REASSESSING FALSE DISCOVERIES IN MUTUAL FUND PERFORMANCE: SKILL, LUCK, OR LACK OF POWER?*

Angie Andrikogiannopoulou[†] King's College London

Filippos Papakonstantinou[‡] King's College London

January 2018

Abstract

Barras, Scaillet, Wermers (2010) propose the False Discovery Rate to separate skill (alpha) from luck in fund performance. Using simulations with parameters informed by the data, we find that this methodology is overly conservative and underestimates the proportion of nonzero-alpha funds. E.g., 65% of funds with economically large alphas of $\pm 2\%$ are misclassified as zero-alpha. This bias arises from the low signal-to-noise ratio in fund returns and the consequent low statistical power. Our results raise concerns regarding the FDR's applicability in performance evaluation and other domains with low power, and can materially change its conclusion that most funds have zero alpha.

Keywords: Mutual Funds, Skill, Performance, False Discovery Rate, Simulation

JEL Classification: G11, G23, C52, C58

*For helpful comments, we would like to thank Alex Michaelides, Enrico Biffis, Markus Brunnermeier, Pasquale Della Corte, Michael Dempster, Alex Kostakis, Robert Kosowski, Elias Papaioannou, Fabio Trojani, and Dimitri Vayanos, seminar participants at BI Norwegian Business School, Citadel LLC, Imperial College London, Lancaster University, Manchester Business School, and the University of Geneva, and conference participants at the 2012 Cambridge-Princeton conference, the 2014 European Finance Association meeting, and the 2014 European Seminar on Bayesian Econometrics. This manuscript has grown out of simulations and ideas that were previously part of a paper circulated under the title "A Direct and Full-information Estimation of the Distribution of Skill in the Mutual Fund Industry."

[†]King's Business School, Department of Banking & Finance, Bush House, 30 Aldwych, London WC2B 4BG, UK, e-mail: aandriko@kcl.ac.uk.

[‡]King's Business School, Department of Banking & Finance, Bush House, 30 Aldwych, London WC2B 4BG, UK, e-mail: fpapakon@kcl.ac.uk.

In an influential study, Barras, Scaillet and Wermers (2010) — hereafter BSW — propose the False Discovery Rate (FDR) as an advantageous methodology for separating skill from luck and precisely estimating the proportions of funds that generate true 'alpha'. Applying this approach to U.S. equity mutual funds, they find that the vast majority (75%) of funds have zero alpha net of expenses, a sizable minority (24.4%) have negative alpha, and only a negligible proportion (0.6%) beat the benchmarks. These findings have been widely cited in the literature as evidence of no skill in the industry, and they have been interpreted as being consistent with the Berk and Green (2004) equilibrium.[1] But the contribution of BSW stretches beyond the mutual fund literature and extends to introducing and popularizing the FDR methodology in finance. The remarkable accuracy of the FDR estimator as shown by BSW in a simulation — together with the approach's simplicity — has spurred a number of subsequent studies to apply it not only in the context of fund performance but also in other contexts. For example, it has been used to assess the performance of trading strategies, to estimate the proportion of takeovers that experience abnormal trading volume, and to detect jumps in asset returns.[2]

In this study, we reassess whether the FDR methodology can successfully distinguish skill from luck in mutual funds. We expand the simulation of BSW and we find that, for data generating processes that are informed by the mutual fund data, the FDR estimator becomes markedly biased. This bias arises from the fact that the pivotal assumptions behind the estimator fail due to the low signal-to-noise ratio in fund return data and the consequent lack of statistical power in tests of fund alpha. In particular our simulations show that, given the information in the data, the FDR methodology misclassifies as zero-alpha many funds with economically large alphas (e.g., $\pm 2\%$ per year), and as a result, it may greatly underestimate the proportion of nonzero-alpha funds. We also find that, while the number of observations per fund affects the estimator's accuracy, the number of funds itself does not, as it does not affect the signal-to-noise ratio. This distinction is important, as most applications of the FDR in finance involve panels with a large N but small T dimension. We note that, though the simulation in BSW is a valuable first step in assessing the FDR methodology for fund performance evaluation, it does not diagnose these limitations, because it is conducted under the specific assumptions that all nonzero alphas are very large (around 3.5% per year) and there is a large number of observations per fund.

[1] E.g., see Busse, Goyal and Wahal (2010), Ben-Rephael, Kandel and Wohl (2012), Jiang, Verbeek and Wang (2014).

[2] Cuthbertson, Nitzsche and O'Sullivan (2012) and Criton and Scaillet (2014) apply the FDR in the context of UK mutual funds and of hedge funds respectively; Bajgrowicz and Scaillet (2012) apply it in the context of trading strategies and Augustin, Brenner and Subrahmanyam (2015) in the context of takeovers; Patton and Ramadorai (2013) use it to assess funds' risk exposures; and Bajgrowicz, Scaillet and Treccani (2015) use it to detect jumps in asset returns.

The bias we demonstrate brings into question the economic conclusions of the FDR approach for fund alpha. Specifically, the finding that most mutual funds have (almost) zero alpha may not be due to a lack of skill in the industry and may not support the Berk and Green (2004) model in which decreasing returns to scale and rational capital reallocation drive fund alphas to zero. Rather it is likely an artifact of an estimation methodology that has low power to detect nonzero-alpha funds.[3] Overall, our results raise concerns about the applicability of the FDR in fund performance evaluation and more widely in areas in finance where the signal-to-noise ratio in the data is similarly low.

The FDR. The FDR approach was developed by Benjamini and Hochberg (1995) in statistics to control the proportion of null hypotheses that are falsely rejected when conducting multiple tests. As a less conservative alternative to previous approaches such as the Bonferroni correction, the FDR has become widespread in biology-related fields where multiple testing is common. The idea behind this approach is simple. Let I be the number of hypothesis tests and $\hat{p}_i$ the p-value for test i. Assuming that i) the p-values corresponding to true nulls are independent and uniformly distributed on [0, 1] and ii) the p-values corresponding to the alternatives are near 0, one can estimate the proportion of true nulls as follows. Since, by assumption (ii), all p-values above some threshold $\lambda \in (0, 1)$ correspond to true nulls, the proportion π^0 of true nulls can be estimated by counting these p-values, extrapolating to the entire [0, 1] interval, and dividing by the number of tests, i.e., $\hat{\pi}^0 = \frac{\frac{1}{1-\lambda} \cdot \#\{\hat{p}_i : \hat{p}_i > \lambda\}}{I}$. Then, the proportion of nulls expected to be falsely rejected at significance $\gamma \in (0, 1)$ is $\hat{\pi}^0(\lambda) \cdot \gamma$.

While assumption (i) is relatively innocuous,[4] assumption (ii) is quite strong and may easily fail. Essentially, if (some) individual tests have low power to detect the alternative, the p-values that correspond to the alternatives will be distributed over the entire [0, 1] interval. Then, some of the p-values above the threshold λ will correspond to alternatives, so the proportion of nulls (alternatives) will be overestimated (underestimated). Thus, to get meaningful estimates from the FDR approach, it is crucial to assess its performance in the context in which it is applied.

In the context of fund performance, the literature typically separates funds into three groups: 1) those with negative alpha, e.g., because they suffer from exploitable biases or have high costs/fees, 2)

[3] To clearly see that the FDR analysis of the real mutual fund data yields biased estimates, one need only compare the proportion of funds it classifies as skilled/unskilled on the basis of returns *before* and *after* expenses (which average 1% per year). It estimates that 75% of funds have *zero* alpha *after* expenses, which would imply that at least as many have *positive* alpha *before* expenses. But it estimates that only 10% of funds have positive alpha before expenses.

[4] To be precise, uniformity also fails in the presence of dependent data. Benjamini and Yekutieli (2001) have refined the FDR to work for arbitrary dependence, at the expense of over-conservativeness. We examine this issue in the fund performance context in Section 2.4.

those with (almost) zero alpha, consistent with the Berk and Green (2004) equilibrium, and 3) those with positive alpha, e.g., because they possess superior information or trading skill. The FDR methodology can be used to estimate the proportions of these groups while accounting for false discoveries, i.e., lucky/unlucky zero-alpha funds for which the zero-alpha null is incorrectly rejected. But the aforementioned assumption (ii) of the FDR is equivalent here to assuming that nonzero alphas are sufficiently large and/or alpha is estimated with great precision. This is unlikely to hold for the *real* data: Not only is the true alpha distribution likely to have complex features, with some less extreme alphas, but also we know that the amount of information in fund returns varies widely across funds.[5]

Our analysis. First, we investigate the FDR estimates' sensitivity to variations in the distribution of fund alphas. Our starting point is the data generating process that BSW use in their simulations, which is a discrete distribution with large nonzero alphas: a 75% mass at $\alpha = 0$, a 23% mass at $\alpha = -3.2\%$, and a 2% mass at $\alpha = 3.8\%$, per year. Then, we vary the proportions of funds with zero, negative, and positive alpha as well as the location and spread of the nonzero alphas. We find that, as the proportion of zero-alpha funds gets smaller and nonzero alphas become less extreme, the FDR estimator becomes inaccurate: the point estimates are far from the truth and their confidence intervals rarely contain the true proportions. Importantly, the FDR methodology does not misclassify as zero-alpha only funds with *small* nonzero alphas (which might be reasonably expected), but also funds that have economically *large* alphas. For example, 90% (65%) of funds with an alpha of 1% (2%) per year are misclassified as zero alpha. Due to this bias, in many cases the FDR does not outperform the naive approach of simply counting the null rejections without performing a correction for false discoveries.

Second, we explore how the number of observations — hence the amount of information — in the data affects the FDR estimator. BSW are conservative in the cross-sectional dimension of their simulated data but less so in the time-series dimension. That is, they generate balanced panels of fund returns in which the number of funds (1,400) is smaller than in the real data (over 2,000), but the number of observations per fund (384) is equal to the maximum — and more than two times the mean — number of observations across all funds in the data. On the one hand, we find that in fact the number of funds has no effect on the estimator's accuracy, as it does not affect the power of each individual test, nor the distribution of alpha p-values in the sample. It is a little like estimating

[5]Indeed, it has been estimated (see Jones and Shanken, 2005; Andrikogiannopoulou and Papakonstantinou, 2016; Harvey and Liu, 2016) that very few funds have very large alphas of the order that would ensure accurate estimation and a small p-value, while a large proportion of funds have less extreme — but economically large — alphas (e.g., about half of all fund alphas are between 1% and 2.5%, in absolute value, annualized). Furthermore, the number of monthly return observations per fund ranges across funds from fewer than 100 to almost 400, with an average value close to 150.

the mean weight of individuals by separately weighing them with a biased scale; no matter how many individuals are weighed, the result will be equally biased. On the other hand, we find that the number of observations per fund has a strong effect, but convergence to the true proportions can be very slow as this number increases. Alarmingly, given the relatively short time-series dimension of fund data, the FDR methodology is inaccurate even if nonzero alphas are as large as in the BSW simulation; for example, 30% of funds with alphas of $\pm 3.5\%$ are misclassified as zero-alpha. Furthermore, we find that even two hundred years of data may not be sufficient to get accurate estimates from the FDR, which is quite discouraging regarding its applicability in this context.

Finally, we study the effect that cross-sectional correlation in fund return errors has on the FDR estimator. Like BSW, we allow for a latent linear factor error structure, using realistic parameters estimated from the data. We find that error correlation increases the estimator's variability by a factor of five. This means that, even if alphas are large and the FDR estimator is unbiased, it will often be far from the truth.

The remainder of the paper is structured as follows. In Section 1, we describe in more detail the FDR approach and its assumptions. In Section 2, we use simulations to examine the FDR approach's accuracy in the mutual fund setting as we vary the characteristics of the data generating process. In Section 3, we conclude.

1 Fund Performance and the False Discovery Rate

Mutual fund skill, 'alpha', is typically estimated using a linear factor model of fund returns (e.g., Carhart, 1997).[6] A natural approach to estimating the proportions of funds with zero, negative, and positive alpha is to i) perform multiple fund-level regressions of fund returns on factor returns to calculate fund-level alpha estimates $\hat{\alpha}_i$; ii) calculate alpha p-values $\hat{p}_i$ for the zero-alpha null hypothesis $H_i^0 : \alpha_i = 0$ against the alternative $H_i^A : \alpha_i \neq 0$, for each fund $i = 1, \ldots, I$; and iii) count the proportion of funds for which the null is rejected at some significance level. However, this approach (henceforth the 'no-luck' approach) does not control for the probability of incorrectly rejecting true nulls (Type I error). To account for the problem of false discoveries in multiple testing in the context

[6]This definition of skill follows BSW and much of the related literature (e.g., Baks, Metrick and Wachter, 2001; Kosowski et al., 2006; Fama and French, 2010). We note, however, that other definitions of skill have been proposed. Berk and Green (2004) define skill as alpha before costs (including, importantly, information acquisition), Pastor, Stambaugh and Taylor (2015) define it as alpha adjusted for fund and industry size (i.e., the alpha on the first dollar invested in the fund and industry), while Koijen (2014) defines it, under market efficiency, as the price of the active-portfolio risk (i.e., the compensation for holding assets that earn a risk premium).

of mutual fund performance evaluation, BSW apply the False Discovery Rate (FDR) approach.

The FDR methodology estimates the proportion π^0 of true nulls (funds with zero alpha) as

$$\hat{\pi}^0(\lambda) = \frac{\frac{1}{1-\lambda} \cdot \#\left\{\hat{p}_i : \hat{p}_i > \lambda\right\}}{I},$$

where $\lambda \in (0, 1)$ is some threshold p-value. This estimate relies on the following crucial assumptions: i) funds satisfying the zero-alpha null have estimated p-values that are independent and uniformly distributed on the interval [0, 1], and ii) all the p-values that lie above λ correspond to true null hypotheses. Then, given this estimate $\hat{\pi}^0(\lambda)$, the proportion of nulls that are expected to be falsely rejected at any significance level $\gamma \in (0, 1)$ can be calculated as $\hat{\pi}^0(\lambda) \cdot \gamma$. Finally, for a significance level $\gamma \leq \lambda$ high enough that the null hypothesis is rejected for all alternatives (funds with nonzero alpha), we can calculate the proportion of null rejections and adjust downward for the proportion of false null rejections. That is, we can estimate the true proportions π^- and π^+ of negative-alpha and positive-alpha funds as

$$\begin{aligned}
\hat{\pi}^-(\gamma, \lambda) &= \frac{\#\left\{\left(\hat{\alpha}_i, \hat{p}_i\right) : \hat{\alpha}_i < 0, \hat{p}_i < \gamma\right\}}{I} - \hat{\pi}^0(\lambda) \cdot \frac{\gamma}{2} \\
\hat{\pi}^+(\gamma, \lambda) &= \frac{\#\left\{\left(\hat{\alpha}_i, \hat{p}_i\right) : \hat{\alpha}_i > 0, \hat{p}_i < \gamma\right\}}{I} - \hat{\pi}^0(\lambda) \cdot \frac{\gamma}{2}.
\end{aligned}$$

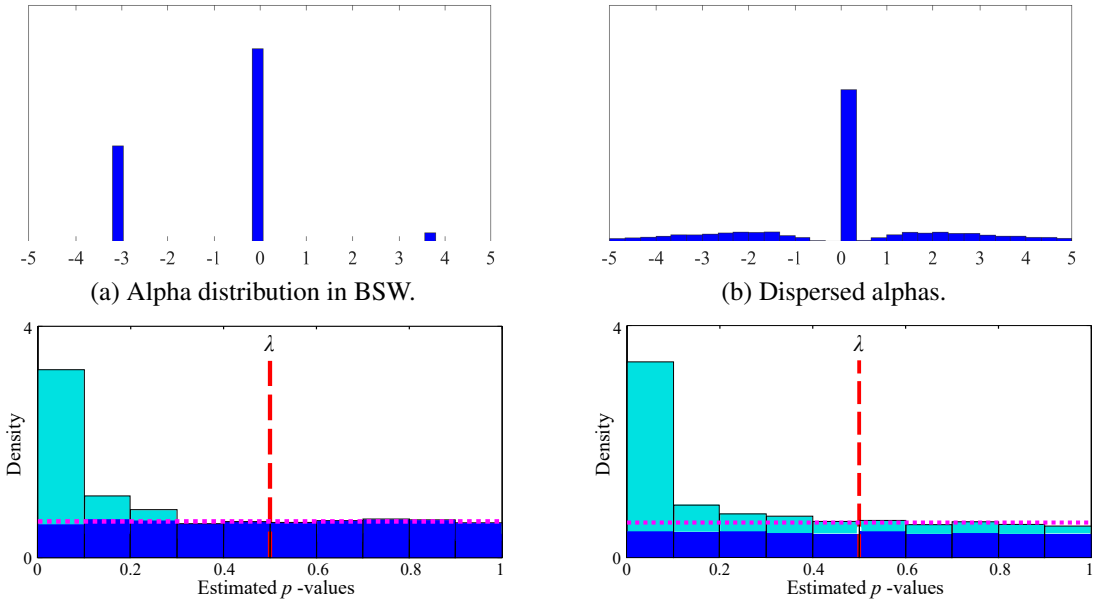
If the assumption holds that all p-values above λ correspond to true nulls, then choosing a large γ effectively also deals with the problem of failing to reject the zero-alpha null when it is false (Type II error). The values for λ and γ can be optimally selected using a bootstrap procedure to minimize the estimated mean squared error of $\hat{\pi}^0$ and of $\hat{\pi}^-$ and $\hat{\pi}^+$ respectively.[7] We note that this is the procedure used in the simulations by BSW as well as in our simulations in Section 2 below.

The crucial assumption that all p-values above the threshold λ correspond to zero-alpha funds is met when nonzero alphas are far from 0 and/or the amount of information in the data is large. But it is likely that some nonzero alphas are not very far from zero, and moreover the information contained in each fund's returns is known to vary widely across funds. As a result, in the context of mutual funds, it is unlikely that the density of p-values satisfies the assumption on which the FDR correction

[7] In particular, Storey (2002) suggests the following procedure for selecting λ. First, set a range of possible values for λ, e.g., $\lambda \in \{0.30, 0.35, \ldots, 0.70\}$, and compute $\hat{\pi}^0(\lambda)$ for each. Second, for each λ, form a number of bootstrap replications of $\hat{\pi}^0(\lambda)$ by drawing with replacement from the p-values in the sample of funds; e.g., form $\hat{\pi}_b^0(\lambda)$ for $b \in \{1, \ldots, 1{,}000\}$. Then, for each λ, calculate the estimated Mean Squared Error (MSE) $\frac{1}{1{,}000} \sum_{b=1}^{1{,}000} \left[\left(\hat{\pi}_b^0(\lambda) - \min_\lambda \hat{\pi}^0(\lambda)\right)^2\right]$, where $\min_\lambda \hat{\pi}^0(\lambda)$ is used as a proxy of π^0 as the latter is unknown. Finally, select the value of λ that minimizes the estimated MSE. BSW suggest an analogous procedure for selecting the value of γ (for details, see the online appendix of BSW).

relies. In this case, the FDR methodology overestimates the proportion of true null hypotheses.[8]

To illustrate this limitation of the FDR approach, we conduct two preliminary simulations using the framework described in Section 2. We generate two samples of fund returns: one in which alphas are generated from the distribution used by BSW — a discrete distribution with 75% mass at $\alpha = 0$, 23% mass at $\alpha = -3.2\%$, and 2% mass at $\alpha = 3.8\%$, annualized — and one in which alphas are generated from a distribution with 45% mass at $\alpha = 0$ and equal weight on two continuous distributions with negative and positive support respectively, as shown in Figure 1b; the remaining simulation parameters are as in BSW (for details, see Section 2.1 below).



(a) Alpha distribution in BSW.

(b) Dispersed alphas.

(c) Alpha p-values in BSW.

(d) p-values for dispersed alphas.

Figure 1: Illustration of the FDR approach's intuition and potential failure. We simulate funds with αs drawn from the distribution used by BSW (in Panel a) and an alternative in which αs are dispersed (in Panel b). In Panels (c) and (d), we plot for each distribution the histogram of p-values $\hat{p}$ from fund-by-fund tests of the zero-α null; the dark blue (light blue) areas correspond to zero-α (nonzero-α) funds. The FDR approach assumes all $\hat{p}$ above some threshold — e.g., $\lambda = 0.5$, indicated by the dashed vertical line — belong to zero-α funds. Since $\hat{p} \sim \mathcal{U}(0, 1)$ under the null, it then estimates the proportion of zero-α (nonzero-α) funds as the area of the histogram below (above) the dotted horizontal line. Thus, the light-blue-shaded area below the dotted horizontal line corresponds to the proportion of nonzero-α funds that are misclassified as zero α.

[8] In Section 2.4, we also investigate the impact of p-value dependence on the performance of the FDR estimator in the fund performance context; it greatly increases the variability of the FDR estimator but not its bias.

In Figure 1, we plot the two alpha distributions and the corresponding histograms of estimated p-values from fund-by-fund tests of the zero-alpha null. We see that, under the first distribution, nonzero alphas are sufficiently far from zero so all p-values above λ belong to zero-alpha funds. But under the second distribution, some nonzero alphas are closer to zero so about a quarter of p-values above λ belong to nonzero-alpha funds. Thus, in the former case the FDR approach estimates the true proportions of funds accurately, while in the latter it overestimates (underestimates) the true proportion of zero-alpha (nonzero-alpha) funds by about a third.

We note that, for illustration purposes, in these simulations we have used $\lambda = 0.5$. We also note that it is not (generally) possible to increase the accuracy of the FDR estimates by increasing λ or by selecting it optimally. This can clearly be seen in Panel (d) of the figure: Since the distribution of p-values is flat above 0.5, the FDR estimate of zero-α funds is essentially the same for all values of λ above it. The intuition for this is that, while the *number* of Type II errors falls as the threshold λ increases, their *proportion* remains the same as they occur in the smaller interval $(\lambda, 1)$. Indeed, BSW note that — in the mutual fund setting — the FDR estimate of zero-α funds is not very sensitive to the value of λ, and intermediate values such as 0.5 produce estimates that are very close to those produced by the optimal approach described above.

Next, we use additional simulations to methodically investigate the effect that the lack of statistical power may have on the estimation of the proportions of fund types, and how this effect varies as we vary the characteristics of the data generating process.

2 Simulation Analysis

2.1 Simulation Setup

To examine the accuracy of the FDR estimator in the context of fund performance, we generate samples of fund returns by combining various data generating processes (DGPs) for alpha with a model for returns. In each sample, we then use the FDR methodology to estimate the proportions of funds with zero, negative, and positive alpha. Here we describe our simulation framework, which closely follows that in BSW.

Model of Returns We generate samples of fund returns according to the linear factor model

$$r_{it} = \alpha_i + F_t'\beta_i + \varepsilon_{it}, \tag{1}$$

with r_{it} the month-t net return of fund i in excess of the risk-free rate, α_i the fund-specific alpha, F_t the month-t factor returns, β_i the fund-specific factor loadings, and ε_{it} the month-t error for fund i.

Mutual Fund Data To select plausible simulation parameters, we use the same mutual fund data sources and apply the same data filters as BSW. Specifically, we obtain monthly fund return data from the CRSP Survivorship-Bias-Free US Mutual Fund Database, for the period 1975–2006.[9] We focus on actively managed open-end US equity funds, so we exclude index, fixed income, international, accrual, money market, and sector funds. We identify funds' share classes using the MFLINKS database, and we compute the monthly return for each fund as the weighted average of its classes' returns, with weights equal to the beginning-of-month total net asset value of each class. To improve data accuracy, we omit any return that directly follows a missing one, as it may compound multiple months' returns. We also keep funds with at least 60 return observations, not necessarily contiguous but with no gaps greater than a year. The benchmarks against which we measure performance are as in Carhart (1997), i.e., the vector of factor returns, F_t, contains the excess return of the market portfolio and the returns of zero-investment factor-mimicking portfolios for size, book-to-market, and momentum. To construct these benchmarks, we use the CRSP NYSE/Amex/NASDAQ value-weighted index as the market factor and the one-month Treasury bill rate as the risk-free rate, while monthly returns for the factor-mimicking portfolios are downloaded from Kenneth French's website.

Simulation Parameters Our starting point is a simulation in which the sample size and the DGPs for the alphas, betas, factor returns, and errors in Equation 1 are as in BSW. In detail, in this baseline a balanced panel of fund returns is generated, with cross-sectional dimension $N = 1{,}400$ and time-series dimension equal to the total number of months in the data ($T = 384$). Alphas are drawn from a discrete distribution with 3 point masses representing funds with zero, negative, and positive alpha. This distribution's parameters are chosen as follows. Initially, the FDR approach is applied to the real data at the end of each year from 2002 to 2006, using all funds' returns up to that point. Next, the proportions of zero-, negative-, and positive-alpha funds are set to the mean estimated proportions (75%, 23%, and 2% respectively). And then, the location of the nonzero alphas (at -3.2% and 3.8% per year) is calibrated to be consistent with these proportions.[10] Factor returns (F_t) are drawn from a normal with parameters equal to their sample counterparts in the data, and factor loadings (β_i) are drawn from a normal with parameters equal to their sample counterparts from fund-level estimations of Equation 1 in the data. Finally, the errors are assumed

[9] Extending the sample to 2011 has no material effect on our results regarding the accuracy of the FDR estimator.

[10] In detail, assuming the negative (and similarly the positive) alphas come from a point mass implies that the distribution of the t-statistics corresponding to these alphas follows a noncentral t distribution with noncentrality that depends on the point mass's location. If, in addition, the FDR approach is accurate, this location can be calculated by comparing the estimated proportions of negative alphas at different significance levels; for details, see the online appendix in BSW.

to be homogeneous, homoscedastic, and cross-sectionally independent, i.e., $\varepsilon_{it} \sim \mathcal{N}\left(0, \sigma^2\right)$, with standard deviation equal to its sample counterpart ($\sigma = 0.021$).

Subsequently, we consider a variety of alternative simulation parameters. First, the aforementioned alpha DGP may miss the mark, because it is calibrated under the assumptions that nonzero alphas come from a point mass and that the FDR methodology is accurate when applied to the real data. However, either of these assumptions may not hold in reality. Thus, in Section 2.2 we examine the estimator's sensitivity to alternative alpha DGPs, in which we vary both the proportions of fund types and the distribution (location and spread) of nonzero alphas. Second, since the real data sample is an unbalanced panel with 2,076 funds and a median (mean) of 150 (180) observations per fund, in Section 2.3 we examine the FDR estimator's performance on an unbalanced panel similar to the one in the data. We also explore how the number of observations in the data — both in the time series and in the cross section — affects the FDR estimator. Third, in Section 2.4 we allow for cross-sectional error correlation, using parameters estimated from the data.

To mitigate the effect of simulation noise, in all simulations we calculate the mean estimate for the proportion of each alpha type across 1,000 repetitions.

2.2 Variation in DGP for alpha

2.2.1 Discrete Nonzero alphas

In our first simulations, we generate alphas from a discrete distribution with three point masses (δ^0, $\delta^-_{\bar{\alpha}}$, and $\delta^+_{\bar{\alpha}}$), i.e.,

$$\alpha \sim \pi^0 \delta^0 + \pi^- \delta^-_{\bar{\alpha}} + \pi^+ \delta^+_{\bar{\alpha}}, \tag{2}$$

varying the proportions π^0, π^-, and π^+ of funds with zero, negative, and positive alpha, and the location ($-\bar{\alpha}$ and $+\bar{\alpha}$) of the nonzero alphas. Specifically, we vary π^0 from 93.75% to 6.25%, keeping the ratio $\frac{\pi^-}{\pi^+}$ equal to 11.5 (as in BSW), and we vary $\bar{\alpha}$ from 1% to 3.5% per year.[11]

In Table 1 we report, for each DGP, the true and the mean (across 1,000 repetitions) estimated proportions of funds with zero, negative, and positive alpha from the FDR methodology, along with the estimates' standard deviations. Each cell of the table is shaded, with darker (lighter) shades corresponding to more (less) biased estimates.[12]

[11] In the online appendix, we present results where the ratio $\frac{\pi^-}{\pi^+}$ equals 6, and our conclusions remain the same.

[12] To be precise, unbiased estimators have no shading and estimators with the maximum possible bias across all DGPs in the table have the darkest shading. For example, across all DGPs in Table 1, the minimum true π^0 is 6.25% (the maximum true π^+ is 7.50%) so the maximum possible bias is 93.75% (7.50%), hence estimators of π^0 (π^+) with bias closer to 93.75% (7.50%) have darker shades. We note that light shading does not indicate *small* bias, but rather bias that is substantially *smaller than the maximum* possible. E.g., for DGP *D*-3 and $\bar{\alpha} = 2.5\%$, the FDR estimates are lightly shaded even though they are very biased ($\pi^0 = 37.50\%$ but $\hat{\pi}^0 = 52.58\%$). All tables in the paper are shaded similarly.

Table 1: Estimates of Alpha Group Proportions — Discrete Nonzero αs

Results from simulations in which nonzero αs (expressed as annualized %ages) are generated from various discrete distributions, i.e., $\alpha \sim \pi^0\delta^0 + \pi^-\delta^-_{\bar{\alpha}} + \pi^+\delta^+_{\bar{\alpha}}$. Across rows, we vary the true proportions π^0, π^-, π^+, and across columns we vary the distance $\bar{\alpha}$ of nonzero αs from 0. Other simulation parameters are as in BSW. In the table, we report the mean (across 1,000 repetitions) estimated proportions from the FDR approach and their standard deviations (in parentheses). Each cell is shaded, with darker shades corresponding to more biased estimates. Results corresponding to the DGP used by BSW are enclosed in a border.

	$\bar{\alpha} = 1.0$	$\bar{\alpha} = 1.5$	$\bar{\alpha} = 2.0$	$\bar{\alpha} = 2.5$	$\bar{\alpha} = 3.0$	$\bar{\alpha} = 3.5$
DGP D-1: $\pi^0 = 93.75\%$	97.95 (2.95)	97.25 (2.95)	96.06 (2.94)	94.99 (2.94)	94.54 (2.94)	94.08 (2.93)
$\pi^- = 5.75\%$	1.25 (2.62)	1.90 (2.62)	2.96 (2.61)	3.86 (2.61)	4.39 (2.61)	4.77 (2.61)
$\pi^+ = 0.50\%$	0.81 (1.30)	0.85 (1.30)	0.98 (1.29)	1.15 (1.29)	1.07 (1.29)	1.15 (1.29)
DGP D-2: $\pi^0 = 75.00\%$	94.97 (2.94)	89.96 (2.91)	85.00 (2.89)	80.85 (2.86)	78.35 (2.84)	76.33 (2.82)
$\pi^- = 23.00\%$	4.61 (2.59)	9.35 (2.57)	13.75 (2.55)	17.27 (2.53)	19.65 (2.52)	21.41 (2.50)
$\pi^+ = 2.00\%$	0.42 (1.24)	0.69 (1.22)	1.25 (1.21)	1.89 (1.20)	2.00 (1.20)	2.26 (1.20)
DGP D-3: $\pi^0 = 37.50\%$	86.71 (2.90)	74.42 (2.80)	62.35 (2.67)	52.58 (2.52)	45.22 (2.39)	40.94 (2.30)
$\pi^- = 57.50\%$	13.18 (2.53)	24.96 (2.45)	35.47 (2.34)	43.75 (2.24)	49.82 (2.14)	53.75 (2.07)
$\pi^+ = 5.00\%$	0.10 (1.11)	0.63 (1.06)	2.18 (1.03)	3.66 (1.01)	4.95 (1.00)	5.31 (1.00)
DGP D-4: $\pi^0 = 6.25\%$	80.41 (2.85)	61.95 (2.66)	43.50 (2.35)	28.03 (1.97)	17.77 (1.61)	11.62 (1.32)
$\pi^- = 86.25\%$	19.59 (2.46)	37.66 (2.30)	53.28 (2.06)	65.99 (1.77)	74.71 (1.51)	80.48 (1.31)
$\pi^+ = 7.50\%$	0.01 (0.99)	0.39 (0.90)	3.22 (0.85)	5.98 (0.82)	7.51 (0.80)	7.90 (0.79)

We see that for the DGP similar to that used in the BSW simulations — $\pi^0 = 75\%$, $\pi^- = 23\%$, $\pi^+ = 2\%$, and $\bar{\alpha} = 3.5\%$ — the FDR yields proportions (76.3%, 21.4%, 2.3%, respectively) close to the true ones. But it becomes biased as the proportion of zero-alpha funds becomes lower and nonzero alphas become less extreme but remain economically large. For example, for the DGP close to the Andrikogiannopoulou and Papakonstantinou (2016) estimates from the real data — $\pi^0 = 6.25\%$, $\pi^- = 86.25\%$, $\pi^+ = 7.5\%$, and $\bar{\alpha} = 1\%$ — the FDR approach is very conservative: It estimates more than ten times as many zero-alpha funds ($\hat{\pi}^0 = 80\%$), a quarter as many negative-alpha funds ($\hat{\pi}^- = 20\%$), and virtually no positive-alpha funds ($\hat{\pi}^+ = 0.01\%$), massively underestimating their presence. Notably, our results also raise concerns about the accuracy of the FDR estimates calculated from the real data ($\hat{\pi}^0 = 75\%$, $\hat{\pi}^- = 24.4\%$, $\hat{\pi}^+ = 0.6\%$). As we see in Table 1, these estimates could have arisen from several DGPs that are economically very different, e.g., both from DGP D-2, for which zero-alpha funds are a large majority, as well as from DGPs D-3 and D-4, for which zero-alpha funds are a minority and the majority of funds have alpha of 1% to 2%.

Importantly, the mis-estimation we document is not because the FDR approach is *economically* conservative, but because it is *statistically* conservative. That is, it does not misclassify as zero-alpha funds those with alpha close to zero but rather those for which, given the noise in the data, there is insufficient power to reject the null. To see this, it is useful to consider a back-of-the-envelope calculation. In the real (hence in our simulated) data, the average fund has return volatility of about 5% per month and about 80% of this is explained by the factor model. So, with 384 observations, the standard error of a fund's $\hat{\alpha}$ is about 0.115% per month ($\sqrt{\frac{0.05^2 \cdot 0.2}{384}}$).[13] With a threshold p-value $\lambda = 0.5$, a fund with an economically significant alpha of, e.g., $\pm 1\%$ ($\pm 2\%$) annualized has $\hat{\alpha}$ p-value that exceeds λ with probability 40% (20%).[14] This implies that the FDR methodology misclassifies $\frac{40\%}{1-\lambda} = 80\%$ ($\frac{20\%}{1-\lambda} = 40\%$) of funds with $\alpha = \pm 1\%$ ($\alpha = \pm 2\%$). Indeed, in the first (third) column of Table 1 we see that, across DGPs, about 80% (40%) of nonzero-alpha funds are misclassified as having zero alpha.[15]

A further important point is that, as a consequence of the FDR estimator's bias, the resulting confidence intervals can become highly misleading. To see this, we construct 95% confidence intervals for the proportions — π^0, π^-, π^+ — for all DGPs of Table 1, and we present in Table 2 their actual coverage probabilities, i.e., the percentage (across 1,000 repetitions) of samples in which the true proportion is in the corresponding interval. For example, for π^0 we calculate the proportion of samples such that the true π^0 is in the interval $\hat{\pi}^0 \pm 1.96 \cdot \hat{\sigma}_{\hat{\pi}^0}$, where $\hat{\pi}^0$ is the FDR estimate of π^0 and $\hat{\sigma}_{\hat{\pi}^0}$ is the estimate's standard deviation. We find that, for most DGPs, the actual coverage probabilities are very different from the nominal. In particular, for about half our DGPs, the confidence intervals *never* contain the true proportions.[16]

[13] The standard error SE can be written as $\sqrt{\hat{\sigma}_r^2(1-R^2)/N}$, with $\hat{\sigma}_r$ the sample standard deviation of returns. Also, the power of rejecting at significance level λ the zero-alpha null against a two-sided alternative when the true value is α can be roughly approximated by $1 - \Phi\left(\Phi^{-1}\left(1 - \frac{\lambda}{2}\right) - |\alpha|/\text{SE}\right)$, where Φ is the standard normal cdf.

[14] As noted in BSW and our discussion above toward the end of Section 1, $\lambda = 0.5$ yields close to optimal results for the FDR methodology in the mutual fund context. To be clear, $\lambda = 0.5$ is used for illustration purposes here; in all our simulations, we select λ optimally (see Footnote 7).

[15] Following BSW, in these simulations we use samples of 384 observations per fund. As we have noted above, in the real data the average number of observations per fund is half that, so the standard error of the $\hat{\alpha}$ estimate is even higher and the FDR approach's accuracy is even lower than shown in this rough calculation. In a sample with 180 observations per fund, the FDR approach misclassifies 90% (65%) of nonzero-alpha funds with $\alpha = \pm 1\%$ ($\alpha = \pm 2\%$) as zero-alpha funds.

[16] Similar results for 90% and 99% confidence intervals are presented in Section A.6 in the online appendix. The coverage probabilities of the confidence intervals are similarly low for all subsequent simulations in Sections 2.3 and 2.4; while we do not present a table of coverage probabilities for each of them, it is easy to see that taking an interval that extends twice the standard deviation on either side of the FDR estimates of the proportions rarely contains the true proportions.

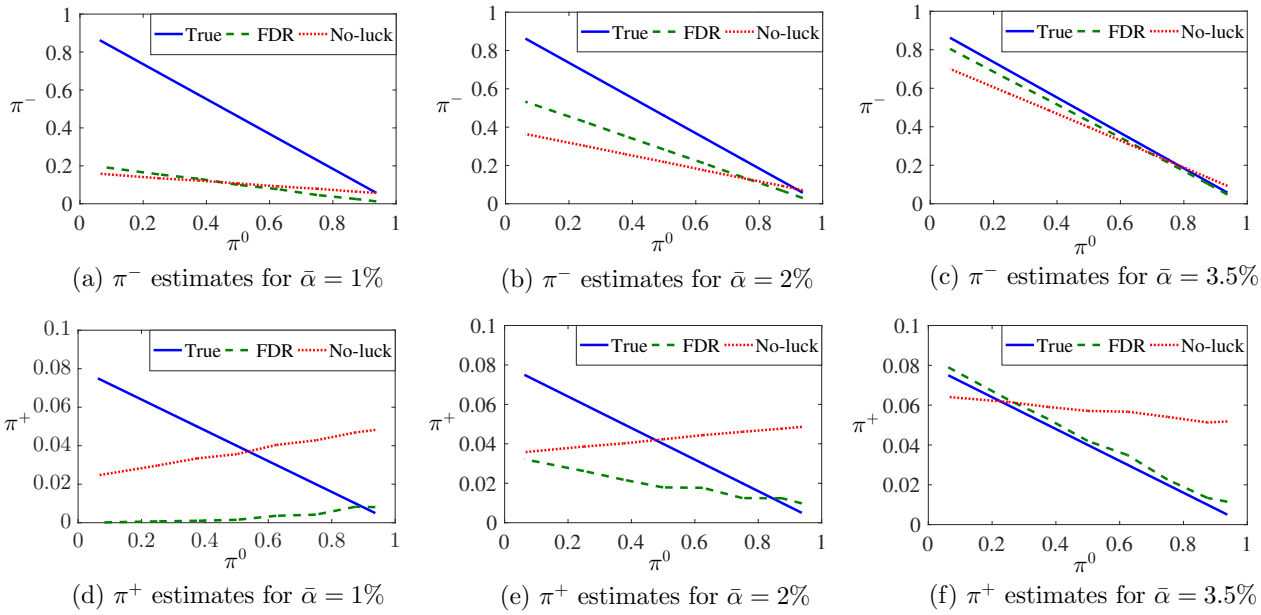
Table 2: Actual Coverage Probabilities of 95% Confidence Intervals of FDR Estimates

Results corresponding to the simulations in Table 1, where nonzero αs come from discrete distributions ($\alpha \sim \pi^0\delta^0 + \pi^-\delta^-_{\bar{\alpha}} + \pi^+\delta^+_{\bar{\alpha}}$), with the true proportions π^0, π^-, π^+ varying across rows, and the distance $\bar{\alpha}$ of nonzero αs from 0 varying across columns. Here, we report the percentage (across 1,000 repetitions) of samples such that the true proportion is contained in the 95% confidence interval constructed by the FDR methodology (e.g., for π^0, it is $\hat{\pi}^0 \pm 1.96\ \hat{\sigma}_{\hat{\pi}^0}$). Each cell of the table is shaded, with darker (lighter) shades corresponding to cases in which the *actual* percentage of confidence intervals containing the true value is farther (closer) to the *nominal* coverage of 95%. Results for the DGP used by BSW are enclosed in a border.

		$\bar{\alpha} = 1.0$	$\bar{\alpha} = 1.5$	$\bar{\alpha} = 2.0$	$\bar{\alpha} = 2.5$	$\bar{\alpha} = 3.0$	$\bar{\alpha} = 3.5$
DGP *D*-1:	$\pi^0 = 93.75\%$	63.50%	76.00%	89.50%	92.50%	95.00%	98.00%
	$\pi^- = 5.75\%$	57.50%	73.50%	88.50%	92.50%	94.50%	97.50%
	$\pi^+ = 0.50\%$	92.00%	92.50%	89.50%	88.50%	88.50%	89.50%
DGP *D*-2:	$\pi^0 = 75.00\%$	0.00%	0.50%	5.50%	48.00%	79.50%	94.00%
	$\pi^- = 23.00\%$	0.00%	0.00%	0.50%	29.50%	85.50%	97.50%
	$\pi^+ = 2.00\%$	99.50%	98.50%	96.00%	93.00%	91.50%	87.50%
DGP *D*-3:	$\pi^0 = 37.50\%$	0.00%	0.00%	0.00%	0.00%	8.00%	72.00%
	$\pi^- = 57.50\%$	0.00%	0.00%	0.00%	0.00%	0.00%	67.00%
	$\pi^+ = 5.00\%$	0.50%	6.00%	30.50%	63.00%	79.00%	80.00%
DGP *D*-4:	$\pi^0 = 6.25\%$	0.00%	0.00%	0.00%	0.00%	0.00%	0.50%
	$\pi^- = 86.25\%$	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	$\pi^+ = 7.50\%$	0.00%	0.00%	4.50%	53.50%	89.00%	93.00%

Next, we check if the FDR approach offers a significant improvement relative to the previously used 'no-luck' approach, which simply counts the proportion of funds with significant alphas at the 0.1 significance level without correcting for Type I errors. In Figure 2, we compare the two approaches for different levels of $\bar{\alpha}$. We plot the true and estimated proportions of negative- and positive-alpha funds from each approach as a function of the true proportion of zero-alpha funds, for three levels of $\bar{\alpha}$: 1%, 2%, and 3.5%. In Panels (c), (f) of the figure we see that, for extreme alphas ($\bar{\alpha} = 3.5\%$), the FDR approach estimates the proportions accurately for all levels of π^0, while the no-luck approach is less accurate as it fails to account for lucky and unlucky funds. But in the other panels we see that for less extreme, yet economically large, alphas — $\bar{\alpha} = 1\%$ and $\bar{\alpha} = 2\%$ — the FDR approach does not perform better than the no-luck approach, and it may even

perform worse on average, due to its large bias.[17]



(a) π^- estimates for $\bar{\alpha} = 1\%$ (b) π^- estimates for $\bar{\alpha} = 2\%$ (c) π^- estimates for $\bar{\alpha} = 3.5\%$

(d) π^+ estimates for $\bar{\alpha} = 1\%$ (e) π^+ estimates for $\bar{\alpha} = 2\%$ (f) π^+ estimates for $\bar{\alpha} = 3.5\%$

Figure 2: Comparison of the FDR and 'no-luck' approaches, for various levels of the true nonzero alphas $\bar{\alpha}$ and the true proportion π^0 of zero-alpha funds. We plot as a function of π^0 the true and estimated proportions of negative-alpha (Panels *a*, *b*, *c*) and positive-alpha (Panels *d*, *e*, *f*) funds. We plot the true proportions in solid blue and the FDR (no-luck) estimates in dashed green (dotted red) lines. Each column of panels corresponds to a different $\bar{\alpha}$: Panels (a), (d) correspond to $\bar{\alpha} = 1\%$, Panels (b), (e) to $\bar{\alpha} = 2\%$, and Panels (c), (f) to $\bar{\alpha} = 3.5\%$.

2.2.2 Normal Nonzero alphas

Next, we generate fund alphas from a mixture of a point mass at 0 and a normal distribution, i.e.,

$$\alpha \sim \pi^0 \delta^0 + \left(\pi^- + \pi^+\right) f_{\mathcal{N}(\mu_\alpha, \sigma_\alpha^2)},$$

where we set $\mu_\alpha = 0$ for ease of exposition, and we vary the proportion of zero-alpha funds (π^0) from 95% to 1% and the standard deviation (σ_α) from 0.5% to 5% per year. This specification is important as it nests the distributions estimated by some studies in the literature (e.g., Jones and

[17] Panels (c) and (f) replicate Figure 3 in BSW, with slight differences. First, in BSW's Figure 3c, the line plotting π^- should be steeper than — not identical to — the line plotting the FDR estimates (as plotted, $\pi^- + \pi^+ \neq 1$ when $\pi^0 = 0$). Second, for expositional purposes, BSW have drawn t-statistics from a mixture of distributions with small variance so that Type II errors do not occur. Instead, we simulate fund returns using the baseline simulation parameters and compute alpha t-statistics by regressing the returns on the factor model; so Type II errors do occur. Third, in our figure, lines plotting the no-luck estimates are constructed using tests of size 10% rather than 20%; this explains why the no-luck line in our Figure 2f is downward-sloping and tends to 5% as π^0 increases, rather than being upward-sloping and tending to 10%.

Shanken, 2005; Fama and French, 2010).[18] We present results from these simulations in Table 3.

Table 3: Estimates of Alpha Group Proportions — Normal Nonzero αs

Results from simulations in which nonzero αs (expressed as annualized %ages) are generated from normals, i.e., $\alpha \sim \pi^0\delta^0 + \left(\pi^- + \pi^+\right) f_{\mathcal{N}(0,\sigma_\alpha^2)}$. Across rows, DGPs differ in the true proportions π^0, π^-, and π^+ — though $\pi^- = \pi^+$ since $\mathcal{N}\left(0, \sigma_\alpha^2\right)$ is symmetric at 0 — and across columns they differ in the standard deviation σ_α of the nonzero αs. Other simulation parameters are as in BSW. In the table, we report the mean (across 1,000 repetitions) estimated proportions from the FDR approach and their standard deviations (in parentheses). Cells are shaded, with darker shades corresponding to more biased estimates.

		$\sigma_\alpha = 0.5$		$\sigma_\alpha = 1.0$		$\sigma_\alpha = 1.5$		$\sigma_\alpha = 2.0$		$\sigma_\alpha = 3.0$		$\sigma_\alpha = 5.0$	
DGP N-1:	$\pi^0 = 95.00\%$	98.67	(2.95)	98.21	(2.95)	98.01	(2.95)	97.46	(2.95)	96.89	(2.95)	96.12	(2.94)
	$\pi^- = 2.50\%$	0.54	(2.63)	0.78	(2.63)	0.79	(2.63)	1.13	(2.63)	1.39	(2.63)	1.74	(2.63)
	$\pi^+ = 2.50\%$	0.80	(1.32)	1.01	(1.32)	1.20	(1.33)	1.41	(1.33)	1.73	(1.34)	2.15	(1.34)
DGP N-2:	$\pi^0 = 75.00\%$	98.18	(2.95)	95.42	(2.94)	92.43	(2.93)	89.58	(2.91)	85.68	(2.89)	81.87	(2.86)
	$\pi^- = 12.50\%$	0.74	(2.63)	2.07	(2.63)	3.39	(2.62)	4.92	(2.62)	6.86	(2.61)	8.90	(2.60)
	$\pi^+ = 12.50\%$	1.09	(1.32)	2.51	(1.34)	4.18	(1.36)	5.50	(1.37)	7.46	(1.39)	9.24	(1.41)
DGP N-3:	$\pi^0 = 20.00\%$	95.67	(2.94)	86.00	(2.89)	76.18	(2.82)	67.58	(2.73)	55.15	(2.56)	42.52	(2.33)
	$\pi^- = 40.00\%$	1.84	(2.63)	6.33	(2.61)	10.97	(2.58)	15.43	(2.53)	21.69	(2.44)	28.24	(2.30)
	$\pi^+ = 40.00\%$	2.49	(1.34)	7.67	(1.39)	12.85	(1.43)	16.99	(1.45)	23.16	(1.47)	29.24	(1.48)
DGP N-4:	$\pi^0 = 10.00\%$	94.94	(2.94)	84.02	(2.88)	72.80	(2.79)	63.46	(2.68)	49.70	(2.47)	35.32	(2.17)
	$\pi^- = 45.00\%$	2.21	(2.63)	7.24	(2.61)	12.73	(2.56)	17.47	(2.50)	24.18	(2.38)	31.80	(2.20)
	$\pi^+ = 45.00\%$	2.85	(1.34)	8.75	(1.39)	14.46	(1.43)	19.07	(1.46)	26.12	(1.47)	32.88	(1.47)
DGP N-5:	$\pi^0 = 1.00\%$	94.52	(2.94)	82.61	(2.87)	70.05	(2.76)	59.56	(2.63)	44.56	(2.38)	29.09	(2.00)
	$\pi^- = 49.50\%$	2.50	(2.63)	7.95	(2.60)	14.08	(2.55)	19.20	(2.47)	26.69	(2.32)	34.78	(2.10)
	$\pi^+ = 49.50\%$	2.98	(1.35)	9.44	(1.40)	15.87	(1.44)	21.23	(1.46)	28.75	(1.48)	36.13	(1.46)

As in the previous simulations, the lower the proportion of zero-alpha funds and the narrower the spread of the distribution, the larger the FDR estimates' bias. For example, for the DGP close to the ones estimated by Jones and Shanken (2005) and Fama and French (2010) on the real data — with $\pi^0 = 1\%$, $\pi^- = \pi^+ = 49.5\%$, and $\sigma_\alpha = 1.5\%$ — the FDR approach grossly overestimates π^0 (70% instead of 1%) and underestimates π^- and π^+ (about 15% instead of 50% each).

[18] Jones and Shanken (2005) find that alphas follow a normal with mean around -0.8% and standard deviation around 2%, and Fama and French (2010) estimate a normal with zero mean and standard deviation around 1% per year. Here, we generate alphas from a normal centered at 0, and in the online appendix we present results from a normal centered at -0.8%; the results are very similar.

Motivated by recent studies which suggest that the distribution of fund alpha may be fat-tailed (Sastry, 2015; Andrikogiannopoulou and Papakonstantinou, 2016), we also examine the sensitivity of the FDR estimators to fat tails by simulating nonzero alphas from a t distribution. We find that fat tails slightly ameliorate the bias in the FDR estimates, but this effect is very weak and the bias is still large even with very fat tails. We present detailed results from this analysis in the online appendix.

2.3 Variation in Simulated Sample Size

The above simulations use a balanced panel with $N = 1{,}400$ funds and $T = 384$ observations per fund, as in BSW. But the real data sample is an unbalanced panel with $N = 2{,}076$ funds and T ranging from 60 to 384 with median/mean of 150/180. As a result, it is important to examine the FDR estimator's sensitivity to the number of funds, to the number of observations per fund, as well as to the use of an unbalanced panel similar to the one observed in reality.

Table 4: Estimates of Alpha Group Proportions — Varying N and T

Results from simulations in which we vary the number of funds N and the number of observations T per fund in the sample, assuming a balanced panel. In all simulations, αs (expressed as annualized %ages) are drawn from $\alpha \sim \pi^0\delta^0 + \pi^-\delta^-_{\bar{\alpha}} + \pi^+\delta^+_{\bar{\alpha}}$, with $\pi^0 = 9\%$, $\pi^- = 78\%$, $\pi^+ = 13\%$ and $\bar{\alpha} = 1\%$. Other simulation parameters are as in BSW. In the table, we report the mean (across 1,000 repetitions) estimated proportions from the FDR approach and their standard deviations (in parentheses). Each cell is shaded, with darker shades corresponding to more biased estimates.

		$T = 180$		$T = 384$		$T = 500$		$T = 750$		$T = 1{,}000$		$T = 2{,}000$	
$N = 1{,}400$:	$\pi^0 = 9.00\%$	90.65	(2.92)	81.29	(2.86)	76.28	(2.82)	66.53	(2.72)	58.89	(2.62)	35.74	(2.18)
	$\pi^- = 78.00\%$	9.31	(2.55)	18.50	(2.49)	23.17	(2.46)	31.51	(2.38)	37.67	(2.31)	55.05	(1.99)
	$\pi^+ = 13.00\%$	0.04	(1.13)	0.21	(1.07)	0.55	(1.06)	1.96	(1.04)	3.44	(1.03)	9.20	(1.00)
$N = 2{,}000$:	$\pi^0 = 9.00\%$	90.79	(2.44)	80.94	(2.39)	76.31	(2.36)	66.79	(2.28)	58.82	(2.19)	35.89	(1.83)
	$\pi^- = 78.00\%$	9.19	(2.13)	18.90	(2.08)	23.25	(2.06)	31.44	(2.00)	37.71	(1.93)	54.95	(1.67)
	$\pi^+ = 13.00\%$	0.02	(0.94)	0.16	(0.90)	0.44	(0.89)	1.77	(0.87)	3.47	(0.86)	9.16	(0.84)
$N = 3{,}500$:	$\pi^0 = 9.00\%$	90.70	(1.85)	81.30	(1.81)	76.18	(1.78)	66.76	(1.72)	58.57	(1.65)	36.11	(1.39)
	$\pi^- = 78.00\%$	9.30	(1.61)	18.64	(1.58)	23.55	(1.56)	31.63	(1.51)	37.77	(1.46)	54.88	(1.26)
	$\pi^+ = 13.00\%$	0.00	(0.71)	0.05	(0.68)	0.27	(0.67)	1.61	(0.65)	3.66	(0.65)	9.01	(0.63)
$N = 5{,}000$:	$\pi^0 = 9.00\%$	90.70	(1.54)	81.31	(1.51)	76.25	(1.49)	66.83	(1.44)	58.62	(1.38)	35.93	(1.16)
	$\pi^- = 78.00\%$	9.30	(1.35)	18.67	(1.32)	23.60	(1.30)	31.61	(1.26)	37.79	(1.22)	54.97	(1.06)
	$\pi^+ = 13.00\%$	0.00	(0.60)	0.02	(0.57)	0.15	(0.56)	1.56	(0.55)	3.58	(0.54)	9.10	(0.53)

In Table 4 we report the mean estimated proportions of funds as a function of the simulated sample size, for a DGP in which the FDR approach performs poorly and which is consistent with

empirical evidence using the real data (see Andrikogiannopoulou and Papakonstantinou, 2016; Harvey and Liu, 2016). Specifically, in this simulation alpha is drawn from a discrete distribution with $\pi^0 = 9\%$, $\pi^- = 78\%$, $\pi^+ = 13\%$, and $\bar{\alpha} = \pm 1\%$. We see that increasing N has virtually no effect on the bias in the FDR estimator, while increasing T makes the estimates more accurate. However, for the DGP we consider, we see that convergence is very slow: even two hundred years of data (i.e., more than 2,000 monthly observations) are not sufficient to get accurate estimates from the FDR approach.[19] Given that varying N has no effect on the accuracy of the FDR estimators, without loss of generality we use $N = 2{,}000$, as in the real data, in all subsequent simulations we conduct.

Next, we study the effect of having an unbalanced panel with characteristics similar to the panel observed in reality. We revisit our first simulations (see Equation 2), but we draw the number of observations per fund from its empirical distribution in the data. We report our results in Table 5.

Table 5: Estimates of Alpha Group Proportions — Unbalanced Panel

Results from simulations in which nonzero αs (expressed as annualized %ages) are generated from discrete distributions ($\alpha \sim \pi^0\delta^0 + \pi^-\delta^-_{\bar{\alpha}} + \pi^+\delta^+_{\bar{\alpha}}$), but additionally the sample of funds is an unbalanced panel with the number of observations per fund drawn from its empirical distribution in the data (whose mean is 180). Across rows, DGPs differ in the true proportions π^0, π^-, and π^+, and across columns they differ in the distance $\bar{\alpha}$ of nonzero αs from zero. Other simulation parameters are as in BSW. We report the mean (across 1,000 repetitions) estimated proportions from the FDR methodology and their standard deviations (in parentheses). Cells are shaded, with darker shades corresponding to more biased estimates.

		$\bar{\alpha} = 1.0$		$\bar{\alpha} = 1.5$		$\bar{\alpha} = 2.0$		$\bar{\alpha} = 2.5$		$\bar{\alpha} = 3.0$		$\bar{\alpha} = 3.5$	
DGP D-1:	$\pi^0 = 93.75\%$	98.66	(2.47)	98.57	(2.47)	97.71	(2.47)	96.72	(2.46)	96.25	(2.46)	95.97	(2.46)
	$\pi^- = 5.75\%$	0.81	(2.19)	1.02	(2.19)	1.61	(2.19)	2.48	(2.19)	2.96	(2.19)	3.27	(2.19)
	$\pi^+ = 0.50\%$	0.53	(1.09)	0.41	(1.08)	0.69	(1.08)	0.80	(1.08)	0.79	(1.08)	0.77	(1.08)
DGP D-2:	$\pi^0 = 75.00\%$	97.11	(2.46)	94.85	(2.46)	91.35	(2.44)	88.14	(2.43)	85.11	(2.42)	82.95	(2.40)
	$\pi^- = 23.00\%$	2.57	(2.18)	4.83	(2.17)	8.14	(2.16)	11.03	(2.15)	13.74	(2.14)	15.60	(2.13)
	$\pi^+ = 2.00\%$	0.32	(1.05)	0.32	(1.04)	0.50	(1.03)	0.82	(1.02)	1.14	(1.01)	1.45	(1.01)
DGP D-3:	$\pi^0 = 37.50\%$	93.56	(2.45)	86.61	(2.42)	78.89	(2.38)	70.75	(2.31)	63.33	(2.24)	57.16	(2.17)
	$\pi^- = 57.50\%$	6.37	(2.15)	13.31	(2.12)	20.91	(2.08)	28.45	(2.02)	34.78	(1.97)	39.96	(1.92)
	$\pi^+ = 5.00\%$	0.06	(0.97)	0.08	(0.93)	0.20	(0.90)	0.80	(0.88)	1.89	(0.87)	2.88	(0.85)
DGP D-4:	$\pi^0 = 6.25\%$	90.60	(2.44)	79.99	(2.38)	68.62	(2.30)	55.76	(2.15)	45.10	(2.00)	35.57	(1.82)
	$\pi^- = 86.25\%$	9.40	(2.12)	20.00	(2.06)	31.31	(1.98)	43.26	(1.87)	52.20	(1.74)	59.96	(1.61)
	$\pi^+ = 7.50\%$	0.00	(0.90)	0.01	(0.83)	0.06	(0.78)	0.99	(0.74)	2.71	(0.72)	4.47	(0.70)

[19] In the online appendix, we show similar effects for a DGP with $\bar{\alpha} = \pm 2\%$, for which the FDR approach performs better. Even though in that case convergence to the true values is, as expected, faster as T increases, it is still very slow.

Comparing Tables 1 and 5 we see that, for all DGPs, the bias is more pronounced in the presence of an unbalanced panel with a shorter average time-series dimension. With the shorter unbalanced panel, the FDR methodology misclassifies 90% (65%) of nonzero-alpha funds with $\alpha = \pm 1\%$ ($\alpha = \pm 2\%$) as zero-alpha funds. Even for large alphas of $\pm 3.5\%$, the FDR approach is not as accurate as shown above: it misclassifies about 30% of nonzero-alpha funds. We note that this increase in bias is not due to the unbalanced nature of the panel, but due to the smaller average number of observations per fund. Indeed, the FDR estimates are almost identical whether we simulate an unbalanced panel with a mean of 180 observations per fund (see Table 5) or a balanced panel with $T = 180$ (see Table A.7 in the online appendix). This is because the empirical distribution of the number of observations per fund has small positive skewness, so alphas that are estimated less precisely are balanced out by those estimated more precisely and the overall proportion of misclassified funds is unchanged.

2.4 Cross-sectional Error Correlation

As we have noted, the FDR approach assumes that p-values, therefore the errors in Equation 1, are independent. BSW show in a simulation that the FDR estimators continue to be very accurate when they introduce cross-sectional correlation in the factor model errors. Here, we revisit this simulation and make several adjustments to better capture the characteristics of the real data.

In detail, BSW replace the assumption $\varepsilon_{it} \sim \mathcal{N}\left(0, \sigma^2\right)$ in Equation 1 with a latent linear factor structure intended to capture the role of non-priced factors. In particular, $\varepsilon_{it} = G_t\delta + G_t^-\delta I_{\alpha_i=-\bar{\alpha}} + G_t^+\delta I_{\alpha_i=\bar{\alpha}} + \xi_{it}$, where all funds load on factor G_t, only negative- and positive-alpha funds load on factors G_t^- and G_t^+ respectively, δ is the *common* loading of funds on these factors, and $\xi_{it} \sim \mathcal{N}\left(0, \sigma^{*2}\right)$ is the cross-sectionally independent part of the error. Assuming each factor follows a normal $\mathcal{N}(0, \sigma_G)$, BSW set rule-of-thumb parameters $\sigma_G = 0.035$ (the mean monthly standard deviation of the Fama-French factors) and $\delta = 0.11$ (the mean loading to the Fama-French factors).

Here, we replace the assumption $\varepsilon_{it} \sim \mathcal{N}\left(0, \sigma^2\right)$ with $\varepsilon_{it} := G_t'\delta_i + \xi_{it}$, where G_t is a vector of four latent factors, one on which all funds may load and three strategy-specific factors on which only funds with a specific investment strategy (Growth & Income, Growth, or Aggressive Growth) may load, δ_i are *fund-specific* loadings, and $\xi_{it} \sim \mathcal{N}\left(0, \sigma_i^{*2}\right)$ with σ_i^* a *fund-specific* standard deviation. Importantly, the fund-specific δ_i and σ_i^* we use in our simulations are estimated from the data (for estimation details, see Geweke and Zhou, 1996 and Andrikogiannopoulou and Papakonstantinou, 2016).[20]

[20] The proportions of funds following each investment strategy in our simulations are as in the data (strategy classifications derived from the Thomson database): 22%, 66%, and 12%, respectively. The fund-specific δ_i and σ_i^* are estimated from the data; the means and standard deviations of δ_i are $10^{-2} \cdot \left[\begin{array}{cccc} 0.60 & 0.33 & 1.07 & 2.01 \end{array}\right]'$ and $10^{-2} \cdot \left[\begin{array}{cccc} 0.78 & 0.46 & 1.08 & 1.07 \end{array}\right]'$, respectively, and the mean of σ_i^* is 0.014. Also, $G_t \sim \mathcal{N}(0, I)$ without loss of generality.

Table 6: Interval Estimates of Alpha Group Proportions — Independent vs. Correlated Errors

Interval estimates corresponding to simulations in which the sample of funds is an unbalanced panel and nonzero αs (expressed as annualized percentages) are generated from a variety of discrete distributions ($\alpha \sim \pi^0\delta^0 + \pi^-\delta^-_{\bar{\alpha}} + \pi^+\delta^+_{\bar{\alpha}}$), and the errors in the model of returns are heterogeneous across funds and *independent* (Panel A) or *correlated* (Panel B). Across rows, the DGPs differ in the true proportions π^0, π^-, and π^+, and across columns they differ in the distance $\bar{\alpha}$ of nonzero αs from zero. For each DGP, we conduct 1,000 simulations and calculate the estimated proportions from the FDR approach. In the table, we report the interval that contains 90% of the estimated proportions across the 1,000 repetitions. Note: For mean point estimates corresponding to the interval estimates in Panel A (B), see Table A.2 (A.3) in the online appendix.

Panel A: Cross-sectionally Independent Errors

	$\bar{\alpha} = 1.0$	$\bar{\alpha} = 1.5$	$\bar{\alpha} = 2.0$	$\bar{\alpha} = 2.5$	$\bar{\alpha} = 3.0$	$\bar{\alpha} = 3.5$
DGP D-1: $\pi^0 = 93.75\%$	(94.9, 99.9)	(93.9, 99.9)	(93.5, 99.9)	(92.7, 99.9)	(92.2, 99.9)	(92.4, 99.7)
$\pi^- = 5.75\%$	(0.0, 2.7)	(0.0, 3.7)	(0.0, 4.3)	(0.0, 5.0)	(0.0, 5.4)	(0.3, 5.4)
$\pi^+ = 0.50\%$	(0.0, 2.8)	(0.0, 3.3)	(0.0, 2.8)	(0.0, 3.1)	(0.0, 2.9)	(0.0, 3.1)
DGP D-2: $\pi^0 = 75.00\%$	(92.6, 99.9)	(89.5, 96.7)	(85.9, 93.2)	(83.7, 90.6)	(80.6, 88.4)	(78.8, 86.3)
$\pi^- = 23.00\%$	(0.0, 6.2)	(3.3, 8.9)	(6.8, 11.8)	(9.4, 14.3)	(11.6, 16.2)	(13.7, 18.0)
$\pi^+ = 2.00\%$	(0.0, 1.6)	(0.0, 2.1)	(0.0, 3.0)	(0.0, 3.2)	(0.0, 3.6)	(0.0, 3.7)
DGP D-3: $\pi^0 = 37.50\%$	(86.7, 94.9)	(78.7, 87.1)	(71.3, 78.7)	(63.8, 71.2)	(58.2, 65.3)	(54.0, 60.4)
$\pi^- = 57.50\%$	(5.1, 12.8)	(12.9, 20.4)	(21.3, 27.8)	(28.7, 33.8)	(34.1, 38.5)	(38.2, 41.9)
$\pi^+ = 5.00\%$	(0.0, 0.4)	(0.0, 1.6)	(0.0, 2.4)	(0.0, 3.3)	(0.0, 4.4)	(0.3, 4.8)
DGP D-4: $\pi^0 = 6.25\%$	(81.6, 89.1)	(69.8, 78.3)	(58.4, 65.6)	(47.7, 55.0)	(39.4, 46.1)	(31.9, 38.3)
$\pi^- = 86.25\%$	(10.9, 18.3)	(21.7, 30.1)	(34.4, 40.3)	(44.2, 49.5)	(52.4, 56.8)	(58.3, 63.0)
$\pi^+ = 7.50\%$	(0.0, 0.0)	(0.0, 0.0)	(0.0, 2.5)	(0.0, 3.6)	(0.8, 4.9)	(2.5, 5.6)

Panel B: Cross-sectionally Correlated Errors

	$\bar{\alpha} = 1.0$	$\bar{\alpha} = 1.5$	$\bar{\alpha} = 2.0$	$\bar{\alpha} = 2.5$	$\bar{\alpha} = 3.0$	$\bar{\alpha} = 3.5$
DGP D-1: $\pi^0 = 93.75\%$	(74.8, 99.9)	(76.8, 99.9)	(75.9, 99.9)	(77.1, 99.9)	(72.5, 99.9)	(76.1, 99.9)
$\pi^- = 5.75\%$	(0.0, 18.0)	(0.0, 19.4)	(0.0, 17.7)	(0.0, 19.4)	(0.0, 23.9)	(0.0, 16.6)
$\pi^+ = 0.50\%$	(0.0, 16.2)	(0.0, 15.3)	(0.0, 17.8)	(0.0, 16.5)	(0.0, 19.1)	(0.0, 16.9)
DGP D-2: $\pi^0 = 75.00\%$	(75.4, 99.9)	(73.8, 99.9)	(74.7, 99.9)	(71.6, 99.9)	(69.2, 99.9)	(63.8, 98.3)
$\pi^- = 23.00\%$	(0.0, 21.7)	(0.0, 22.9)	(0.0, 21.2)	(0.0, 27.8)	(0.0, 27.8)	(1.7, 33.8)
$\pi^+ = 2.00\%$	(0.0, 10.4)	(0.0, 12.3)	(0.0, 15.1)	(0.0, 14.2)	(0.0, 18.5)	(0.0, 20.5)
DGP D-3: $\pi^0 = 37.50\%$	(70.9, 99.9)	(58.2, 99.9)	(59.8, 91.6)	(52.0, 82.6)	(49.7, 76.1)	(43.6, 67.5)
$\pi^- = 57.50\%$	(0.0, 28.0)	(0.0, 40.3)	(6.6, 40.2)	(14.3, 46.9)	(20.7, 48.8)	(26.8, 55.7)
$\pi^+ = 5.00\%$	(0.0, 10.8)	(0.0, 10.5)	(0.0, 9.7)	(0.0, 10.2)	(0.0, 8.9)	(0.0, 8.2)
DGP D-4: $\pi^0 = 6.25\%$	(63.7, 99.9)	(57.6, 94.3)	(42.2, 85.2)	(34.8, 69.3)	(27.9, 64.6)	(23.8, 56.7)
$\pi^- = 86.25\%$	(0.0, 36.3)	(5.1, 42.3)	(14.4, 56.8)	(26.2, 63.3)	(32.9, 69.8)	(39.7, 72.2)
$\pi^+ = 7.50\%$	(0.0, 5.8)	(0.0, 5.0)	(0.0, 6.1)	(0.0, 5.6)	(0.0, 5.9)	(1.2, 6.2)

We find that the FDR estimator's bias is unaffected but its variability is greatly increased in the presence of this correlation. For the baseline DGP, the 90% intervals for the estimates of $\pi^0/\pi^-/\pi^+$ are 4.6/7.5/5.5 times wider than in the independence case (compare results for DGP D-2 with $\bar{\alpha} = 3.5\%$ in Panels A and B of Table 6). For other alpha DGPs, this difference becomes even larger.[21]

A consequence of this increased estimator variability is that, even if nonzero alphas are large enough that the FDR estimator is not biased, the estimator will often be far from the truth. For DGP D-2 with large alphas of $\pm 3.5\%$, the FDR methodology has little bias but it estimates the proportions of zero-, negative, and positive-alpha funds to be above 90%, below 10%, and 0%, respectively, in more than a quarter of the simulated samples. Another consequence of this increased estimator variability is that, when the true proportions are near the natural boundaries of 0 and 1, the FDR estimator becomes biased, even for very large $\bar{\alpha}$. For example for a DGP with $\pi^+ = 0.5\%$, the estimates of π^+ are 5 to 7 times larger than the truth, even for $\bar{\alpha} = 3.5\%$ (see Table A.3 in the online appendix).

3 Concluding Discussion

BSW propose the FDR approach to precisely estimate the proportion of funds with negative, zero, and positive alpha in the population of mutual funds. Owing to its simplicity and its remarkable accuracy as demonstrated in their simulations, this methodology has subsequently been widely adopted in finance. In this study, we expand the BSW simulations to gain a deeper understanding of the FDR's performance. We find that the FDR performs poorly under certain conditions, and that the accuracy found in past simulations in the mutual fund setting can be traced to potentially unrealistic simulation assumptions such as that i) fund alphas follow a discrete distribution with point masses very far from zero, ii) there is a large number of observations per fund, and iii) there is no or limited cross-sectional correlation across funds.

First, we show that, for a wide range of DGPs, the FDR approach produces estimates for the proportion of zero-alpha (nonzero-alpha) funds that are upward (downward) biased and that the corresponding confidence intervals almost never contain the true proportions. We also show that the FDR estimator's accuracy is not improved as the number of funds in the sample increases, and that convergence to the true values is very slow as the number of observations per fund increases. Finally, we show that realistic levels of cross-sectional correlation increase the estimator's variability by as

[21] For the same DGP, BSW find a more modest increase of 1.5/2.0/2.2 in the confidence intervals. BSW also consider other types of error correlation (e.g., block dependence), one of which they term "extreme." In this case, they find that the 90% intervals for the estimates of $\pi^0/\pi^-/\pi^+$ are 2.8/2.7/2.5 times wider than in the baseline. This effect is still considerably smaller than the one we find.

much as an order of magnitude. Overall, our results indicate that the FDR approach is unlikely to offer a substantial improvement over simpler methodologies in settings where the signal-to-noise ratio in the data is low and consequently individual tests have low power. Specifically in the context of mutual funds, the bias we demonstrate can materially change the economic conclusions regarding the prevalence of skill in the mutual fund industry, the rationality of investing in mutual funds, and the validity of the Berk and Green (2004) theory, which suggests most funds have zero alpha.

Our results highlight that, while the FDR methodology can be very advantageous in some settings, it needs to be applied with caution. Therefore, it is important that researchers in finance examine the data at hand to check, e.g., using simulations, that individual tests have sufficient power to render the FDR accurate. Another possibility would be to consider approaches that aim to mitigate the problem of low power by focusing on the alternative hypothesis and/or by pooling information from all individuals (e.g., funds) to learn about the cross-sectional distribution of the parameters of interest (e.g., alpha). This is a potentially fruitful avenue that is being pursued in recent working papers in the context of fund performance evaluation (Ferson and Chen, 2015; Andrikogiannopoulou and Papakonstantinou, 2016; Harvey and Liu, 2016).

References

Andrikogiannopoulou, A, and F Papakonstantinou. 2016. "Estimating mutual fund skill: A new approach." Unpublished Paper.

Augustin, P, M Brenner, and M. G Subrahmanyam. 2015. "Informed options trading prior to M&A announcements: Insider trading?" Unpublished Paper.

Bajgrowicz, P, and O Scaillet. 2012. "Technical trading revisited: False discoveries, persistence tests, and transaction costs." *Journal of Financial Economics*, 106(3): 473–491.

Bajgrowicz, P, O Scaillet, and A Treccani. 2015. "Jumps in high-frequency data: Spurious detections, dynamics, and news." *Management Science*, 62(8): 2198–2217.

Baks, K, A Metrick, and J Wachter. 2001. "Should investors avoid all actively managed mutual funds? A study in Bayesian performance evaluation." *Journal of Finance*, 56(1): 45–85.

Barras, L, O Scaillet, and R. R Wermers. 2010. "False discoveries in mutual fund performance: Measuring luck in estimated alphas." *Journal of Finance*, 65(1): 179–216.

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

Benjamini, Y, and Y Hochberg. 1995. "Controlling the false discovery rate: A practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1): 289–300.

Ben-Rephael, A, S Kandel, and A Wohl. 2012. "Measuring investor sentiment with mutual fund flows." *Journal of Financial Economics*, 104(2): 363 – 382. Special Issue on Investor Sentiment.

Berk, J. B, and R. C Green. 2004. "Mutual fund flows and performance in rational markets." *Journal of Political Economy*, 112(6): 1269–1295.

Busse, J. A, A Goyal, and S Wahal. 2010. "Performance and persistence in institutional investment management." *Journal of Finance*, 65(2): 765–790.

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

Criton, G, and O Scaillet. 2014. "Hedge fund managers: Luck and dynamic assessment." *Bankers, Markets & Investors*, 129: 28–38.

Cuthbertson, K, D Nitzsche, and N O'Sullivan. 2012. "False discoveries in UK mutual fund performance." *European Financial Management*, 18(3): 444–463.

Fama, E. F, and K. R French. 2010. "Luck versus skill in the cross-section of mutual fund returns." *Journal of Finance*, 65(5): 1915–1947.

Ferson, W, and Y Chen. 2015. "How many good and bad fund managers are there, really?" Unpublished Paper.

Geweke, J, and G Zhou. 1996. "Measuring the pricing error of the arbitrage pricing theory." *Review of Financial Studies*, 9(2): 557–587.

Harvey, C. R, and Y Liu. 2016. "Rethinking performance evaluation." Unpublished Paper.

Jiang, H, M Verbeek, and Y Wang. 2014. "Information content when mutual funds deviate from benchmarks." *Management Science*, 60(8): 2038–2053.

Jones, C, and J Shanken. 2005. "Mutual fund performance with learning across funds." *Journal of Financial Economics*, 78(3): 507–552.

Koijen, R. S. 2014. "The cross-section of managerial ability, incentives, and risk preferences." *Journal of Finance*, 69(3): 1051–1098.

Kosowski, R, A Timmermann, R Wermers, and H White. 2006. "Can mutual fund stars really pick stocks? New evidence from a bootstrap analysis." *Journal of Finance*, 61(6): 2551–2595.

Pastor, L, R Stambaugh, and L. A Taylor. 2015. "Scale and skill in active management." *Journal of Financial Economics*, 116(1): 23–45.

Patton, A. J, and T Ramadorai. 2013. "On the high-frequency dynamics of hedge fund risk exposures." *Journal of Finance*, 68(2): 597–635.

Sastry, R. 2015. "The cross-section of investing skill." Unpublished Paper.

Storey, J. 2002. "A direct approach to false discovery rates." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3): 479–498.