N-1)]^{-1}\, \Sigma_j\, \Sigma_{i\neq j}\, \rho_{ij}$ is the average correlation of the tests, where ρ_{ij} is the correlation between the tests for fund *i* and fund *j*.

[22] The coefficients are $D = (-\delta_b +\gamma/2)/G$, $E = (\beta_g -\gamma/2)/G$, $B = (\beta_b -\gamma/2)/G$, $C = (-\delta_g +\gamma/2)/G$, with $G = - (\delta_g -\gamma/2)(\delta_b - \gamma/2) + (\beta_b -\gamma/2)(\beta_g -\gamma/2)$. Setting the β parameters equal to 1.0 and the δ parameters equal to 0.0, then $1- \pi_b- \pi_g$ in (A.1) is equal to the estimator used by BSW.

We proxy for the correlation ρ by the average of the pairwise correlations of the mutual fund returns, adjusted for the extent of data overlap among the fund returns. The adjustment to the average correlation assumes that the correlations of tests for funds with no overlapping data are zero. The estimated correlation ρ is 0.044 for the mutual fund sample and 0.086 in the hedge fund sample. BSW estimate the same average correlation in their mutual fund sample, adjusted for data overlap (p. 193), of 0.08 (0.55) = 0.044.

To derive the standard errors we introduce indicator variables for tests rejecting the null hypothesis that fund *i* has a zero alpha in favor of the alternative that fund *i* is a good fund: $x_{ig} = I(t_i > t_g)$, where t_i is the t-statistic for fund *i*'s alpha and t_g is the empirical critical value for the one-sided t-test, computed by simulation under the null hypothesis. Similarly, $x_{ib} = I(t_i < t_b)$, where t_b is the empirical critical value for the alternative of a bad fund. Then, $F_b = (1/N)\Sigma x_{ib}$ and $F_g = (1/N)\Sigma x_{ig}$ and the variances and covariances of the sums are computed as functions of the variances and covariances of the x_{ib} and x_{ig}. We generalize $E(x_i)$ above to consider the conditional expectations of x_{ig} and x_{ib} given each of the three hypothesized values for the alpha parameters. The unconditional expectations of the x's are then computed as the averages of the conditional expectations, given the three subpopulations, weighted by the estimated π fractions. We find that the use of an asymptotic normal approximation in these calculations provides improved finite sample performance for the standard errors. The Internet Appendix provides the details.

References

Amihud, Y. and R. Goyenko, 2013, Mutual fund R^2 as a predictor of performance, *Review of Financial Studies* 26, 667-694.

Ardia, David and Kris Boudt, 2018, The peer performance ratios of hedge funds, *Journal of Banking and Finance* 87, 351–368.

Asness C., R. Krail, and J. Liew, 2001, Do hedge funds hedge? *Journal of Portfolio Management* 28, 6-19.

Avramov, Doron, and Russ Wermers, 2006, Investing in mutual funds when returns are predictable, *Journal of Financial Economics* 81, 339-377.

Bajgrowicz, Pierre, and Olivier Scaillet, 2012, Technical trading revisited: False discoveries, persistence tests, and transactions cost, *Journal of Financial Economics* 106, 473-491.

Bajgrowicz, Pierre, Olivier Scaillet, and Adrien Treccani. 2016, Jumps in high-frequency data: Spurious detections, dynamics, and news, *Management Science* 62, 2198–2217.

Baks, Klaas, Andrew Metrick, and Jessica Wachter, 2001, Should investors avoid all actively managed mutual funds? A study in Bayesian performance evaluation, *Journal of Finance* 56, 45-85.

Barras, Laurent, Olivier Scaillet and Russ Wermers, 2010, False discoveries in mutual fund performance: Measuring luck in estimated alphas, *Journal of Finance* 65, 179-216.

Barras, Laurent, Olivier Scaillet and Russ Wermers, 2010, Internet Appendix to: "False discoveries in mutual fund performance: Measuring luck in estimated alphas," *Journal of Finance* 65, 179-216, http://www.afajof.org/supplements.asp.

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

Berk, J., and J. H. van Binsbergen, 2015, Measuring skill in the mutual fund industry, *Journal of Financial Economics* 118, 1–20.

Brown, Stephen J., 1979, Optimal portfolio choice under uncertainty: A Bayesian approach. In: Bawa, V.S., Brown, S.J., Klein, R.W. (Eds.), Estimation Risk and Optimal Portfolio Choice. North Holland, Amsterdam, pp. 109–144.

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

Chen, Y., M. Cliff, and H. Zhao, 2017, Hedge funds: The good, the bad, and the lucky, *Journal of Financial and Quantitative Analysis* 52, 1081–1109.

Cremers, M. and A. Petajisto, 2009, How active is your mutual fund manager? A new measure that predicts performance. *Review of Financial Studies* 22, 3329-3365.

Criton, Gilles and Olivier Scaillet, 2014, Hedge fund managers: Luck and dynamic assessment, *Bankers, Markets & Investors* 129, 1–15.

Cuthbertson, Keith, D. Nitzsche and N. O'Sullivan, 2012, False discoveries in UK mutual fund performance, *European Financial Management* 19, 444-463.

Daniel, K., M. Grinblatt, S. Titman, and R. Wermers, 1997, Measuring mutual fund performance with characteristic-based benchmarks, *Journal of Finance* 52, 1035-1058.

Das, Sanjiv Ranjan, 2013, *Data science: Theories, models, algorithms and analytics*, a web book.

Dewaele, B., H. Pirotte, N. Tuchschmid and E. Wallerstein, 2011, Assessing the performance of funds of hedge funds, working paper, Solvay Brussels School.

Efron, B., and R. J. Tibshirani, 1993, *An introduction to the bootstrap*. Chapman & Hall.

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

Evans, R.B., 2010. Mutual fund incubation, *Journal of Finance* 65, 1581-1611.

Fama, Eugene F., and Kenneth R. French, 1996, Multifactor explanations of asset pricing anomalies, *Journal of Finance* 51, 55-87.

Fama, Eugene F., and Kenneth R. French, 2010, Luck versus skill in the cross section of mutual fund returns, *Journal of Finance* 65, 1915-1947.

Farnsworth, H., W. Ferson, D. Jackson, and S. Todd, 2002, Performance evaluation with stochastic discount factors, *Journal of Business* 75, 473-504

Ferson, W., T. Henry, and D. Kisgen, 2006, Evaluating government bond fund performance with stochastic discount factors, *Review of Financial Studies* 19, 423-456.

Ferson, W., and J. Lin, 2014, Alpha and performance measurement: The effects of investor disagreement and heterogeneity, *Journal of Finance* 69, 1565-1596.

Ferson, W., and H. Mo, 2016, Performance measurement with market and volatility timing and selectivity, *Journal of Financial Economics* 121, 93-110.

Fung, William, and David A. Hsieh, 2001, The risk in hedge fund strategies: Theory and evidence for trend followers, *Review of Financial Studies* 14, 313-341.

Fung, William, and David A. Hsieh, 2004, Hedge fund benchmarks: A risk based approach, *Financial Analysts Journal* 60, 65-80.

Genovese C., and L. Wasserman, 2002, Operating characteristics and extensions of the FDR procedure, *Journal of the Royal Statistical Society B* 64, 499–517.

Genovese C., and L. Wasserman, 2004, A stochastic process approach to false discovery control, *Annals of Statistics* 32, 1035-1061.

Getmansky M, Lo A, and Makarov I., 2004, An econometric model of serial correlation and illiquidity in hedge fund returns, *Journal of Financial Economics* 74, 529-610.

Harvey, C. and Y. Liu, 2018, Detecting repeatable performance, *Review of Financial Studies*, forthcoming.

Jones, Christopher, and Haitao Mo, 2016, Out-of-sample performance of mutual fund predictors, working Paper.

Jones, Christopher, and Jay Shanken, 2005, Mutual fund performance with learning across funds, *Journal of Financial Economics* 78, 507-552.

Kim, M.S., 2011, Changes in mutual fund flows and managerial incentives, *working paper*, University of New South Wales.

Kosowski, R., A. Timmerman, H. White and R. Wermers, 2006, Can mutual fund "stars" really pick stocks? Evidence from a Bootstrap Analysis, *Journal of Finance* 61, 2551-2569.

Pástor, Lubos, and Robert F. Stambaugh, 2002, Investing in equity mutual funds, *Journal of Financial Economics* 63, 351–380.

Pástor, L., R. F. Stambaugh, and L. A., Taylor, 2015, Scale and skill in active management, *Journal of Financial Economics* 116, 23-45.

Romano, Joseph, Azeem Shaikh and Michael Wolf, 2008, Control of the false discovery rate under dependence using the bootstrap and subsampling, *TEST* 17, 417-442.

Scholes, M. and J. Williams, 1977, Estimating betas from nonsynchronous data, *Journal of Financial Economics* 5, 309-328.

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

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

Titman, S. and C. Tiu, 2011, Do the best hedge funds hedge? *Review of Financial Studies* 24, 123-168.

White, H., 1980, A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica* 48, 817-838.

Table 1 Summary Statistics

Monthly returns are summarized for mutual funds and hedge funds, stated in monthly percentage units. The values at the cutoff points for various fractiles of the cross-sectional distributions of the sample of funds are reported. Each column is sorted on the statistic shown. Nobs is the number of available monthly returns, where a minimum of 8 are required. Mean is the sample mean return, Std is the sample standard deviation of return, reported as monthly percentages, and Rho1 is the first order sample autocorrelation in raw units. The alpha estimates are based on OLS regressions using the Fama-French three factors for mutual funds (3,619 mutual funds in the sample), and the Fung and Hsieh seven factors for hedge funds (3,620 hedge funds in the sample). The alpha t-ratios are based on heteroskedasticity-consistent standard errors.

		Monthly Returns (%)			Alphas (%)	Alpha T-ratios
Fractile	Nobs	Mean	Std	Rho1		
Panel A: Mutual Fund Returns: January 1984–December 2011 (336 months)						
0.01	335	1.95	9.88	0.38	0.505	2.407
0.05	276	1.30	7.72	0.28	0.270	1.474
0.10	214	1.11	7.04	0.24	0.170	1.020
0.25	148	0.88	6.13	0.19	0.031	0.194
Median	89	0.62	5.35	0.12	-0.094	-0.665
0.75	39	0.33	4.71	0.00	-0.234	-1.511
0.90	26	0.04	4.15	-0.09	-0.415	-2.381
0.95	23	-0.18	3.68	-0.13	-0.573	-2.952
0.99	12	-1.25	2.50	-0.21	-1.109	-4.024
Panel B: Hedge Fund Returns: January 1994–March 2012 (219 months)						
0.01	208	2.60	14.44	0.61	2.671	12.471
0.05	153	1.47	8.68	0.50	1.251	4.175
0.10	122	1.15	6.70	0.43	0.861	3.108
0.25	79	0.74	4.22	0.30	0.437	1.617
Median	46	0.37	2.62	0.16	0.112	0.428
0.75	25	-0.04	1.73	0.03	-0.210	-0.656
0.90	15	-0.63	1.21	-0.12	-0.732	-1.809
0.95	11	-1.25	0.95	-0.21	-1.356	-2.640
0.99	8	-3.20	0.52	-0.37	-3.697	-14.136

Table 2 Finite Sample Properties of the Estimators

In each of 1,000 bootstrap simulation trials artificial data are generated from a mixture of three fund distributions. The "population" values of the fractions of funds in each group, π, shown here in the first row, determine the mixture, combined with the good, zero or bad alpha values that we estimate as the best-fitting values for the full sample period. Hedge fund data over January 1994–March 2012 are used, and the values of the bad and good alphas are -0.138 and 0.262% per month. For each simulation draw from the mixture distribution we run the estimation by simulation with 1,000 trials, to generate the parameter and standard error estimates. Standard error estimates are removed for a given trial, when an estimated fraction is on the boundary of a constraint. The empirical SD's are the standard deviations taken across the remaining simulation draws. The Avg. estimates are the averages over the 1,000 draws. The Root MSE's are the square root of the average over the 1,000 trials, of the squared difference between an estimated and true parameter value. γ/2 indicates the size of the tests (the area in one tail of the two-tailed tests). The Classical FDR estimators follow Storey (2002) and BSW, except with a fixed test size.

	π_{zero}	π_{good}	π_{bad}
	Panel A: γ/2 = 0.05		
Population values	0.100	0.600	0.300
Our Avg. Estimates	0.417	0.408	0.175
Classical FDR Avg. Estimates	0.955	0.052	-0.007
Empirical SDs	0.270	0.246	0.240
Avg. Reported SDs	0.059	0.409	0.421
Root MSE	0.416	0.312	0.270
Classical Empirical SD	0.039	0.038	0.014
Classical Avg. Reported SD	0.007	0.008	0.013
Classical Root MSE	0.856	0.549	0.307
	Panel B: γ/2 = 0.10		
Population values	0.100	0.600	0.300
Our Avg. Estimates	0.266	0.523	0.211
Classical FDR Avg. Estimates	0.859	0.133	0.008
Empirical SDs	0.208	0.195	0.200
Avg. Reported SDs	0.247	0.313	0.414
Root MSE	0.266	0.210	0.219
Classical Empirical SD	0.056	0.062	0.030
Classical Avg. Reported SD	0.010	0.012	0.018
Classical Root MSE	0.761	0.471	0.293

Table 2, continued.

	π_{zero}	π_{good}	π_{bad}
	Panel C: $\gamma/2 = 0.20$		
Population values	0.100	0.600	0.300
Our Avg. Estimates	0.214	0.542	0.245
Classical FDR Avg. Estimates	0.784	0.195	0.022
Empirical SDs	0.173	0.161	0.170
Avg. Reported SDs	0.531	0.285	0.515
Root MSE	0.207	0.172	0.178
Classical Empirical SD	0.060	0.072	0.042
Classical Avg. Reported SD	0.014	0.015	0.024
Classical Root MSE	0.686	0.412	0.281
	Panel D: $\gamma/2 = 0.30$		
Population values	0.100	0.600	0.300
Our Avg. Estimates	0.203	0.549	0.248
Classical FDR Avg. Estimates	0.748	0.226	0.026
Empirical SDs	0.167	0.151	0.166
Avg. Reported SDs	0.753	0.321	0.628
Root MSE	0.196	0.160	0.174
Classical Empirical SD	0.066	0.081	0.050
Classical Avg. Reported SD	0.020	0.020	0.030
Classical Root MSE	0.651	0.382	0.279

Table 3 Analysis of Discovery Rates

This table presents simulated discovery rates for three methods. The values of the good alpha, α_g, in the simulated populations are shown in the first column and the fractions of good funds, π_g are shown in the second column. The discovery rates are the fractions of funds in the simulated sample that a test finds to be a positive-alpha fund, averaged over the 1,000 simulation trials. The total fraction discovered to be good is the sum of the False Discoveries and the Correct discoveries. The three methods are the simple t-test, the classical false discovery rate method (FDR Method) and our approach (CF Method). The symbol $\pi_{g,C}$ denotes the classical FDR estimate of the fraction of positive-alpha funds averaged across simulation trials. The symbol $\pi_{g,CF}$ denotes our average estimate. The simulations are based on a parametric bootstrap from a sample of 3,620 hedge funds during January 1994–March 2012.

Good α_g	Fraction	Simple t-test		FDR Method			CF Method		
	Good,π_g	Correct	False	$\pi_{g,C}$	Correct	False	$\pi_{g,CF}$	Correct	False
0.252	0.00	0.000	0.108	0.008	0.000	0.108	0.078	0.000	0.042
0.252	0.10	0.032	0.097	0.031	0.031	0.095	0.131	0.013	0.059
0.252	0.20	0.069	0.085	0.059	0.067	0.082	0.218	0.045	0.095
0.252	0.30	0.107	0.074	0.090	0.103	0.070	0.325	0.109	0.139
0.252	0.40	0.132	0.065	0.106	0.126	0.061	0.385	0.166	0.155
0.252	0.50	0.159	0.054	0.126	0.154	0.051	0.457	0.243	0.174
0.252	0.60	0.198	0.044	0.158	0.191	0.041	0.569	0.377	0.196
0.252	0.70	0.231	0.031	0.181	0.224	0.028	0.649	0.513	0.175
0.252	0.80	0.271	0.020	0.213	0.262	0.019	0.743	0.659	0.135
0.252	0.90	0.301	0.011	0.236	0.291	0.011	0.801	0.798	0.078
0.252	1.00	0.333	0.000	0.260	0.324	0.000	0.851	0.907	0.000
0.861	0.00	0.000	0.109	0.009	0.000	0.110	0.031	0.000	0.013
0.861	0.10	0.069	0.096	0.078	0.073	0.088	0.106	0.042	0.031
0.861	0.20	0.138	0.087	0.148	0.143	0.078	0.202	0.117	0.046
0.861	0.30	0.205	0.077	0.233	0.210	0.070	0.291	0.192	0.056
0.861	0.40	0.287	0.062	0.284	0.298	0.055	0.404	0.298	0.065
0.861	0.50	0.335	0.053	0.363	0.366	0.048	0.493	0.389	0.075
0.861	0.60	0.420	0.044	0.429	0.426	0.041	0.589	0.483	0.079
0.861	0.70	0.492	0.032	0.500	0.521	0.033	0.687	0.597	0.083
0.861	0.80	0.576	0.020	0.587	0.617	0.022	0.806	0.742	0.086
0.861	0.90	0.635	0.011	0.645	0.680	0.012	0.884	0.856	0.070
0.861	1.00	0.709	0.000	0.719	0.758	0.000	0.971	0.990	0.000

Table 4 Estimated Fractions of Mutual Funds

The fractions of funds in the population with specified values of zero, good or bad alphas, are estimated using simulation. The symbol π_g denotes the estimated fraction of good funds and π_0 denotes the fraction of zero-alpha funds. Alphas are stated in monthly percentage units. The power parameters of the test are β_g, the power to reject against the alternative of a good fund, and β_b, the power to reject against the alternative of a bad fund. The confusion parameters are δ_b, the probability of finding a good fund when it is bad, and δ_g, the probability of finding a bad fund when it is good. All fractions except for the test sizes are stated as (monthly, for the alphas) percentages. Empirical standard errors for the pie fractions are indicated in parentheses, except when a constraint is binding (na). Panel A presents the estimates with pre-set alpha values and various test sizes. Panel B summarizes the joint estimation of the alphas and π fractions using the 10% test size. Fit is the goodness of fit based on the Pearson χ^2 statistic. The sample period for mutual funds is January 1984–December 2011 (336 months).

Alphas (%)		Size	FDR Calcs.		Powers		Confusions		Fractions	
Good α_g	Bad α_b	$\gamma/2$	π_0	π_g	β_g	β_b	δ_g	δ_b	π_0	π_g
Panel A: Estimates for Given Alpha Values and Various Test Sizes										
0.317	-0.267	0.025	90.9	-0.8	50.0	40.7	0.3	0.3	74.7	0.0
									(5.8)	(na)
0.317	-0.267	0.05	87.0	-1.4	60.4	51.2	0.5	0.4	70.4	0.0
									(8.1)	(na)
0.317	-0.267	0.10	81.3	-1.7	73.5	66.9	0.9	0.7	65.6	0.0
									(10.8)	(na)
0.317	-0.267	0.20	77.0	-3.9	81.7	80.0	2.1	1.5	61.0	0.0
									(14.9)	(na)
0.317	-0.267	0.30	72.3	-4.4	87.7	83.7	3.4	2.5	56.9	0.0
									(22.0)	(na)
0.317	-0.267	0.40	67.9	-4.4	91.7	88.5	5.4	3.9	53.2	0.0
									(42.1)	(na)

Alphas (%)		Size	Fit	Powers		Confusions		Fractions	
Good α_g	Bad α_b	$\gamma/2$		β_g	β_b	δ_g	δ_b	π_0	π_g
Panel B: Joint Estimation of Alphas and Fractions in the Populations									
Unconstrained Alpha Domains, 3-Group Model									
-0.034	-0.204	0.10	2155	6.9	52.4	1.5	15.4	50.7	6.9
								(26.7)	(37.2)
Constrained Alpha Domains ($\alpha_g \geq 0$, $\alpha_b \leq 0$), 3-Group Model									
0.001	-0.173	0.10	3816	10.3	45.7	1.8	10.3	0.0	55.2
								(na)	(44.5)
2-Group Model									
0.0	-0.172	0.10	2862	na	45.5	na	na	48.4	0.0
								(37.8)	(na)
Single-Alpha Model									
na	-0.205	0.10	3401	na	na	na	na	0.0	0.0
								(na)	(na)

Table 5 Estimated Fractions of Hedge Funds

The fractions of funds in the population with specified values of zero, good or bad alphas, are estimated using simulation. The symbol π_g denotes the estimated fraction of good funds and π_0 denotes the fraction of zero-alpha funds. Alphas are stated in monthly percentage units. The power parameters of the test are β_g, the power to reject against the alternative of a good fund, and β_b, the power to reject against the alternative of a bad fund. The confusion parameters are δ_b, the probability of finding a good fund when it is bad, and δ_g, the probability of finding a bad fund when it is good. All fractions except for the test sizes are stated as (monthly, for the alphas) percentages. Empirical standard errors for the pie fractions are indicated in parentheses, except when a constraint is binding (na). Panel A presents the estimates with pre-set alpha values and various test sizes. Panel B summarizes the joint estimation of the alphas and π fractions using the 10% test size. Fit is the goodness of fit based on the Pearson χ^2 statistic. The sample period for hedge funds is January 1994–March 2012 (219 months).

Alphas (%)		Size	FDR Calcs.		Powers		Confusions		Fractions	
Good α_g	Bad α_b	$\gamma/2$	π_0	π_g	β_g	β_b	δ_g	δ_b	π_0	π_g
Panel A: Estimates for Given Alpha Values and Various Test Sizes										
0.317	-0.267	0.025	100	0.0	4.1	4.9	1.8	1.9	100.0	0.0
									(na)	(na)
0.317	-0.267	0.05	91.3	7.9	20.9	17.8	2.9	2.9	40.4	48.2
									(24.2)	(21.2)
0.317	-0.267	0.10	85.9	12.3	35.7	35.5	5.0	5.0	41.0	46.8
									(19.7)	(16.1)
0.317	-0.267	0.20	77.7	19.5	56.6	52.2	8.0	7.0	41.6	45.4
									(18.6)	(14.4)
0.317	-0.267	0.30	76.1	20.6	64.4	62.2	12.1	11.2	37.3	47.1
									(16.9)	(12.7)
0.317	-0.267	0.40	73.3	22.6	73.1	71.2	16.7	15.5	36.9	47.3
									(20.4)	(14.7)

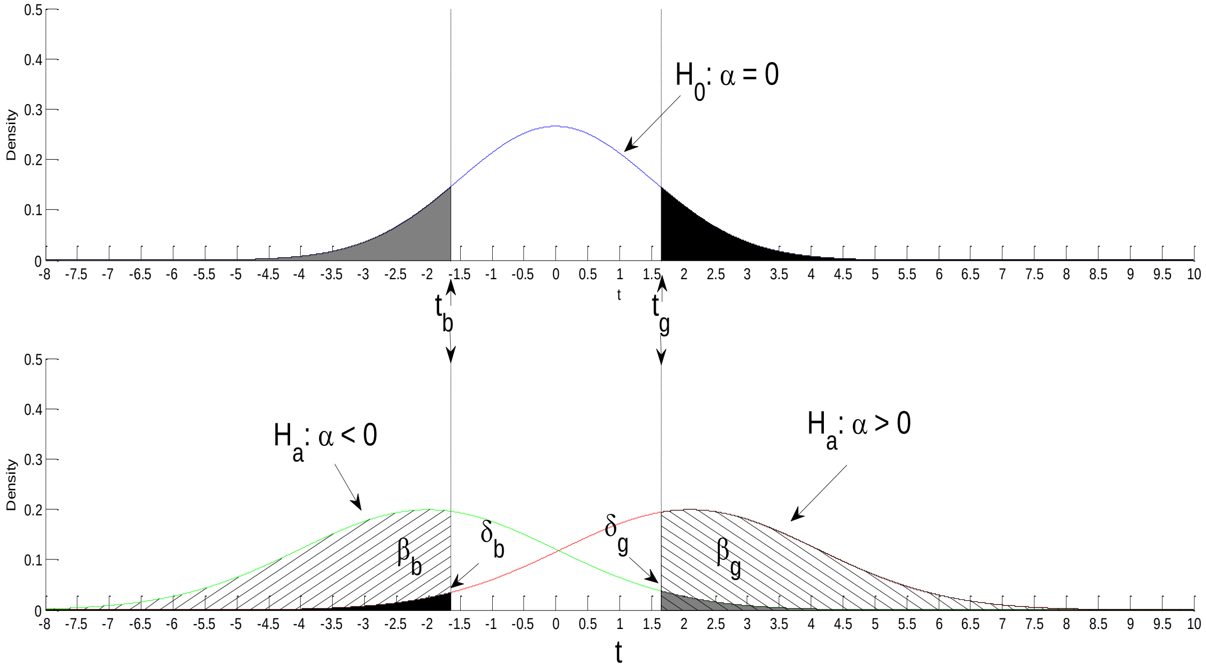
Alphas (%)		Size	Fit	Powers		Confusions		Fractions	
Good α_g	Bad α_b	$\gamma/2$		β_g	β_b	δ_g	δ_b	π_0	π_g
Panel B: Joint Estimation of Alphas and Fractions in the Populations									
Unconstrained Alpha Domains, 3-Group Model:									
0.252	-0.108	0.10	2225	31.4	17.4	7.0	5.0	0.0	53.2
								(na)	(16.8)
2-Group Model:									
0.252	na	0.10	4807	31.9	na	na	na	44.6	55.4
								(28.2)	(33.5)
Single-Alpha Model									
0.434	na	0.10	3120	na	na	na	na	0.0	100.0
								(na)	(na)

Table 6 Holding Period Returns after Bayesian Fund Selection

A 60-month rolling formation period is used to estimate the model of good, zero alpha and bad funds. The π fractions are estimated by simulation using a test size of 10% in each tail, recursive estimation and 1,000 simulation trials. Bayesian selection is used to assign funds to one of three groups, held for evaluation during the next four years. Good is the equal weighted portfolio of funds detected to have high alphas during the formation period, Zero is the portfolio of funds found to have zero alphas and Bad is the portfolio of low-alpha funds. G-B is the excess return of the good over the bad funds, during the months when both exist. N is the average number of funds in the holding portfolio returns during the holding period, taken over all of the formation periods, μ is the sample mean portfolio return during the holding periods, and α is the portfolio alpha, formed using the Fama-French factors for mutual funds and the Fung and Hsieh factors for hedge funds during the holding period. The holding period is one year in length and follows the formation period by one to four years. Mean returns and alphas are stated in monthly percentage units. T is the heteroskedasticity-consistent t-ratio.

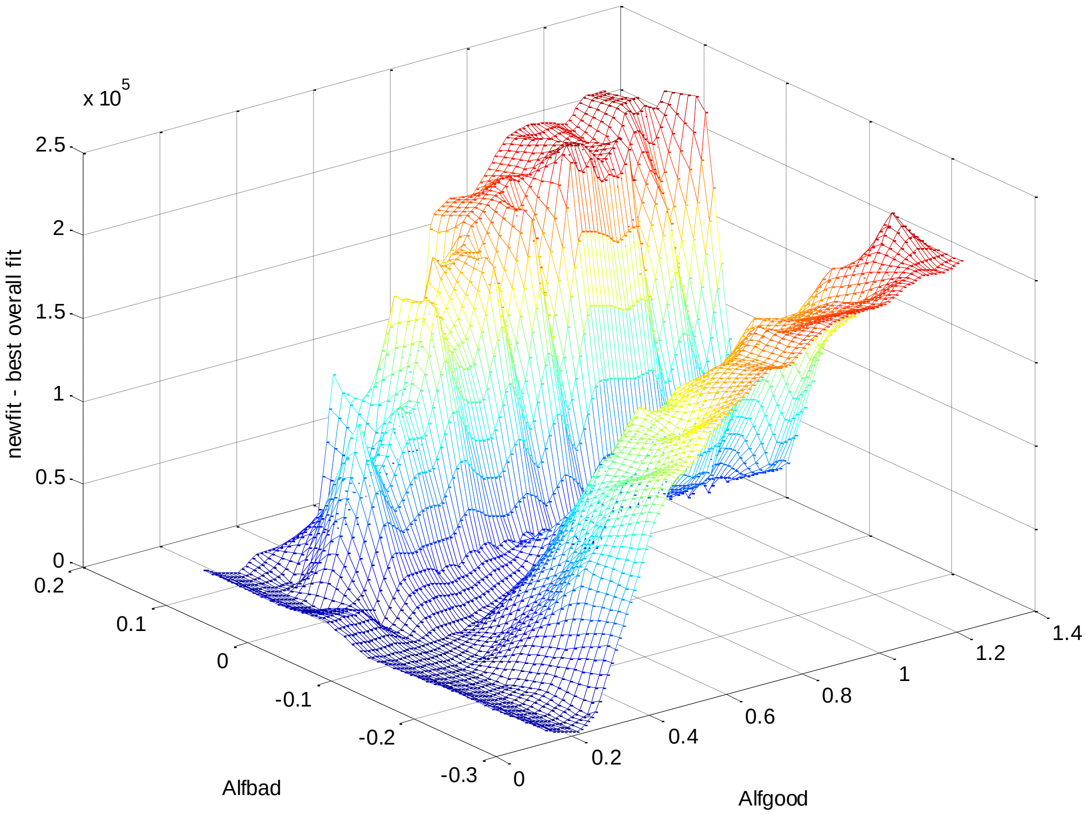
Portfolio		Years after Formation Period:											
			1			2			3			4	
	N	μ	α	T	μ	α	T	μ	α	T	μ	α	T
Panel A: Hedge Funds during January 1999–March 2012													
Good	377.7	0.35	0.24	3.0	0.23	0.20	2.4	0.26	0.16	1.9	0.28	0.17	2.0
Zero	245.0	0.38	0.28	3.6	0.12	0.11	1.3	0.21	0.13	1.3	0.40	0.24	2.3
Bad	7.8	-0.68	-0.57	-2.3	0.55	0.24	0.9	-0.08	0.16	0.5	0.96	0.78	2.1
G-B		0.41	0.59	2.3	0.06	-0.03	-0.1	-0.15	-0.10	-0.4	-0.50	-0.72	-2.0
Panel B: Mutual Funds during January 1989–December 2011													
Good	216.2	0.19	-0.01	-0.1	0.43	-0.11	-1.9	1.10	-0.07	-1.3	0.71	0.07	1.1
Zero	510.8	0.59	-0.07	-1.9	0.59	-0.03	-0.6	0.44	-0.13	-3.3	0.66	-0.06	-1.3
Bad	290.8	0.45	-0.13	-2.3	0.40	-0.15	-2.6	0.50	-0.11	-2.7	0.37	-0.17	-3.5
G-B		-0.01	0.09	1.0	0.10	0.14	1.8	0.19	-0.07	-1.1	0.23	0.29	3.8

Figure 1: Hypothetical Distributions of Alpha t-ratio for Three Subpopulations of Funds



The t-ratios for the three subpopulations are centered around a negative value (bad funds), zero, and a positive value (good funds), respectively. t_g and t_b are two critical values for the t-ratios corresponding to a test size of 10%, from a simulation under the null that the alpha is zero. The parameters β_g and β_b denote the power of the tests for good and bad funds, respectively, from the simulations where the alphas are at the good or bad values. The parameter δ_g (δ_b) captures the confusion, the probability that a bad (good) fund is mistaken for a good (bad) fund.

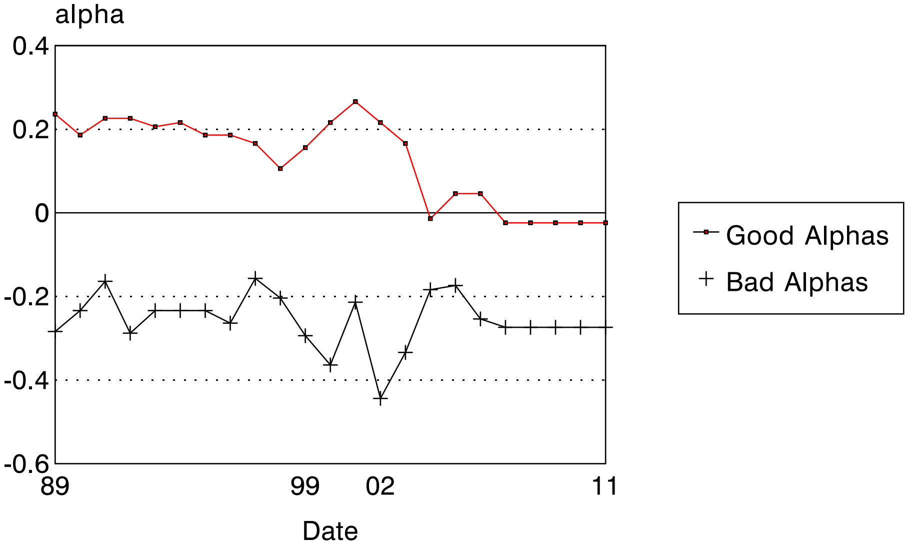
Figure 2: Simultaneous Estimation of α's and π's for Hedge Funds



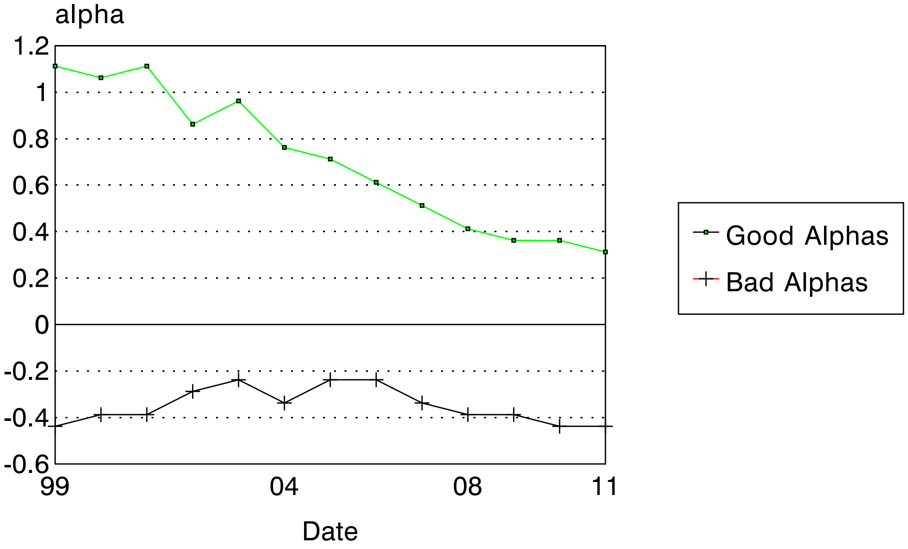
This figure depicts the results of grid search for the good and bad alpha parameters for the hedge fund sample. The vertical axis is the goodness of fit based on the Pearson χ^2 statistic in Equation (3). The sample period for hedge funds is January 1994–March 2012 (219 months).

Figure 3: 60-Month Rolling Alphas

Panel A: Mutual Funds



Panel B: Hedge Funds



Panel A depicts the time series of formation period estimates of good and bad mutual fund alphas. Panel B depicts the time series of formation period estimates of good and bad hedge fund alphas. The fund alphas are estimated jointly with the π-fractions over 60 month rolling windows. The date shown is the last year of the 60-month formation period. Alphas are stated in monthly percentage units.

Internet Appendix for:

How Many Good and Bad Funds Are there, Really?

Wayne Ferson and Yong Chen*

This version: April 18, 2018

* Ferson is the Ivadelle and Theodore Johnson Chair of Banking and Finance and a Research Associate of the National Bureau of Economic Research, Marshall School of Business, University of Southern California, 3670 Trousdale Parkway Suite 308, Los Angeles, CA. 90089-0804, ph: (213) 740-5615, ferson@marshall.usc.edu, www-rcf.usc.edu/~ferson/. Chen is Associate Professor of Finance at Mays Business School, Texas A&M University, ph: (979) 845-3870, ychen@mays.tamu.edu. This appendix is to accompany the working paper.

Introduction

This Internet Appendix contains ancillary results for the paper. Section A describes how our approach is related to previous methods. Section B summarizes the two-distribution version of the probability model. Section C describes our kernel density estimation. Section D provides details about the standard error calculations and further analysis of the standard errors. Section E compares our standard errors to those of the classical False Discovery Rate (FDR) methods similar to Barras, Scaillet and Wermers (BSW, 2010). Section F presents the results of additional robustness checks on the empirical results in the main paper. Section G presents experiments in simulating the bootstrap simulations for a wider range of parameter values than in the main paper. Section H describes results for mutual fund trading strategies that use the assignment of funds to the three groups based on the estimated alphas and π fractions, as well as strategies that use our method. Section I presents an analysis of the effects of the missing value patterns in the data on the simulations. Section J presents an analysis of the impact of errors in the δ and β parameters on the standard errors of the π fractions. Section K explores the sensitivity to alternative goodness-of-fit measures. Section L contains the Internet Appendix Tables.

A. Relation to Previous Approaches

This section describes in more detail than in the main text how our approach modifies the framework of BSW (2010) and Storey (2002). Consider the case of a two-tailed test with size γ and power $\beta < 1$, and let $F = (F_b + F_g)$, where F_b and F_g are the fractions of funds where the test rejects the null hypothesis of zero alphas in favor or bad or good alphas, respectively. Then:

$$\begin{aligned} 1- E(F) \quad &= P(\text{don't reject} \mid H_o)\, \pi_0 \; + \; P(\text{don't reject} \mid H_a)(1- \pi_0) \\ &= (1- \gamma)\, \pi_0 \; + (1- \beta)(1- \pi_0). \qquad (A.1) \end{aligned}$$

Solving (A.1) for π_0 gives the estimator:

$$\pi_0^* = [\beta - E(F)] / (\beta - \gamma). \tag{A.2}$$

BSW estimate the fraction of zero-alpha funds, π_0, following Storey (2002). Storey's classical estimator for π_0 is, in our notation:[1]

$$\pi_{0,C} = [1- (F_b + F_g)]/(1-\gamma). \tag{A.3}$$

As the Equation (A.3) indicates, this estimator assumes that the fraction of zero-alpha funds in the population, multiplied by the probability that the test will not reject the null of zero alpha when it is true, is equal to the fraction of funds for which the null of zero alpha is not rejected in the actual sample. But a test will also not reject the null for cases where alpha is not zero, if the power of the tests is below 100%. This motivates our modification to the case where $\beta < 1$.

Equation (A.3) is a special case of (A.2) when $\beta=1$. Storey motivates $\beta=1$ as a "conservative" choice, justified by choosing the size of the tests to be large enough. Comparing the two estimators, $E(\pi_{0,C} - \pi_0^*) >0$ when $E(F) > \gamma$ and $\beta>\gamma$, so this classical estimator is likely biased in favor of large values of π_0 when the power of the tests is below one.

Our estimators in the three-group model are found by solving the two equations (1) and (2) in the main paper, subject to the constraints that the π fractions are probabilities. When the constraints are not binding, the solutions are:

[1] In the estimation of π_0, BSW follow Storey (2002), choosing the value of γ such that in simulations, the sum of squares of the π_0 estimator in (A.1) is minimized around the minimum $\pi\hat{}_0$ value found for all γ values. This obviously reduces the estimate and the bias in the estimator of the fraction of zero alpha funds. BSW find that setting γ to 0.5 or 0.6 produces similar results, and that these results are not highly sensitive to the γ value. In our paper, the value γ is also used in Equation (A.3) as the threshold (denoted as λ in Storey (2002) and BSW) to estimate $\pi_{0,C}$, as BSW suggest a similarly large value (such as 0.5 or 0.6) for the threshold.

$$\pi_g = B\,(F_g - \gamma/2) + C\,(F_b - \gamma/2) \qquad (A.4)$$
$$\pi_b = D\,(F_g - \gamma/2) + E\,(F_b - \gamma/2),$$

where the coefficients are $D = (-\delta_b + \gamma/2)/G$, $E = (\beta_g - \gamma/2)/G$, $B = (\beta_b - \gamma/2)/G$, $C = (-\delta_g + \gamma/2)/G$, with $G = -(\delta_g - \gamma/2)(\delta_b - \gamma/2) + (\beta_b - \gamma/2)(\beta_g - \gamma/2)$. Our estimator for π_0 derives from (A.4) as $1 - \pi_b - \pi_g$.

Setting the β parameters equal to 1.0 and the δ parameters equal to 0.0, then $1 - \pi_b - \pi_g$ in (A.4) is equal to the classical FDR estimator $\pi_{0,C}$. At these parameter values, assuming $F_b > \gamma/2$ and $F_g > \gamma/2$, we find that $\partial\pi_0/\partial\delta_b > 0$ and $\partial\pi_0/\partial\delta_g > 0$. Setting the δ-parameters to zero biases the estimator of π_0 towards zero. There are two offsetting biases. Setting the β-parameters to 1.0 creates an upward bias in the estimator, while setting the δ-parameters to 0.0 creates a downward bias in the estimator.

In the BSW analysis a central focus is estimating the fractions of skilled and unskilled managers, π_g and π_b. We would call this high and low-performance funds, acknowledging that the after-cost alphas used in both of our studies are better measures of the performance available to investors than they are measures of fund skill. BSW estimate the fractions by subtracting the expected fraction of "lucky" funds, $P(\text{reject at } t_g | H_o)\,\pi_0 = (\gamma/2)\,\pi_0$, from the observed fraction of funds where the null hypothesis of zero alpha is rejected in favor of the alternative of a positive alpha. In our notation the classical estimator for π_g is $F_g - (\gamma/2)\,\pi_{0,C,}$ where $\pi_{0,C}$ is the estimator in (A.3).

The classical estimator separates skill from luck by subtracting from F_g, the fraction of "lucky" funds, that the test says have positive alphas but which actually have zero alphas, $(\gamma/2)\,\pi_{0,C}$. Our approach can better separate skill from luck, because we also consider the expected fraction of "extremely lucky" bad funds, where the test is confused and indicates that the fund is good.

B. A Two-Distribution Model

With two distributions we use two-sided tests of the null hypothesis. Solving the probability model for an estimator of π_0 in the two-distribution model with only a zero alpha and a positive alpha, the parameters β_b, δ_b and δ_g are no longer relevant and we obtain:

$$\pi_0 = [\beta_g - E(F_g)] / (\beta_g - \gamma), \qquad \text{(A.5)}$$

where γ is the size of the one-tailed test of the null of zero alpha against the alternative of a positive alpha. We use this version of the model in our simulations that examine the discovery rates in the main paper. The two-distribution model with only a zero and a bad alpha modifies (A.5) in the obvious way. The standard error of the estimator follows from (A.5) and the variance of F_b, described in the next section.

C. Kernel Smoothing

We employ a kernel smoothing function to estimate the simulation-generated empirical densities of the alphas, conditional on each subpopulation of funds. Let the sample of the relevant cross-section of funds' alpha estimates be $\{x_i\}_{i=1,\ldots n}$. The estimator for the density function evaluated at some particular value, x, is:

$$f(x) = (1/nh)\, \Sigma_i\, K((x-x_i)/h), \qquad \text{(A.6)}$$

where the kernel function K(.) must be strictly positive and integrate to 1.0. The symbol h denotes the bandwidth parameter and n is the number of observations. We choose the Epanechnikov optimal kernel function:

$$K(u) = (3/4)(1-u^2)\, I(|u|<1), \qquad \text{(A.7)}$$

with the bandwidth parameter that approximately minimizes the mean integrated squared error of the kernel approximation for second order kernels (see Hansen (2009), Section 2.7), which is 0.374 in our application when the number of funds in the sample is 3865.

When the conditional distributions are estimated from simulations and there are 1,000 trials, there are 3,865,000 observations of the x_i, which is unwieldy. To keep the problem of manageable size, we use the 3865 observations from the first simulation trial. We experiment with concatenating the observations from the first k simulation runs, and adjust the bandwidth of the kernel according to Silverman's rule of thumb, multiplying it by $k^{(-1/5)}$. We also experiment with using the means of the simulated alphas for each fund taken across the simulation trials to characterize the conditional distributions. Neither of these alternatives changes the results much in the full sample. In the rolling, 60-month analysis, k=1 produces some instability across simulation trials. We find that recursive estimation with k=2 produces more stable results, so we recursive estimation with k=2.

D. Analysis of the Standard Errors

We introduce indicator variables for the event where a test rejects the null hypothesis that fund *i* has a zero alpha in favor of the alternative that fund *i* is a good fund: $x_{ig} = I(t_i > t_g)$, where t_i is the t-statistic for fund *i*'s alpha and t_g is the empirical critical value for the one-sided t-test, computed by simulation under the null hypothesis. Similarly, $x_{ib} = I(t_i < t_b)$, where t_b is the empirical critical value for the alternative of a bad fund. Then, $F_b = (1/N)\Sigma x_{ib}$ and $F_g = (1/N)\Sigma x_{ig}$ and the variances and covariances are computed as functions of the variances and covariances of the x_{ib} and x_{ig}. We use the fact that the expectation of an indicator variable is the probability that it takes the value 1.0, and the expected value is the expected value of the square. We compute $\text{Var}(x_{ig}) = E(x_{ig}^2) - E(x_{ig})^2 = E(x_{ig}) - E(x_{ig}),^2$ and $\text{Cov}(x_{ib},x_{ig}) = - E(x_{ib}) E(x_{ig})$.

Since our calculations allow for dependence across funds, the standard errors depend on covariance terms for funds $i\neq j$: $\text{Cov}(x_{ib},x_{jb})$, $\text{Cov}(x_{ig},x_{jg})$ and $\text{Cov}(x_{ib},x_{jg})$. We make an asymptotic normality assumption for the t-ratios and use the bivariate normal probability function with correlation ρ described in the main text to compute these covariances, as well as the expectations like $E(x_{ib})$. The unconditional expectations of the x's are computed as the averages of the conditional expectations, given the three subpopulations, weighted by the estimated π fractions. To compute the probability that the t-ratio exceeds a critical value we require the expected t-ratio given the hypothesized value of alpha. These conditional expected values of the t-ratios in the subpopulations are approximated as $\mu_g = \alpha_g/\sigma(\alpha)$, where α_g is the alpha value assumed for the good funds, and $\sigma(\alpha)$ is the average across all funds, of the consistent standard error estimate for alpha. The expected values for the t-ratios of the bad funds are similarly approximated as $\mu_b = \alpha_b/\sigma(\alpha)$.[2]

Let F(x,y) be the lower tail region of the bivariate normal cumulative distribution function with correlation equal to ρ. The calculations for the covariances are illustrated with the following example.

$$\begin{aligned}
\mathrm{Cov}(x_{ib},x_{jg}) &= E(x_{ib}\, x_{jg}) - E(x_{ib})\, E(x_{jg}) \qquad\qquad (A.8)\\
&= F(t_b, -t_g)\, \pi_0^2 + F(t_b, \mu_g - t_g)\, \pi_0\, \pi_g + F(t_b, \mu_b - t_g)\, \pi_0\, \pi_b\\
&\quad + F(t_b - \mu_b, -t_g)\, \pi_b\, \pi_0 + F(t_b - \mu_b, \mu_g - t_g)\, \pi_b\, \pi_g + F(t_b - \mu_b, \mu_b - t_g)\, \pi_b^2\\
&\quad + F(t_b - \mu_g, -t_g)\, \pi_g\, \pi_0 + F(t_b - \mu_g, \mu_g - t_g)\, \pi_g^2 + F(t_b - \mu_g, \mu_b - t_g)\, \pi_g\, \pi_b\\
&\quad - E(x_{ib})\, E(x_{jg}).
\end{aligned}$$

Note that in (A.8) we have used the symmetry of the normal density, implying $\Pr(t>t_g) = \Pr(-t<-t_g)$. When symmetry is used for only one of the arguments of F(.,.), the correlation is $-\rho$ instead of ρ. We evaluate the expectations like $E(x_{ib}) = F(\infty,t_b)\, \pi_0 + F(\infty,t_b - \mu_b)\, \pi_b + F(\infty,t_b - \mu_g)\, \pi_g$. We use the asymptotic normality assumption here in the calculations, instead of the estimated β and δ parameters. The bivariate normal probabilities do not exactly match the empirical β and δ parameters, which are estimated under non normal distributions, and using them we would have either to empirically estimate all of the joint probabilities in (A.8) by

[2] Since we assign the nonzero alphas to funds randomly in our bootstrap simulations we use the unconditional standard errors of the alphas here. In the robustness section in the main text, we

simulation, or make other strong simplifying assumptions.

We also develop a version of our standard error estimator that avoids the asymptotic normality assumption for the t-ratios. We estimate $E(F_b) = E(x_{ib}) = (\gamma /2)\ \pi_0 + \delta_b\ \pi_g + \beta_b\ \pi_b$, and similarly for $E(F_g)=E(x_{ig})$. We estimate $Var(x_{ib}) = E(x_{ib})\ \{1-E(x_{ib})\}$ and use the correlation, ρ, described above to approximate $Cov(x_{ib},x_{jg})$ and $Cov(x_{ib},x_{jb})$ as the correlation times the product of the standard deviations. However, when we simulate the simulations to evaluate the finite sample performance of this version of the standard errors, we find that they perform much worse.

In the two-distribution example for mutual funds, the standard errors of π_0 and π_b are equal. The standard errors follow from Equation (A.4) with $Var(F_b) = Var(x_{ib})[(1/N) - \rho(1-1/N)]$, $Var(x_{ib}) = E(x_{ib})[\ 1-E(x_{ib})]$ and $E(x_{ib}) = \pi_0\ CDFN(t_b) + CDFN(t_b - \mu_b)\ (1-\pi_0)$, where CDFN(.) is the standard normal cumulative distribution function.

These expressions for the variances of the pie fractions hold when the constraints that the fractions are positive and sum to less than 1.0 are not binding. When the constraints are binding the distribution of the estimators is complicated by the truncation, and it involves the sampling variance of the Lagrange multipliers. We do not report standard errors when the constraints are binding.

E. A Comparison to BSW's Standard Errors

The standard errors in BSW are based on Equation (A.1), which implies:

$$Var(\pi_{0,C}) = Var(F)/(1-\gamma)^2, \qquad \text{(A.9)}$$

where Var(F) is computed as the sum of Bernoulli random variables, specialized to the case of a two-sided test, with $F = F_g + F_b$, and ignoring dependence across the tests (ρ=0). Thus, BSW use $Var(F) = \gamma(1-\gamma)/N$ when computing their standard errors. Similarly, from (A.1), their

build in a relation between the standard deviations and the alphas in the different fund groups.

estimate of π_g is F_g - $(\gamma/2)$ $\pi_{0,BSW}$, and they use $Var(\pi_{g,C}) = Var(F_g) + Var(\pi_{0,C})^* (\gamma/2)^2 - 2(\gamma/2) Cov(F_g, \pi_{0,C})$, with $Var(F_g) = F_g (1-F_g)/N$ and $Cov(F_g, \pi_{0,C}) = F_g (1-F)/[N(1-\gamma)]$. It follows that $Var(\pi_{b,C}) = Var(F_g) + Var(\pi_{0,C}) (1 - \gamma/2)^2 + (1-\gamma/2) Cov(F_g, \pi_{0,C})$.

The BSW calculations do not use the asymptotic normal approximations that we employ, and they do not arrive at the unconditional expectations by averaging over the conditional expectations given each fund group, weighted by the estimated fractions in each group, as in our standard error estimators.

F. Robustness Results Mentioned in the Main Text

Tables A.1–A.5 present ancillary results described in the main text. This section provides additional information for interpreting these tables.

Table A.1 presents summary statistics for the hedge fund returns, comparing the effects of different benchmarks, different numbers of required observations and the effects of return smoothing. In the main text, Table I, the minimum number of observations is 8, while here it is 12. The summary statistics show that the cross-sectional distributions of the means, standard deviations and autocorrelations are quite similar. The cross-sectional distributions of the alpha estimates and t-alphas are presented for hedge funds using the Fama and French (1996) three-factor model. For ease of comparison, the results using the Fund and Hsieh seven-factor model as in the main text are shown here as well. However, the minimum number of observations here is 12, so comparing these figures with the main text shows the small impact of requiring more observations. The three-factor and seven-factor models produce similar cross-sectional distributions of alphas and their t-ratios. Finally, the impact of return smoothing is examined by running the seven-factor model using the Scholes Williams beta estimator. The cross-sectional distribution is very similar to that of the seven-factor model without the lagged betas included.

Table A.2 reports the results of two additional analyses mentioned in the main text. The first one examines the effects of the alpha locations on the estimation results; and the second reports the robustness of the joint estimation results to the use of the 5% test size.

In Panel A of Table A.2 we fix the size of the tests and vary the locations of the good and bad alphas. We use the preferred sizes: $\gamma/2 = 0.10$ for our estimator, and $\gamma/2 = 0.30$ for the classical estimator. We first examine mutual funds and then hedge funds. The first row examines the case where we set the alpha values to the median of the estimated alphas across all of the mutual funds, plus or minus 0.01%. For these values the power parameters of the tests are small (4.1% and 29.5%) because the null and alternatives are very close, and the δ errors can be large. The value of δ_b is 25.6%, indicating a high risk of rejecting the null in favor of a bad fund when the fund is truly good. Thus, in settings where the null and alternatives are close together, our modification of the model to account for nonzero confusion parameters could be important. Then, we examine values for the alphas that correspond to the estimates at the boundaries of the 25%, 10% and 5% tail areas in the actual sample, as shown in Table 1. The classical FDR estimates of the fraction of zero alpha mutual funds are very similar across the rows, at 72-73%, as they do not refer to the locations of the good and bad fund alphas. Our estimates of the π fractions and the standard errors are highly sensitive to the alpha values. The π_0 estimate varies from 0.0 to 78% as the alpha values move from the center of the distribution to the extreme tails. In panel B of Table A.2, we report the findings for the hedge fund sample, and the impression from hedge funds here is similar to that from mutual funds. The dependence of the inferences about the fractions on the choice of the alpha locations of the good and bad funds motivates our simultaneous estimation of the alphas and the π fractions.

In Table A.2, we also examine the robustness of the joint estimation results to the choice of the test size, 5% and 10%. As explained in the main text, the test size of 10% is our preferred test size, and the results from the 10% test size are reported in the main text. Here, we find that the results are similar in general when a test size of 5% is used.

Table A.3 repeats the analysis of Table IV using the Fama and French (1996) three-factor model for hedge funds, and finds similar results those reported in the main paper. The difference is that the reported standard errors, not the empirical standard errors from the simulations, are displayed here. This provides a feel for the biases in the reported standard errors and shows the increase in the reported standard errors as the test size increases.

Table A.4 presents results for joint estimation of the three-group model, with mutual fund results in Panel A and hedge fund results in Panel B. The joint estimation here uses 100 simulation trials, but in a few untabulated experiments we find that the results for 100 and 1,000 trials usually differ only in the last decimal place. Panels C-E summarize robustness checks. Panel C reproduces results from the baseline case in the main text for comparison. Table D presents results where the alpha parameters are associated with active management, measured by low R-squares in factor model regressions for the funds returns. Funds are ranked on their R-squares and when assigned to the three alpha groups, the low R-square funds go into the high-alpha group and the high R-square funds go into the low-alpha group.

Panel E of Table A.4 presents results where we account for estimation error in mutual funds' alphas, as described in the main text. After we subtract the random alpha to account for estimation error, we rescale the adjusted simulated fund returns so that they still match the standard deviations of the actual fund returns in the data, and we add a constant to preserve the means of the transformed simulated fund returns. The transformation produces $r^* = wr + x$, where r is the simulated return with the random alpha, $w = [\sigma^2(r)/(\sigma^2(\alpha)+ \sigma^2(r))]^{1/2}$, where $\sigma^2(r)$ is the variance of the fund returns before adjustment and $\sigma^2(\alpha)$ is the variance of the estimated alpha, and $x = E(r)(1-w)$.

Table A.5 presents results from joint estimation of the two-group model for mutual funds, using a range of test sizes. The result that the fraction of mutual funds having the bad alpha is 100% in the two-group model is robust to the size of the test, and the estimate of the bad alpha varies between -0.136% per month and -0.203% per month as the size of the test is varied between 1% and 40%.

G. Simulations of the Simulations: Additional Experiments

Panels A-D of Table A.6 present simulation results that evaluate the standard errors when the population values of the π-fractions are each set to 1/3. The simulated samples are of the same size as the samples in our actual data, and 1,000 simulation trials are used. The results are similar to those in Table II of the main paper. Our point estimates are typically within one empirical standard deviation of the true values of the π fractions at the 5% test size, and are even closer to the true values than the example in the main text, at all sizes above 10%. The standard errors are the most accurate at the 10% test size, but understated, and the average reported standard errors get larger with the larger test sizes. Unlike the case in the main paper, there seems to be no benefit to using test sizes larger than 10%. The results show that the findings in the main paper are conservative, in the sense that the fractions of zero alpha funds are likely even smaller than our estimates indicate, and the fractions of good and bad hedge funds are likely even larger.

Panels E-H of Table A.6 present the results of experiments where we set the true π fractions equal to the banner values reported by BSW: π_0=0.75, π_g=0.01 and π_b=0.24. When the test size is 5% in each tail, the BSW estimator of π_0 is upwardly biased, with an average estimate of 1.01. The upward bias remains at the larger test sizes, and the average estimate is 0.93 at the 30% size. Our point estimate is also upwardly biased, averaging 0.829 at size 5%, but becomes more accurate at the 10% size, averaging 0.740 when the true value is 75%. The average values of our estimates are always within one empirical standard deviation of the true value, and the patterns in both the reported and the empirical standard errors are similar to what we observed at the other parameter values. Our standard errors are understated for the larger π_0 estimates and overstated for the other two fractions. Moving to larger test sizes beyond 10% our point estimates do not improve, but the empirical standard errors get larger and the reported standard errors get much larger, resulting in dramatic overstatement in the standard errors at the largest test sizes. Overall, the results confirm the patterns reported in the main text, and suggest that the 10% test size results in the best overall performance of the estimators.

Panel I presents the results of experiments in which we extend the use of the asymptotic normality assumption in computing the standard errors, using it for the computation of the point estimates of the π fractions. We use the expectations of the x's described above in place of the simulation-generated β and δ parameters. This simpler approach has the advantage that joint estimation of the π fractions and alpha parameters can be conducted using only one instead of three simulations. The results show that the point estimates are similar, and the standard error patterns are similar to what we report for the base case in the main paper. The last part of the panel runs the sample size up to T=5,000, but with only 100 simulation trials. This suggests that the point estimates making use of the asymptotic normality assumption are consistent estimators, but the standard errors under the simpler approach remain overstated relative to the variation across the bootstrap simulation trials.

Panels J and K of Table A.6 present the results of experiments where we increase the size of the time-series samples used in the simulations to 5,000 observations for the base simulation results. We use again only 100 simulation trials in these experiments, given the computational requirements. In panel J the test size is 5% in each tail and the true π fractions are set equal to (.10, .60, .30). In panel L each of the true fractions is set equal to 1/3. Our point estimates are within 3% of the true values. This is about the magnitude of the simulation errors that we experience using 100 trials in the simulations. The average BSW estimate of π_0 is 26.7% when the true value is 10%, indicating that the upward bias in the estimator remains in large samples, and the estimated fractions of bad funds is about 10% too low. All of the standard errors approach zero as the sample sizes grow, of course, and at these sample sizes the classical standard errors are quite accurate. Our standard errors remain overstated at the large sample sizes when $(\gamma/2)=0.05$. In panel K when the true values of the π fractions are each 1/3, and the classical estimates of π_0 remain overstated, averaging 49%. Our average estimates are closer, but can be off by as much as 13%. Our standard errors remain overstated at the 5% test size, and the other results are similar.

Panels M and N report the results of experiments where we use the large sample sizes, T=5,000, and use the estimators which maximize the use of the asymptotic normality assumption. The results are similar to those for the baseline estimator. Our point estimates are within about 3% of the true values, while the classical estimates of π_0 are overstated, and our standard errors remain overstated.

H. Trading Strategies

H.1. Using False Discovery Rate Methods

The analysis of BSW (2010) finds that by controlling for the rate of false discoveries it is possible – at least early in the sample -- to better identify positive performance mutual funds and profit thereby. The False Discovery Rate (FDR) is the expected fraction of lucky funds, in the set of funds where the tests reject the null of zero alphas in favor of a good fund. BSW compute the false discovery rate as FDR = $\pi_{0,BSW}$ $(\gamma/2)/F_g$, in our notation. The false discovery rate is a natural extension of the idea of the size of a test to a multiple-comparisons setting. Controlling the FDR involves using simulations, searching for the value of γ that delivers the desired value of the FDR, where both $\pi_{0,BSW}$ and F_g depend on the chosen γ. Bootstrapping under the null hypothesis of zero alphas determines a critical value for the alpha t-statistic for a test of the optimal size, and all funds with t-statistics in the sample above this critical value are selected as good funds. Portfolios of good funds are formed in this way during a series of rolling formation periods, and their performance is examined during subsequent holding periods.

We implement a version of this strategy using our model. The evidence in BSW for economic significance when selecting good mutual funds using the FDR approach is not very strong (alphas of 1.45% per year or less, with p-values for the alphas of 4% or larger). Given our evidence that there are no significant positive-alpha mutual funds, such weak results are not surprising. However, we do find that not all mutual funds have zero alphas. There are plenty of bad mutual funds and investors might benefit from attempting to avoid the bad funds. Previous studies suggest that the most significant persistence in fund performance is that of the bad funds (e.g. Carhart, 1997). We therefore also examine the performance of strategies that use our results to detect and avoid bad mutual funds.

We modify the FDR approach to detecting bad funds by controlling the expected fraction of funds that the tests find to be bad, but which are not really not bad. The fraction of these unlucky funds, as a ratio to the total number of funds that the tests find to be bad, is the false discovery rate for bad funds:

$$FDR_b = [(\gamma/2)\pi_0 + \delta_b \pi_g]/F_b. \tag{A.10}$$

This modification of the FDR considers the "unlucky" funds with zero alphas, as in the classical FDR approach, and also the "very unlucky" funds with positive alphas, that were confused with bad funds by the tests. We also form strategies that attempt to find good funds by controlling the false discovery rate for good funds, which in our more general model takes the following form:

$$FDR_g = [(\gamma/2)\pi_0 + \delta_g \pi_b]/F_g. \tag{A.11}$$

Following BSW, we pick a rolling, 60 month formation period. The first formation period ends in December of 1998 for hedge funds and in December of 1988 for mutual funds. The cross section during a formation period includes every fund with at least 8 observations (12 for hedge funds) during the formation period.

We first jointly estimate the alphas and the π-fractions for each formation period and we hold these alphas fixed for that formation period. [3] Figures 3 and 4 in the main text illustrate the rolling good and bad alpha values. For a given level of the false discovery rate, we run a

[3] This exploits our observation that the alphas we find with joint estimation are not very sensitive to the value of γ, and saves considerable computation time.

grid search over the test size, ($\gamma/2$), estimating the π fractions at each point in the grid, along with the δ's and β's, to find the choice that makes the FDR expressions in Equations (A.10) and (A.11) the closest to the target value of the FDRs, minimizing the absolute deviation between the target and the value of the expressions. These estimates by simulation over the formation period determine critical values for the alpha t-ratios, and a set of funds are chosen to be good or bad funds based on those critical values. An equally weighted portfolio of the selected funds is examined during a holding period. If a fund ceases to exist during a holding period, the portfolio allocates its investment equally among the remaining funds.

The holding period is a one-year future period: Either the first, second, third or fourth year after formation. The 60-month formation period is rolled forward year by year. This gives us a series of holding period returns for each of the first four years after formation, starting in January of 1999 for the hedge funds and in January of 1989 for the mutual funds. The holding period returns for the fourth year after formation start in January of 2002 for the hedge funds, and in January of 1992 for the mutual funds. We show results for equally-weighted portfolios of the selected good funds (Good), the zero-alpha funds (Zero) and the selected bad funds (Bad). We also present the excess return difference between the good and the bad (G-B) using only those months where both good and bad funds are found.

The results of the trading strategies are summarized in Table A.7. The first columns show the averages of the size of the test chosen to control the false discovery rates, where the FDR is chosen following BSW to be 30% in each tail. The test sizes that best match are between 10% and 26%. Also shown is the average over the formation periods of the number of funds in each portfolio. For the hedge funds in Panel A, many more good than bad funds are found. For the mutual funds in Panel B, there are many more zero-alpha funds and bad funds than there are good funds, but some good funds are found. Early in the evaluation period there are more good than bad funds, while later in the sample there are more bad funds than good funds.

For the hedge funds in Panel A, both the mean excess returns and the alphas are ordered roughly as expected across the good, zero and bad groups during the first two years after portfolio formation, although many of the groups have positive point estimates of alpha. In the first year the G-B excess return alpha is 0.13% per month, or about 1.5% per year, with a t-ratio of 1.8. During the second year the G-B excess return is similar in magnitude but the t-ratio is only 1.5. During the third and fourth years after portfolio formation there is no economic or statistically significant difference between the fund groups.

The mutual funds in Panel B also display mean excess returns and alphas that are ordered as expected across the groups during the first year, but most of the alphas are negative. The bad group has an alpha t-ratio of -2.4, and the G-B excess alpha is 0.08% per month, with a t-ratio of 1.4. During the second through fourth years after portfolio formation some groups have statistically significant negative alphas, but there is no economic or statistically significant difference between the mutual fund groups.

H.2. Using Fund Group Membership

We examine a trading strategy based on assigning funds to the various alpha groups in the simplest possible way. We form portfolios of funds based on the estimated fractions in each subpopulation, and on the alpha estimates for each fund, in the formation period. Funds are sorted each year from low to high on the basis of their formation period alphas, and they are assigned to one of the three groups according to the current estimates of the π fractions. Equally-weighted portfolios of the selected funds are examined during a subsequent holding period, just like in the previous exercise. The average returns and their alphas during the holding periods are shown in Table A.8.

For the hedge funds in Panel A, many more good than bad funds are found. For the mutual funds in Panel B, there are many more zero-alpha funds and bad funds then there are good funds. But, some good funds are found. Early in the evaluation period there are more good than bad funds, while later in the sample there are more bad funds than good funds.

For the hedge funds in Panel A, both the mean excess returns and the alphas are ordered roughly as expected across the good, zero and bad groups during the first two years after portfolio formation, although many of the groups have positive point estimates of alpha. In the first year the G-B excess return alpha is 0.33% per month, or about 4% per year, with a t-ratio of 2.1. During the second year the G-B excess return is 0.26% per month and t-ratio is 1.3. During the third and fourth years after portfolio formation there is no economic or statistically significant difference between the fund groups.

The mutual funds in Panel B also display mean excess returns and alphas that are ordered as expected across the groups during the first year, but all of the alphas are negative. The bad group has an alpha t-ratio of -2.2, and the G-B excess alpha is 0.12% per month, with a t-ratio of 1.8. During the second through fourth years after portfolio formation some groups have statistically significant alphas, but there is no economic or statistically significant difference between the mutual fund groups.

I. The Impact of Missing Data Values on the Simulations

Our baseline simulations use a cross-sectional bootstrap method similar to Fama and French (2010). There is a potential issue of inconsistency in this procedure. Since we draw a row from the data matrix at random, the missing values will be distributed randomly through "time" in the artificial sample, while they tend to occur in blocks in the original data. The number of missing values will be random and differ across simulation trials. The bootstrap can be inconsistent under these conditions.

In this experiment we exploit the fact that the beta and alpha estimates for funds are the results of a seemingly-unrelated regression model (SURM), with the same right-hand side variables for each fund. Thus, equation-by-equation OLS produces the same point estimates of the alphas as does estimation of the full system. We bootstrap artificial data for each fund, *i*, separately, drawing rows at random from the data matrix (f, rf, r_i), which concatenates the factors (f), the risk-free rate (rf) and the returns data for fund *i*, r_i. If we encounter a missing value for a fund, we keep searching randomly over the rows until we find one that is not missing, and we include the nonmissing value with its associated monthly observation for (f, rf). In this way, we preserve the relation between r_i, the risk-free rate and the vector of factors for each fund. This continues until the time-series of the proper length has been filled out for a particular fund, resulting in an artificial sample with no missing values. We then form a "Hole Matrix," H, which is the same size as the original fund sample, and which contains zeros where the original fund data are missing and ones elsewhere. We apply the H matrix to assign missing values for the same months in which they appear in the original data for each fund. We estimate the alphas treating this simulated data the same way we treat the original data and the baseline simulation data.

The simulations using the H matrix guarantee that each fund has the same number of missing values in each simulation draw, appearing at each point in "time" as it does in the original data. The artificial data for each fund should replicate the statistical properties of the original data, on the assumption of independence over time. This approach also preserves cross-sectional dependence and conditional heteroscedasticity to some extent, through the common dependence of the fund returns on the factors.

We compare the results of this approach with that of our baseline simulation method in Table A.9. Simulations are conducted under the null hypothesis that all of the alphas are zero. Baseline Sims. refers to the simulation method used in the main paper. Hole-preserving Sims is based on the alternative simulation methodology. Each case has the same number of fund and time-series observations as in the original data. In Panel A the statistics are drawn from the first simulation trial. Mean is the average across the mutual funds, Min is the minimum and Max is the maximum. A fund is required to have at least 8 observations in the simulated samples to be included in the summary statistics.

Panel A of Table A.9 shows that the Baseline and Hole-preserving simulations deliver similar statistical properties for funds' residual standard deviations and factor model R-squares. Either method closely reproduces the statistical properties of the original data.

In panel B of Table A.9, the values at each fractile of the cross-sectional distributions of mutual funds' alphas and alpha t-ratios are shown for the two simulation methods. These are the averages across 100 simulation trials, thus estimating the expected outcomes. Alpha is the average alpha value at each fractile of the cross-sectional fund distribution of alphas. T-ratio is the heteroscedasticity-consistent t-ratio for alpha at each fractile of the cross-section of alpha t-ratios. Alphas are in percent per month. The results show that the cross-sectional distributions of alphas and alpha t-ratios that is produced by the two simulation methods are very similar.

J. Analysis of the Impact of Variation in the δ and β Parameters on the Standard Errors

Our standard error calculations assume that the δ and β parameters are estimated without error. As the number of simulation trials gets large, these errors should be negligible, but it is useful to evaluate their impact when we use 1,000 simulation trials.

Consider the estimated fractions $\pi = \pi(\phi,F)$ as a function of $\phi=(\delta_g, \delta_b, \beta_g, \beta_b)$ and the fractions rejected, F. Note that we report and use the average values of the ϕ parameters over 1,000 simulation trials, so it is the variance of the mean value that we are concerned with here. Our current estimate for the variance of π may be written as Q Vf Q′, where the $Q = Q(\phi)$ is defined by Equation (A.4). Vf is the variance matrix of the fractions rejected, F. To consider the impact of variation in the ϕ parameters, we expand $\pi = \pi(\phi,F)$ using the delta method and assume that the covariance matrix of (ϕ,F) is block diagonal. The standard errors in the main paper include the part due to Vf, but not the part due to $V(\phi)$. Equivalently, we can assume that our current estimate is the expected value of the conditional variance of π given ϕ, and that we are missing the variance of the conditional mean, taken with respect to the variation in ϕ. Either approach leads to the same missing term due to variation in ϕ:

$$E(\partial\pi/\partial\phi)\ V(\phi^{\wedge}-\phi)\ E(\partial\pi/\partial\phi)' \qquad \text{(A.12)}$$

The analytical derivatives are evaluated at the average parameter values across the 1,000 simulation trials, and the variance matrix $V(\phi\hat{}-\phi)$ is estimated from the covariances of the (δ,β) estimates across the 1,000 trials. These covariances are estimated assuming independent simulation trials.

The results for hedge fund data and 1,000 simulation trials, with $(\gamma/2)=0.10$, and evaluated at the best fitting alpha values are:

$Q\ Vf\ Q' =$

$$\begin{matrix} 0.0327520 & -0.0175781 \\ -0.0175781 & 0.0488524 \end{matrix}$$

$E(\partial\pi/\partial\phi)\ V(\phi\hat{}-\phi)\ E(\partial\pi/\partial\phi)' =$

$$\begin{matrix} 0.000831679 & 0.000259822 \\ 0.000259822 & 9.87724e\text{-}005. \end{matrix}$$

We conclude that the variation in the δ and β parameters has a trivial impact on our results. Even if the covariance of (ϕ,F) is not block diagonal, the cross terms should be small compared with the included $Q\ Vf\ Q'$ terms.

K. Alternative Goodness-of-fit Measures

The alternative goodness-of-fit measures are the two-sample, Kolmogorov-Smirnov distance: $KS = \text{Sup}_x \mid F_1(x) - F_2(x) \mid$, where $F_1(.)$ and $F_2(.)$ are the two empirical cumulative distribution functions (CDFs) of alpha t-ratios; one from the data and one from the model. This measure looks at the maximum vertical distance between the CDFs. The second alternative measure is the Cramer-von Mises distance: $CvM = E_x \{[F_1(x) - F_2(x)]^2\}$, which looks at the mean squared distance. To implement the alternative measures we combine the observations of the alpha t-ratios from the original data and from a model, rank the values and calculate the two CDFs at each of the discrete points.

Table A.10 presents the analysis. The first row replicates the Pearson Chi-square case, because these exercises are based on 100 simulation trials at each point in the grid, whereas the results in the main text use 1,000 trials. The difference between 100 and 1,000 here is a maximum difference of 2 basis points per month in the alpha parameters. The difference across the goodness-of-fit measures is less than 8 basis points per month in the alpha parameters, 7% in the power parameters, less than one percent in the confusions and within about one standard error on the π parameters. While the alternative measures find a few more zero-alpha hedge funds, the results are otherwise very similar. We conclude that results are robust to the use of the different goodness-of-fit measures.

L. Appendix Tables

Table A.1: Summary Statistics with Alternative Benchmarks

Monthly returns are summarized for hedge funds, stated in monthly percentage units. The values at the cutoff points for various fractiles of the cross-sectional distributions of the sample of funds are reported. Each column is sorted on the statistic shown. Nobs is the number of available monthly returns, where a minimum of 12 are required. Mean is the sample mean return, Std is the sample standard deviation of return, reported as monthly percentages, and Rho1 is the first order sample autocorrelation in raw units. The alpha (α) estimates and their T-ratios are based on OLS regressions using the Fama-French three factors (denoted 3-fac) or the Fung and Hsieh seven factors (denoted 7-fac) and heteroskedasticity-consistent standard errors for the hedge funds. The t-ratios denoted 7-fac SW use Scholes Williams betas, and increase the minimum number of required observations to 19.

Hedge Fund Returns: January 1994–March 2012 (219 months)

		Monthly Returns (%)			α Estimates (%)		α T-ratios		
Fractile	Nobs	Mean	Std	Rho1	3-fac	7-fac	3-fac	7-fac	7-fac SW
0.010	208	2.52	14.31	0.61	2.147	2.200	6.395	6.787	7.326
0.050	154	1.46	8.57	0.50	1.042	1.140	3.369	3.380	3.932
0.100	125	1.15	6.65	0.43	0.736	0.811	2.485	2.959	2.919
0.250	83	0.75	4.21	0.30	0.342	0.431	1.287	1.598	1.682
Median	48	0.38	2.63	0.17	0.030	0.119	0.110	0.456	0.504
0.750	28	-0.01	1.74	0.03	-0.307	-0.186	-0.931	-0.605	-0.668
0.900	18	-0.56	1.23	-0.11	-0.813	-0.672	-1.850	-1.642	-1.806
0.950	15	-1.07	0.97	-0.20	-1.296	-1.079	-2.436	-2.248	-2.486
0.990	12	-2.61	0.58	-0.33	-2.626	-2.786	-4.031	-3.742	-4.012

Table A.2: Estimated Fractions of Funds with Various Alpha Values

The fractions of managers in the population with specified values of zero, good or bad alphas are estimated using simulation. The symbol π_g denotes the estimated fraction of good funds and π_0 denotes the fraction of zero-alpha funds. Alphas are stated in monthly percentage units. The power parameters of the test are β_g, the power to reject against the alternative of a good fund, and β_b, the power to reject against the alternative of a bad fund. The confusion parameters are δ_b, the probability of finding a good fund when it is bad, and δ_g, the probability of finding a bad fund when it is good. All fractions except for the test sizes are stated as (monthly, for the alphas) percentages. Empirical standard errors for the pie fractions are indicated in parentheses, except when a constraint is binding (na). * denotes that the test sizes for the classical FDR estimator is 30% in each tail, and the size is 10% for our estimators. In Panel A, the sample period for mutual funds is January 1984–December 2011 (336 months). In Panel B, the sample period for hedge funds is January 1994–March 2012 (219 months).

Panel A: Mutual Funds

Monthly Alphas (%)		**Size**	**FDR Calcs:**		**Powers**		**Confusions**		**Fractions**	
Good α_g	**Bad α_b**	**$\gamma/2$**	**π_0**	**π_g**	**β_g**	**β_b**	**δ_g**	**δ_b**	**π_0**	**π_g**
Estimates for Given Test Sizes and Various Alpha Values										
-0.084	-0.104	*	72.6	-4.6	4.1	29.5	3.3	25.6	0.0	0.0
									(na)	(na)
0.031	-0.234	*	72.4	-4.5	14.3	57.9	1.1	7.3	63.2	0.0
									(34.1)	(na)
0.170	-0.415	*	71.7	-4.3	45.2	81.5	0.4	1.8	75.9	0.0
									(6.1)	(na)
0.270	-0.573	*	72.7	-4.5	64.6	91.1	0.2	0.9	77.7	0.0
									(4.1)	(na)

Table A.2, page 2

Alphas (%)		**Fit**	**Size**	**Powers**		**Confusions**		**Fractions**	
Good α_g	**Bad α_b**		**$\gamma/2$**	**β_g**	**β_b**	**δ_g**	**δ_b**	**π_0**	**π_g**
Joint Estimation of Alphas and Fractions in the Populations									
Unconstrained Alpha Domains, 3-Group Model:									
-0.087	-0.305	2509	0.05	2.1	58.7	0.4	15.6	38.9	4.4
								(14.6)	(22.1)
-0.034	-0.204	2155	0.10	6.9	52.4	1.5	15.4	50.7	6.9
								(26.7)	(37.2)
Constrained Alpha Domains ($\alpha_g \geq 0$, $\alpha_b \leq 0$), 3-Group Model									
0.004	-0.162	3925	0.05	5.6	30.0	1.0	4.8	33.4	12.2
								(28.3)	(35.3)
0.001	-0.173	3816	0.10	10.3	45.7	1.8	10.3	0.0	55.2
								(na)	(44.5)
2-Group Model:									
0.00	-0.139	4694	0.05	na	24.0	na	na	29.3	0.0
								(23.2)	(na)
0.0	-0.172	3862	0.10	na	45.5	na	na	48.4	0.0
								(37.8)	(na)
Single-Alpha Model									
na	-0.220	3431	0.05	na	na	na	na	0.0	0.0
								(na)	(na)
na	-0.205	3401	0.10	na	na	na	na	0.0	0.0
								(na)	(na)

Table A.2, page 3

Panel B: Hedge Funds

Monthly Alphas (%)		**Size**	**FDR Calcs:**		**Powers**		**Confusions**		**Fractions**	
Good α_g	**Bad α_b**	**$\gamma/2$**	**π_0**	**π_g**	**β_g**	**β_b**	**δ_g**	**δ_b**	**π_0**	**π_g**
Estimates for Given Test Sizes and Various Alpha Values										
0.040	0.020	*	75.8	22.3	12.8	9.4	11.6	8.7	0.0	100
									(na)	(na)
0.342	-0.307	*	76.6	21.8	40.4	39.2	4.2	4.1	53.3	39.8
									(18.8)	(13.7)
0.736	-0.813	*	75.0	22.6	68.2	73.5	2.1	2.4	77.6	20.6
									(9.1)	(7.2)
1.042	-1.296	*	76.5	22.1	80.0	85.8	1.4	1.8	81.5	17.1
									(6.4)	(4.7)

Table A.2, page 4

Alphas (%)		**Fit**	**Size**	**Power**		**Confusions**		**Fractions**	
Good α_g	**Bad α_b**		**$\gamma/2$**	**β_g**	**β_b**	**δ_g**	**δ_b**	**π_0**	**π_g**
Joint Estimation of Alphas and Fractions in the Populations									
Unconstrained Alpha Domains, 3-Group Model:									
0.237	-0.098	2633	0.05	15.9	7.6	4.0	3.1	0.0	51.7
								(na)	(24.5)
0.252	-0.108	2225	0.10	31.4	17.4	7.0	5.0	0.0	53.2
								(na)	(16.8)
2-Group Model:									
0.287	na	4073	0.05	18.2	na	na	na	45.5	54.5
								(32.9)	(37.1)
0.252	na	4807	0.10	31.9	na	na	na	44.6	55.4
								(28.2)	(33.5)
Single-Alpha Model									
0.452	na	3160	0.05	na	na	na	na	0.0	100.0
								(na)	(na)
0.434	na	3120	0.10	na	na	na	na	0.0	100.0
								(na)	(na)

Table A.3: Estimated Fractions of Hedge Funds with Specified Values of Nonzero Alphas: Using the Fama and French Factors

The fractions of managers in the population with specified values of alphas, which are either zero, good or bad, are estimated using simulation as described in the text. Π_g denotes the estimated fraction of good funds, Π_b denotes the fraction of bad funds and Π_0 denotes the fraction of zero-alpha funds. All alphas are stated in monthly percentage units. The power parameters of the test are denoted as β_g, the power to reject against the alternative of a good fund, and β_b, the power to reject against the alternative of a bad fund. The confusion parameters are denoted as δ_b, the probability of finding a good fund when it is bad, and δ_g, the probability of finding a bad fund when it is good. All fractions are stated as percentages. The reported asymptotic standard errors are in parentheses. These are not applicable when a constraint is binding (na). The sample period for hedge funds is January 1994–March 2012 (219 months).

Monthly Alphas (%)			**FDR Calcs:**		**Power**		**Confusions**		**Fractions**	
Good α_g	**Bad α_b**	**$\gamma/2$**	**Π_0**	**Π_g**	**β_g**	**β_b**	**δ_g**	**δ_b**	**Π_0**	**Π_g**
0.317	-0.267	0.025	92.2	6.1	17.5	14.2	1.2	1.1	41.3	40.7
									(7.5)	(5.2)
0.317	-0.267	0.05	85.8	9.6	29.5	25.2	2.0	1.8	35.1	39.6
									(15.7)	(9.8)
0.317	-0.267	0.20	71.8	17.2	58.4	54.3	6.6	5.8	25.7	41.6
									(38.7)	(16.0)
0.317	-0.267	0.30	66.4	20.0	68.0	64.5	10.6	9.3	17.5	45.1
									(60.0)	(17.1)
0.317	-0.267	0.40	69.3	18.4	76.0	72.5	15.4	13.1	28.5	39.2
									(121)	(17.6)
0.040	0.020	0.05	85.9	9.6	6.9	4.8	6.1	4.2	0.0	100
									(na)	(na)
0.342	-0.307	0.05	86.4	9.6	31.8	28.6	1.8	1.7	45.2	35.5
									(13.7)	(9.3)
0.736	-0.813	0.05	85.9	9.6	62.2	68.2	0.8	0.9	76.9	16.1
									(6.3)	(4.9)
1.042	-1.296	0.05	85.8	9.6	75.1	83.1	0.5	0.7	81.2	13.1
									(6.3)	(4.8)

Table A.4: Simultaneous Estimation of Alpha Parameters in Three-Distribution Models

The true good and bad alpha parameters, $\boldsymbol{\alpha_g}$ and $\boldsymbol{\alpha_b}$, are estimated for each test size $\boldsymbol{\gamma/2}$, simultaneously with the fractions of managers in the population having the specified alphas. A grid search over the true good and bad alpha parameters looks from the median alpha value in the data from Table 1, to the upper or lower 5% tail values in the data, with a grid size of 0.001% for mutual funds and 0.005% for hedge funds. The best fitting good and bad alpha parameters are shown, where the fit is determined by minimizing the difference between the cross section of fund alphas estimated in the actual data, versus the cross section of alphas estimated from a mixture of return distributions, formed from the good and bad alpha parameters and the estimated Π fractions for each of the three types. At each point in the grid, the Π fractions are estimated using simulation as described in the text, with 100 simulation trials. Π_g denotes the fraction of good funds and Π_b denotes the fraction of bad funds. All alpha parameters are stated in monthly percentage units. The power parameters of the test are denoted as β_g, the power to reject against the alternative of a good fund, and β_b, the power to reject against the alternative of a bad fund. The confusion parameters are denoted as δ_b, the probability of finding a good fund when it is bad, and δ_g, the probability of finding a bad fund when it is good. In Panels C-E, the alphas use 100 trials in the grid search, but the other parameter estimates and the standard errors, conditioning on those alphas, use 1,000 trials.

$\gamma/2$	**Good α_g**	**Bad α_b**	**β_g**	**β_b**	**δ_g**	**δ_b**	**Π_g**	**Π_b**
Panel A: Mutual Funds: January 1984–December 2011 (336 months)								
0.005	-0.074	-0.184	0.001	0.001	0.001	0.001	0.000	1.000
0.010	-0.074	-0.184	0.006	0.134	0.003	0.030	0.287	0.351
0.025	-0.034	-0.184	0.016	0.254	0.006	0.045	0.312	0.426
0.050	-0.087	-0.305	0.021	0.587	0.004	0.156	0.044	0.567
0.100	-0.034	-0.204	0.069	0.524	0.015	0.154	0.069	0.424
0.200	-0.064	-0.224	0.051	0.554	0.012	0.198	0.190	0.349
0.300	-0.064	-0.234	0.174	0.798	0.043	0.476	0.165	0.411
0.400	-0.094	-0.334	0.210	0.928	0.033	0.632	0.280	0.296
Panel B: Hedge Funds: January 1994–March 2012 (219 months)								
0.005	0.260	-0.220	0.002	0.002	0.001	0.001	0.000	1.000
0.010	0.280	-0.030	0.056	0.009	0.011	0.006	0.341	0.266
0.025	0.380	0.030	0.219	0.225	0.030	0.010	0.502	0.000
0.050	0.237	-0.098	0.159	0.076	0.004	0.031	0.517	0.483
0.100	0.252	-0.108	0.314	0.174	0.070	0.050	0.532	0.468
0.200	0.230	-0.200	0.469	0.495	0.076	0.082	0.564	0.436
0.300	0.167	-0.123	0.508	0.449	0.195	0.163	0.652	0.348
0.400	0.230	-0.220	0.702	0.671	0.191	0.167	0.517	0.441

Table A.4, page 2

Alphas (%)		Fit	Size	BSW Calcs:		Power		Confusions		Fractions	
Good α_g	Bad α_b		$\gamma/2$	Π_0	Π_g	β_g	β_b	δ_g	δ_b	Π_0	Π_g
Panel C: Hedge Funds, Baseline Case, Unconstrained Alpha Domains, 3-Group Model:											
0.237	-0.098	2633	0.05	90.9	8.2	15.9	7.6	4.0	3.1	0.0	51.7
										(na)	(24.5)
0.252	-0.108	2225	0.10	84.1	13.4	31.4	17.4	7.0	5.0	0.0	53.2
										(na)	(16.8)
Panel D: Results When Alphas are Associated with Active Management											
0.272	-0.373	1068	0.05	92.1	8.3	18.4	26.7	2.4	2.6	29.7	58.9
										(23.3)	(23.8)
0.282	-0.308	2081	0.10	85.8	13.2	35.0	39.0	4.2	4.5	43.8	51.4
										(17.3)	(17.5)
Panel E: Results Accounting for Error Variance in Estimated Alphas											
0.209	-0.083	1596	0.05	95.2	6.1	19.3	7.5	3.6	2.6	12.4	58.6
										(28.1)	(25.6)
0.292	-0.038	1413	0.10	84.2	13.1	36.7	12.0	8.8	4.3	0.02	28.2
										(29.0)	(13.1)

Table A.5: Simultaneous Estimation of True Alpha Parameters in the Two-Distribution Model for Mutual Funds

The true bad alpha parameters, $\boldsymbol{\alpha_b}$, are estimated for each test size $\boldsymbol{\gamma}$**,** simultaneously with the fractions of managers in the population having zero or bad alphas. A grid search over the bad alpha parameters looks from the median alpha value in the data from Table 1, to the lower 5% tail values in the data, with a grid size of 0.001%. The best fitting bad alpha parameters are shown, where the fit is determined by minimizing the difference between the cross section of fund alphas estimated in the actual data, versus the cross section of alphas estimated from a mixture of return distributions, formed from the zero and bad alpha parameters and the estimated Π fractions for each of the types. At each point in the grid, the Π fractions are estimated using simulation as described in the text, with 100 simulation trials. Π_b denotes the fraction of bad funds. All alpha parameters are stated in monthly percentage units. The power parameters of the test are denoted as β_b, the power to reject against the alternative of a bad fund.

γ	Bad α_b	β_b	Π_b	Π_0
0.010	-0.136	0.073	1.000	0.000
0.025	-0.141	0.173	1.000	0.000
0.050	-0.156	0.296	1.000	0.000
0.060	-0.142	0.292	1.000	0.000
0.070	-0.158	0.353	1.000	0.000
0.100	-0.148	0.401	1.000	0.000
0.150	-0.158	0.507	1.000	0.000
0.200	-0.162	0.573	1.000	0.000
0.300	-0.187	0.723	1.000	0.000
0.400	-0.203	0.824	1.000	0.000

Table A.6: Finite Sample Properties of the Estimators

In each of bootstrap simulation trials artificial data are drawn randomly from a mixture of three fund distributions. The "population" values of the fractions of funds in each group, π, shown here in the first row, determine the mixture, combined with the good, zero or bad alpha values that we estimate as the best-fitting values for the full sample period. Hedge fund data for January 1994–March 2012 are used, and the values of the bad and good alphas are -0.138 and 0.262% per month. For each simulation draw we run the estimation by simulation with 1000 trials, to generate the parameter and standard error estimates. Standard error estimates are removed when an estimated fraction is on the boundary of a constraint. The Avg. estimates are the averages over the 1000 draws from the mixture distribution. The empirical SD's are the standard deviations taken across all of the 1000 trials. The Root MSE's are the square root of the average over the 1000 trials, of the squared difference between an estimated and true parameter value. $\gamma/2$ indicates the size of the tests (the area in one tail of the two-tailed tests). When the sample size is T=5,000, we use only 100 simulation trials.

Panel A: $\gamma/2 = 0.05$

	π_{zero}	π_{good}	π_{bad}
Population values	1/3	1/3	1/3
Our Avg. Estimates	0.548	0.197	0.255
Classical FDR Avg. Estimates	0.980	0.024	-0.003
Empirical SDs	0.290	0.186	0.261
Avg. Reported SDs	0.054	0.357	0.330
Root MSE	0.361	0.231	0.272
Classical Empirical SD	0.029	0.027	0.015
Classical Avg. Reported SD	0.006	0.007	0.011
Classical Root MSE	0.647	0.311	0.337

Panel B: $\gamma/2 = 0.10$

	π_{zero}	π_{good}	π_{bad}
Population values	1/3	1/3	1/3
Our Avg. Estimates	0.417	0.286	0.298
Classical FDR Avg. Estimates	0.910	0.069	-0.020
Empirical SDs	0.233	0.161	0.227
Avg. Reported SDs	0.241	0.354	0.295
Root MSE	0.247	0.168	0.230
Classical Empirical SD	0.048	0.047	0.036
Classical Avg. Reported SD	0.009	0.011	0.017
Classical Root MSE	0.599	0.270	0.315

Table A.6, page 2

Panel C: $\gamma/2 = 0.20$

	π_{zero}	π_{good}	π_{bad}
Population values	1/3	1/3	1/3
Our Avg. Estimates	0.380	0.298	0.322
Classical FDR Avg. Estimates	0.857	0.097	0.056
Empirical SDs	0.230	0.145	0.237
Avg. Reported SDs	0.528	0.406	0.360
Root MSE	0.235	0.149	0.237
Classical Empirical SD	0.061	0.063	0.060
Classical Avg. Reported SD	0.014	0.015	0.023
Classical Root MSE	0.527	0.244	0.293

Panel D: $\gamma/2 = 0.30$

	π_{zero}	π_{good}	π_{bad}
Population values	1/3	1/3	1/3
Our Avg. Estimates	0.368	0.308	0.324
Classical FDR Avg. Estimates	0.830	0.115	0.056
Empirical SDs	0.233	0.146	0.249
Avg. Reported SDs	0.772	0.461	0.488
Root MSE	0.235	0.148	0.249
Classical Empirical SD	0.069	0.073	0.072
Classical Avg. Reported SD	0.020	0.020	0.030
Classical Root MSE	0.500	0.231	0.287

Panel E: $\gamma/2 = 0.05$

	π_{zero}	π_{good}	π_{bad}
Population values	0.750	0.010	0.240
Our Avg. Estimates	0.829	0.027	0.143
Classical FDR Avg. Estimates	1.010	-0.010	-0.003
Empirical SDs	0.285	0.085	0.253
Avg. Reported SDs	0.054	0.319	0.282
Root MSE	0.293	0.087	0.271
Classical Empirical SD	0.024	0.017	0.017
Classical Avg. Reported SD	0.005	0.006	0.009
Classical Root MSE	0.264	0.026	0.244

Table A.6, page 3

Panel F: $\gamma/2 = 0.10$

	π_{zero}	π_{good}	π_{bad}
Population values	0.750	0.010	0.240
Our Avg. Estimates	0.706	0.046	0.248
Classical FDR Avg. Estimates	0.988	-0.008	0.020
Empirical SDs	0.301	0.097	0.274
Avg. Reported SDs	0.284	0.370	0.267
Root MSE	0.304	0.104	0.274
Classical Empirical SD	0.053	0.038	0.044
Classical Avg. Reported SD	0.008	0.010	0.015
Classical Root MSE	0.244	0.042	0.224

Panel G: $\gamma/2 = 0.20$

	π_{zero}	π_{good}	π_{bad}
Population values	0.750	0.010	0.240
Our Avg. Estimates	0.609	0.068	0.322
Classical FDR Avg. Estimates	0.959	-0.007	0.048
Empirical SDs	0.314	0.118	0.294
Avg. Reported SDs	0.693	0.499	0.342
Root MSE	0.344	0.132	0.306
Classical Empirical SD	0.083	0.066	0.076
Classical Avg. Reported SD	0.014	0.015	0.022
Classical Root MSE	0.225	0.068	0.206

Panel H: $\gamma/2 = 0.30$

	π_{zero}	π_{good}	π_{bad}
Population values	0.750	0.010	0.240
Our Avg. Estimates	0.574	0.083	0.333
Classical FDR Avg. Estimates	0.946	-0.006	0.061
Empirical SDs	0.311	0.124	0.301
Avg. Reported SDs	1.084	0.628	0.560
Root MSE	0.357	0.144	0.318
Classical Empirical SD	0.100	0.084	0.095
Classical Avg. Reported SD	0.021	0.021	0.030
Classical Root MSE	0.220	0.086	0.203

Table A.6, page 4

Panel I: Maximizing The Use of Asymptotic Normality

	π_{zero}	π_{good}	π_{bad}
Population values	0.100	0.600	0.300
$\gamma/2 = 0.05$			
Our Avg. Estimates	0.382	0.404	0.214
Empirical SDs	0.271	0.259	0.230
Avg. Reported SDs	0.058	0.416	0.426
Root MSE	0.390	0.323	0.244
$\gamma/2 = 0.10$			
Our Avg. Estimates	0.294	0.512	0.194
Empirical SDs	0.212	0.194	0.163
Avg. Reported SDs	0.237	0.312	0.406
Root MSE	0.286	0.212	0.193
$\gamma/2 = 0.20$			
Our Avg. Estimates	0.186	0.548	0.265
Empirical SDs	0.165	0.165	0.174
Avg. Reported SDs	0.537	0.288	0.523
Root MSE	0.186	0.172	0.177
$\gamma/2 = 0.30$			
Our Avg. Estimates	0.198	0.545	0.247
Empirical SDs	0.160	0.145	0.151
Avg. Reported SDs	0.775	0.332	0.643
Root MSE	0.187	0.151	0.159
$\gamma/2 = 0.30$, USING SAMPLES OF SIZE T=5,000:			
Our Avg. Estimates	0.082	0.595	0.323
Empirical SDs	0.018	0.016	0.025
Avg. Reported SDs	0.054		0.058
Root MSE	0.026	0.016	0.034

Table A.6, page 5

Panel J: $\gamma/2$ = 0.05, USING SAMPLES OF SIZE T=5,000

	π_{zero}	π_{good}	π_{bad}
Population values	0.100	0.600	0.300
Our Avg. Estimates	0.075	0.595	0.329
Classical FDR Avg. Estimates	0.267	0.525	0.208
Empirical SDs	0.016	0.013	0.023
Avg. Reported SDs	0.089	0.014	0.101
Root MSE	0.030	0.014	0.038
Classical Empirical SD	0.009	0.012	0.015
Classical Avg. Reported SD	0.008	0.012	0.019
Classical Root MSE	0.167	0.076	0.093

Panel K: $\gamma/2$ = 0.10, USING SAMPLES OF SIZE T=5,000

	π_{zero}	π_{good}	π_{bad}
Population values	0.100	0.600	0.300
Our Avg. Estimates	0.078	0.599	0.322
Classical FDR Avg. Estimates	0.257	0.521	0.222
Empirical SDs	0.012	0.015	0.020
Avg. Reported SDs	0.059	0.021	0.079
Root MSE	0.025	0.015	0.030
Classical Empirical SD	0.009	0.014	0.015
Classical Avg. Reported SD	0.009	0.012	0.019
Classical Root MSE	0.157	0.080	0.079

Panel L: $\gamma/2$ = 0.05, USING SAMPLES OF SIZE T=5,000

	π_{zero}	π_{good}	π_{bad}
Population values	0.333	0.333	0.333
Our Avg. Estimates	0.199	0.345	0.456
Classical FDR Avg. Estimates	0.490	0.286	0.223
Empirical SDs	0.049	0.020	0.030
Avg. Reported SDs	0.256	0.362	0.413
Root MSE	0.143	0.023	0.126
Classical Empirical SD	0.016	0.017	0.022
Classical Avg. Reported SD	0.009	0.012	0.020
Classical Root MSE	0.158	0.050	0.112

Table A.6, page 6

Panel M: $\gamma/2$ = 0.05, SAMPLES OF SIZE T=5,000, MAXIMIZING USE OF ASYMPTOTIC NORMALITY ASSUMPTION

	π_{zero}	π_{good}	π_{bad}
Population values	0.100	0.600	0.300
Our Avg. Estimates	0.074	0.599	0.328
Classical FDR Avg. Estimates	0.291	0.503	0.206
Empirical SDs	0.013	0.018	0.025
Avg. Reported SDs	0.062	0.038	0.098
Root MSE	0.029	0.018	0.038
Classical Empirical SD	0.009	0.016	0.038
Classical Avg. Reported SD	0.008	0.011	0.019
Classical Root MSE	0.192	0.098	0.096

Panel N: $\gamma/2$ = 0.30, SAMPLES OF SIZE T=5,000, MAXIMIZING USE OF ASYMPTOTIC NORMALITY ASSUMPTION

	π_{zero}	π_{good}	π_{bad}
Population values	0.100	0.600	0.300
Our Avg. Estimates	0.082	0.595	0.323
Classical FDR Avg. Estimates	0.213	0.538	0.249
Empirical SDs	0.018	0.016	0.025
Avg. Reported SDs	0.054	0.038	0.058
Root MSE	0.026	0.016	0.034
Classical Empirical SD	0.014	0.016	0.021
Classical Avg. Reported SD	0.012	0.014	0.024
Classical Root MSE	0.114	0.064	0.055

Table A.7: Holding Period Returns after Fund Selection with FDR Control

A 60-month rolling formation period is used to select good and bad funds, controlling the false discovery rate at 30% in each tail. (γ/2) is the average, over the formation periods, of the significance levels that best control the false discovery rates during the formation periods. Good is the portfolio of funds found to have high alphas, Bad is the portfolio of low alpha funds and G-B is the difference between the Good and Bad alpha fund returns in the months when both exist. N is the average number of funds in each group, μ is the sample mean portfolio return and α is the portfolio alpha, formed using the Fama-French factors for mutual funds and the Fung and Hsieh factors for hedge funds during the holding period. The holding period is one year in length and follows the formation period by one to four years. Mean returns and alphas are percent per month. T is the heteroskedasticity-consistent t-ratio.

Portfolio			Year after Formation Period:												
				1			2			3			4		
	(γ/2)	N	μ	α	T	μ	α	T	μ	α	T	μ	α	T
Panel A: Hedge Funds during January1999–March 2012															
Good	0.11	360.6	0.34	0.23	2.9	0.26	0.22	2.7	0.23	0.13	1.4	0.30	0.19	2.1
Zero		193.7	0.26	0.17	2.1	0.21	0.12	1.3	0.31	0.17	1.8	0.26	0.12	1.0
Bad	0.26	76.6	0.19	0.10	1.1	0.15	0.08	0.7	0.29	0.14	1.4	0.41	0.27	2.3
G-B			0.15	0.13	1.8	0.11	0.14	1.5	-0.06	-0.02	-0.2	-0.10	-0.10	-1.1
Panel B: Mutual Funds during January 1989–December 2011															
Good	0.10	236.6	0.53	-0.02	-0.5	0.44	-0.09	-1.8	0.51	-0.11	-1.9	0.49	-0.04	-0.8
Zero		398.8	0.49	-0.06	-1.8	0.45	-0.07	-1.8	0.52	-0.09	-2.5	0.44	-0.09	-2.4
Bad	0.19	382.4	0.48	-0.10	-2.4	0.49	-0.04	-1.0	0.54	-0.07	-2.0	0.42	-0.11	-2.8
G-B			0.06	0.08	1.4	-0.05	-0.05	-0.8	-0.03	-0.04	-0.8	0.07	0.07	-1.4

Table A.8: Holding Period Returns by Fund Groups after Rolling Estimation

A 60-month rolling formation period is used to select good, zero alpha and bad funds, in which the basic probability model used to estimate the π fractions when the size of the tests is 10% in each tail. The jointly estimated good and bad alpha values are those depicted in figures 3 and 4 of the main paper. Good is the equal weighted portfolio of funds detected to have positive alphas during the formation period, zero is a portfolio of zero alpha funds, Bad is the portfolio of low alpha funds and G-B is the difference between the two, in months where both have nonzero numbers of funds. The groups are formed by ranking funds on their formation period alphas and assigning their group membership according to the π fractions estimated for that formation period. N is the average number of funds over all of the formation periods, μ is the sample mean portfolio return during the holding periods, and α is the portfolio alpha, formed using the Fama-French factors for mutual funds and the Fung and Hsieh factors for hedge funds during the holding period. The holding period is one year in length and follows the formation period by one to four years. Mean returns and alphas are percent per month. T is the heteroskedasticity-consistent t-ratio.

Portfolio		Year after Formation Period:											
			1			2			3			4	
	N	μ	α	T	μ	α	T	μ	α	T	μ	α	T
Panel A: Hedge Funds during January1999–March 2012													
Good	339.3	0.32	0.20	2.5	0.22	0.18	2.1	0.25	0.16	1.9	0.29	0.17	2.0
Zero	263.6	0.30	0.21	2.6	0.22	0.13	1.5	0.28	0.15	1.7	0.31	0.18	1.7
Bad	27.8	0.21	-0.01	-0.1	0.09	0.00	0.01	0.15	0.07	0.3	0.35	0.12	0.6
G-B		0.31	0.33	2.1	0.24	0.26	1.3	-0.01	0.03	0.1	-0.02	0.10	0.5
Panel B: Mutual Funds during January 1989–December 2011													
Good	444	0.56	0.02	-0.3	0.41	-0.11	-2.4	0.49	-0.13	-2.2	0.48	-0.05	-1.0
Zero	294.7	0.52	-0.06	-1.6	0.84	-0.06	-1.7	0.29	-0.11	-2.8	0.61	-0.11	-2.5
Bad	277.5	0.43	-0.10	-2.2	0.51	-0.05	-1.2	0.53	-0.08	-2.2	0.42	-0.11	-2.6
G-B		0.08	0.12	1.8	-0.08	-0.06	-0.1	-0.04	-0.05	-0.8	0.06	0.06	1.2

Table A.9: Impact of Missing Value Patterns on the Simulations

Baseline Sims. refers to the simulation method used in the main paper. Hole-preserving Sims is based on an alternative simulation methodology that exactly reproduces the number of missing values and their location in time for each mutual fund. In Panel A the statistics are drawn from the first simulation trial. Mean is the average across the mutual funds, Min is the minimum and Max is the maximum, requiring at least 8 observations during the January 1984–December 2011 sample period (336 months). In panel B the values at each fractile of the distribution are the averages across 100 simulation trials. Alpha is the average alpha value at each fractile of the cross-sectional fund distribution of alphas. T-ratio is the heteroscedasticity-consistent t-ratio for alpha at each fractile of the cross-section of alpha t-ratios. Alphas are in percent per month.

Panel A: Mutual Fund Statistics in Original Data and two Simulation Methods

Residual Standard Deviations:	Actual Data	Baseline Sims.	Hole-preserving Sims
Mean	1.55	1.46	1.48
Min	0.24	0.00	0.00
Max	21.02	16.58	13.20
Factor Model R-squares			
Mean	90.6	90.8	90.0
Min	3.6	3.6	4.5
Max	99.8	100	100

Panel B: Average Values at Various Fractiles

Fraction of	Baseline Sims.		Hole-preserving Sims.	
funds above	Alphas	T-ratio	Alphas	T-ratio
0.010	0.673220	2.83576	0.669183	2.75717
0.050	0.344044	1.83518	0.336002	1.81665
0.100	0.239758	1.41382	0.230975	1.38223
0.250	0.111415	0.75920	0.104421	0.71177
0.500	0.006508	0.04764	-0.000549	-0.00393
0.750	-0.096898	-0.65971	-0.105605	-0.72110
0.900	-0.222724	-1.32077	-0.231732	-1.39057
0.950	-0.326849	-1.73972	-0.337738	-1.82626
0.990	-0.654136	-2.71572	-0.668404	-2.77866

Table A.10: Joint Estimation of Alphas and Fractions in the Populations with Alternative Goodness of-Fit Measures

All tests are conducted using a size of γ/2 = 0.10. Simulations use 100 trials at each point in the grid search of the good and bad alpha values. Given those alpha values, simulations use 1,000 trials to estimate the other model parameters. The standard errors in parentheses are the asymptotic standard errors. Pearson denotes the Pearson Chi-square goodness-of-fit measure, KS denotes the Kolmogorov-Smirnov distance and CvM denotes the Cramer-von Mises distance.

Alphas (%)		**Measure**	**Power**		**Confusions**		**Fractions**	
Good α_g	**Bad α_b**		**β_g**	**β_b**	**δ_g**	**δ_b**	**π_0**	**π_g**
Unconstrained Alpha Domains, 3-Group Model:								
0.237	-0.128	Pearson	30.7	19.5	6.6	4.9	7.3	63.2
							(20.6)	(20.1)
0.227	-0.148	KS	29.7	21.7	6.2	5.1	10.0	65.9
							(20.3)	(21.3)
0.307	-0.113	CvM	37.5	18.3	6.9	4.3	26.5	46.0
							(19.9)	(14.3)

References

Barras, Laurent, Olivier Scaillet and Russ Wermers, 2010, False discoveries in mutual fund performance: Measuring luck in estimated alphas, *Journal of Finance* 65, 179-216.

Barras, Laurent, Olivier Scaillet and Russ Wermers, 2010, Internet Appendix to: "False discoveries in mutual fund performance: Measuring luck in estimated alphas," *Journal of Finance* 65, 179-216, http://www.afajof.org/supplements.asp.

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

Fama, Eugene F., and Kenneth R. French, 1996, Multifactor explanations of asset pricing anomalies, *Journal of Finance* 51, 55-87.

Fama, Eugene F., and Kenneth R. French, 2010, Luck versus skill in the cross section of mutual fund returns, *Journal of Finance* 65, 1915-1947.

Fung, William, and David A. Hsieh, 2001, The risk in hedge fund strategies: Theory and evidence for trend followers, *Review of Financial Studies* 14, 313-341.

Fung, William, and David A. Hsieh, 2004, Hedge fund benchmarks: A risk based approach, *Financial Analysts Journal* 60, 65-80.

Hansen, Bruce, 2009, Lectures notes on nonparametrics, mimeo, University of Wisconsin.

Scholes, M. and J. Williams, 1977, Estimating betas from nonsynchronous data, *Journal of Financial Economics* 5, 309-328.

Storey, John D., 2002, A direct approach to false discovery rates, *Journal of the Royal Statistical Society B*, part 3, 479-498.