JOURNAL OF FINANCIAL AND QUANTITATIVE ANALYSIS

doi:10.1017/S0022109024000097

Mutual Funds' Conditional Performance Free of Data Snooping Bias

Po-Hsuan Hsu
National Tsing Hua University, College of Technology Management
pohsuanhsu@mx.nthu.edu.tw (corresponding author)

Ioannis Kyriakou
City, University of London, Bayes Business School
ioannis.kyriakou.2@city.ac.uk

Tren Ma
University of Nottingham, Nottingham University Business School
tren.ma@nottingham.ac.uk

Georgios Sermpinis
University of Glasgow, Adam Smith Business School
georgios.sermpinis@glasgow.ac.uk

Abstract

We introduce a test to assess mutual funds' "conditional" performance that is based on updated information and corrects data snooping bias. Our method, named the functional false discovery rate "plus" (fFDR^+), incorporates fund characteristics in estimating fund performance free of data snooping bias. Simulations suggest that the fFDR^+ controls well the ratio of false discoveries and gains considerable power over prior methods that do not account for extra information. Portfolios of funds selected by the fFDR^+ outperform other tests not accounting for information updating, highlighting the importance of evaluating mutual funds from a conditional perspective.

I. Introduction

It is well known that luck plays an important role in mutual funds' performance (Kosowski, Timmermann, Wermers, and White (2006)). In order to appropriately assess fund performance, investors should rely on a multiple hypothesis testing framework to correct for "data snooping" bias or "p-hacking," a major challenge to social science (Sullivan, Timmermann, and White (1999), (2001), White (2000), Hansen (2005), Hsu, Taylor, and Wang (2016), and Chordia, Goyal, and Saretto (2020)).

We would like to thank Hendrik Bessembinder (the editor), Keith Cuthbertson, Jens Perch Nielsen, Olivier Scaillet (the reviewer), Zheng Sun, Qifei Zhu, and the audience of the American Finance Association (AFA) student poster session and the China International Conference in Finance (CICF) for their comments. We appreciate the replies and discussions of Gery Geenens, Kuangyu Wen, and Ximing Wu on their works on density estimations, as well as Laurent Barras and Russ Wermers on the mutual fund data. Hsu acknowledges the Ministry of Education and the Ministry of Science and Technology in Taiwan for financial support (Yushan Fellow Program and MOST109-2628-H-007-001-MY4). All errors are our own.

Prior research has developed multiple hypothesis testing frameworks to correct such bias and control the number of false discoveries, which are, in our context, seemingly profitable funds that are just due to luck. In particular, researchers propose the concept of the false discovery rate (FDR) of Benjamini and Hochberg (1995), Storey (2002), (2003), and Romano and Wolf (2005), that is, the ratio of models that are mistakenly identified as having predictive power. One common feature of the methodologies in this framework is that the rejection criterion *only* depends on information of raw data and predictive models' performance metrics. This feature appears too restrictive or even unrealistic because, in economics and finance research, the economic agents use all available and routinely updated information in assessing models' performance. Extra information sources can assist researchers to more accurately estimate the FDR. Recently, Chen, Robinson, and Storey (CRS) (2021) introduced the functional FDR method that embeds the role of informative covariates (i.e., variables that carry extra information) in forming null hypotheses. This advancement is important in the sense that it enables us to test the "conditional" performance of predictive models, which is more consistent with the rational expectation hypothesis.[1] In the context of mutual funds, if we use prior testing methods that do not account for extra information, we are testing an unconditional zero hypothesis, which corresponds to investors not collecting external information in assessing mutual fund performance. This approach appears inappropriate because mutual funds and their managers are routinely reviewed by investors based on updated information. In other words, a more suitable null hypothesis for a mutual fund's performance should be zero conditional on the updated information set.

In this article, we introduce the functional FDR "plus" (fFDR^+) method. Compared to the work of CRS, it has two distinguishing features in assessing mutual fund performance. First, it allows us to focus on the right tail of the distribution and detect the significantly outperforming funds, which is important for investors (see Barras et al. (2010), hereafter BSW). Second, it is robust to cross-sectional dependencies among performance measures, which are common for mutual funds because their alphas are likely dependent due to common shareholdings and herding in trading behavior (Wermers (1999)). Compared to all earlier methods in the economics literature on control of the FDR, our fFDR^+ method incorporates extra information, has higher power, and controls for noise. In addition, it is easy to implement, does not rely on any strong assumptions, and can handle any continuous fund characteristic.

In examining our method, we use simulated mutual fund performance similar to BSW and Andrikogiannopoulou and Papakonstantinou (AP) (2019). We show that, when an informative covariate (i.e., fund characteristics) is available, our fFDR^+ approach detects more true-positive-alpha funds under different alpha distributions, balanced and unbalanced data, and both cross-sectional independence and dependence in the error terms. The gap in power between the fFDR^+ and prior

[1]Since White (2000), several multiple testing procedures have been proposed to correct for luck in the past (Hansen (2005), Romano and Wolf (2005), Barras, Scaillet, and Wermers (2010), Hsu, Hsu, and Kuan (2010), and Bajgrowicz and Scaillet (2012)); however, they only consider unconditional null hypotheses.

https://doi.org/10.1017/S0022109024000097 Published online by Cambridge University Press

FDR methods, depending on the distribution of the fund alpha population, can be up to about 30%. Our approach is also robust to estimation errors in the covariates.

We then apply our method and construct portfolios in order to evaluate them empirically in selecting outperforming mutual funds. In particular, we explore ten fund characteristics as informative covariates: The first set contains six fund attributes that have been shown in prior studies to convey information on mutual fund performance, and the second set contains four new attributes that are inspired by asset pricing models. The first set includes the R^2 of the asset pricing model (e.g., Carhart 4-factor model) as suggested by Amihud and Goyenko (2013), the return gap of Kacperczyk, Sialm, and Zheng (2008), the active weight of Doshi, Elkamhi, and Simutin (2015), the fund size of Harvey and Liu (2017), the fund flow suggested by Zheng (1999), and the expense ratio. The second set includes the Sharpe ratio, the beta, and Treynor ratio based on the capital asset pricing model (CAPM), and the idiosyncratic volatility of the Carhart 4-factor model (sigma).

We find that the set of mutual funds selected as outperformers by the fFDR$^+$ is usually larger and different from the one obtained by prior FDR methods. As already discussed, earlier studies do not account for information other than mutual funds' returns and performance metrics; thus, their null hypotheses are unconditional and neglect investors' time-varying expectation. The fact that our fFDR$^+$ discovers more outperforming funds suggests that, with more information input, there may exist more profitable mutual funds than researchers have detected.

Based on the funds selected by the fFDR$^+$, we build portfolios that consistently outperform the one generated by prior FDR methods. Our results highlight the economic value of extra information. In particular, the fFDR$^+$ portfolios based on beta are found to be the best with annualized alphas of 1.1%, followed by the fFDR$^+$ portfolios based on expense ratio, R^2, active weight, sigma, fund flow, return gap, Treynor ratio, fund size, and Sharpe ratio, separately achieving annualized alphas of at least 0.17%. We note that this profitability is persistent in our sample and is even strengthened over the recent period prior to the COVID-19 pandemic, a finding that disagrees with part of the recent literature that suggests otherwise (see Jones and Mo (2021)). All our fFDR$^+$ portfolios outperform the one generated by prior FDR methods and a set of portfolios created by single and double sorting the covariates under study. This finding suggests that the relationship between luck and funds' performance with the mentioned fund characteristics is nonlinear and that traditional portfolio approaches that do not control luck may be inadequate.

In additional analysis, we also consider the fFDR$^+$ portfolio based on various ways of combining the 10 covariates, such as the first principal component of the 10 covariates (PC 1), the ordinary least squares (OLS), the least absolute shrinkage and selection operator (LASSO) of Tibshirani (1996), and the ridge regression and the elastic net of Zou and Hastie (2005). We find that the ridge and elastic net deliver the best performance with an annualized alpha of at least 0.86%. Investors may also benefit from such combinations as they result in lower volatility in portfolio performance. In fact, we find that our fFDR$^+$ portfolios based on the combined covariates gain the highest Sharpe ratio in the most recent decade. This is advantageous as, in reality, investors do not know ex ante what covariate is the best.

https://doi.org/10.1017/S0022109024000097 Published online by Cambridge University Press

The literature on mutual funds' performance has two main strands: one that tries to model the distribution of mutual funds in terms of alphas and identify the outperforming funds and managers and another that focuses on identifying covariates that explain mutual funds' performance.[2] Consequently, to identify outperforming funds researchers simply rank the funds' alphas and covariates.[3] Our study contributes to the mutual fund literature as follows: First, our fFDR$^+$ approach estimates the FDR as a function of a fund characteristic that is related to a fund's performance. By designing and implementing suitable Monte Carlo simulation experiments, we illustrate that our fFDR$^+$ approach actually controls for the FDR and delivers higher power than its benchmark under cross-sectional dependence and different distributions of the fund alpha population.[4] In our empirical analyses, fund portfolios based on the fFDR$^+$ consistently dominate benchmark portfolios in terms of generating positive alphas. Our simulations and empirical evidence collectively highlight the importance of evaluating the conditional performance of mutual funds and the persistence of outperforming funds identified by the fFDR$^+$.

Second, our research adds to the mutual fund literature by exploring different information contents of fund characteristics. Based on our fFDR$^+$, we construct portfolios that persistently produce positive alphas for decades. Our portfolios based on 4 new covariates perform well and outrank, in the context of our method, those based on the traditional 6 covariates on several metrics and subsamples. Finally, we move one step further and combine the 10 covariates into single ones via linear combinations with the weights obtained from a principal component analysis and shrinkage regression methods. We find that investors might benefit from such combinations as they offer lower volatility in portfolio performance.

The rest of the article is organized as follows: In Section II, we introduce and explain our methodology. In Section III, we provide a description of our data. Section IV is devoted to our simulation experiment descriptions, whereas in Section V, we present in detail our simulation results. Section VI focuses on the empirical part of our analysis. Section VII concludes the article. All technical

[2]In one of the earliest studies, Jensen (1968) documents that the majority of active equity mutual fund managers are unable to beat passive investment strategies' net of fee. More recent research incorporates cross-sectional information and assesses funds' performance via a Bayesian approach with some prior beliefs about the distribution of fund alphas. For example, Jones and Shanken (2005) assume that the fund alpha population has a normal distribution and use the Gibbs sampling technique to estimate the parameters of the distribution, whereas Harvey and Liu (2018) adopt a mixture of normals and introduce an expectation maximization technique to estimate the weights and the parameters of the component distributions. Others assess funds in different aspects such as the horizon of the return used to estimate alpha (see, e.g., Bessembinder, Cooper, and Zhang (2023)).

[3]For instance, Carhart (1997) constructs a portfolio by sorting mutual funds according to their past performance (e.g., lagged 1-year return and 3-year past 4-factor model alpha). Kacperczyk et al. (2008) discover that the return gap, defined as the difference between the fund's reported return and the return based on previous holdings, can predict the fund's future performance. Similarly, Doshi et al. (2015) present the active weight metric that conveys information about the fund's future performance and demonstrate predictability. Other researchers do multiple sorting on variables related to funds' performance. For example, Amihud and Goyenko (2013) show that a fund's R^2 can predict its performance.

[4]We consider a discrete distribution as in BSW, a mixture of discrete and normal distributions as in AP, a single normal distribution as in Fama and French (2010) and Jones and Shanken (2005), and a mixture of 2 normals studied in Harvey and Liu (2018).

details, simulations, and robustness checks are provided in the Supplementary Material.

II. Methods for Controlling Luck with Fund Characteristics

A. Functional False Discovery Rate (fFDR)

Throughout this article, we use mutual funds to represent predictive models. We define funds' performance based on their net return, that is, the return net of trading costs, fees, and other expenses except loads and taxes. A fund is deemed outperforming if it distributes to investors a net return that generates a positive alpha (i.e., a part of a return series that is unexplained by systematic risk). If the alpha is negative (zero), the fund is said to be underperforming (zero alpha). These definitions of outperforming and underperforming funds coincide with skilled and unskilled funds in BSW, respectively, and reflect the interest of investors.[5]

Suppose that we are assessing m funds and each of them has a net return time series. We also assume that there exists a continuous fund characteristic X, with observed values $(x_1,\ldots,x_m)$, that conveys information about the alpha of each fund. Our fund characteristic corresponds to the informative covariate in the statistical literature (e.g., CRS and Ignatiadis and Huber (2021)). Associated with X, we define Z whose observed value for fund i is $z_i = rank(x_i)/m$, where $rank(x_i)$ is the ranking of x_i in the set of observed values $(x_1,\ldots,x_m)$. As X to Z is an one–one mapping and we work based on Z, we call that the covariate from now on. We introduce our notation by means of a single test, conditional on Z, for the alpha of a mutual fund:

$$H_0 : \alpha = 0, \quad H_1 : \alpha \neq 0. \tag{1}$$

We denote by h the status of the null hypothesis (i.e., $h=0$ if the hypothesis $\alpha=0$ is true and $h=1$ if otherwise). In addition, P is the random variable representation of the p-value of the test, Z, as mentioned previously, is the covariate that is uniformly distributed on $[0,1]$, and $T=(P,Z)$. We suppose that $(h|Z=z)\sim \text{Bernoulli}(1-\pi_0(z))$; that is, conditional on $Z=z$, the fund possesses a zero alpha with probability $\pi_0(z)$; this can be constant if Z does not convey any information about the probability of the fund's alpha being 0. The estimation procedure for $\pi_0(z)$ will be discussed later on. We require that under the true null, $(P|h=0,Z=z)$ is uniformly distributed on $[0,1]$ regardless of the value of z; when

[5]We note that traditional approaches, such as the studies of Carhart (1997), Kosowski et al. (2006), and Fama and French (2010), define fund skill by the alpha that the fund delivers to investor. However, recent literature in mutual funds proposes differently. Berk and van Binsbergen (2015) provide convincing arguments that ones should not use the net alpha nor the gross alpha that the fund delivers to investors as a measure of skill. They show that the value added (i.e., the value that a mutual fund extracts from capital markets) always measures skill. Subsequently, other studies such as Barras, Gagliardini, and Scaillet (2022) further separate the effects of scale from the measure of skill. Thereby, the skill is defined as the gross alpha earned on the first dollar invested in the fund. More specifically, Barras et al. (2022) model a time-varying gross alpha (the alpha calculated based on gross return) of a fund and express it as $a-bq_{t-1}$, where a and b are defined as skill and scale coefficients and the q_{t-1} is lagged fund size. In this study, we consider the net alpha as a measurement of performance.

https://doi.org/10.1017/S0022109024000097 Published online by Cambridge University Press

the null hypothesis is false, the conditional density function of $(P|h=1, Z=z)$ is $f_1(.|z)$.

To assess the performance of m funds in terms of α within our framework, we consider m conditional hypothesis tests like equation (1):

$$H_{0,i}: \alpha_i = 0, \quad H_{1,i}: \alpha_i \neq 0, \quad i = 1, \ldots, m, \tag{2}$$

where α_i is the alpha of fund i. For each i, we have $T_i = (P_i, Z_i)$, and we assume that all the triples (T_i, h_i) are independent and each of them has the same distribution as (T, h).[6] Finally, we denote by $f(p,z)$ the joint density function of (P, Z). We thus have

$$\mathbb{P}(h=0|T=(p,z)) = \frac{\pi_0(z)}{f(p,z)} =: r(p,z) \tag{3}$$

as the posterior probability of the null hypothesis being true given that we observe $T = (p,z)$.[7]

To control the type I error, Storey (2003) introduces the "positive false discovery rate"

$$\text{pFDR} = \mathbb{E}\left(\frac{V}{R}\middle| R > 0\right), \tag{4}$$

where R is the number of rejected hypotheses in m tests and V is the wrongly rejected ones. CRS show that, with a fixed set Γ in $[0,1]^2$, if we reject hypothesis $H_{0,i}$ whenever $T_i \in \Gamma$, then

$$\text{pFDR}(\Gamma) = \mathbb{P}(h=0|T \in \Gamma) = \int_\Gamma r(p,z)dpdz. \tag{5}$$

To maximize the number of rejections, we reject the hypotheses with the smallest statistic $r(p,z)$. Thus, the significance region $\{\Gamma_\theta : \theta \in [0,1]\}$ is defined as

$$\Gamma_\theta = \{(p,z) \in [0,1]^2 : r(p,z) \leq \theta\}, \tag{6}$$

where a larger θ implies more rejected hypotheses. Finally, we recall from Storey (2003) and CRS the definition of the q-value for the observed (p,z):

$$q(p,z) = \inf_{\{\Gamma_\tau | (p,z) \in \Gamma_\tau\}} \text{pFDR}(\Gamma_\tau) = \text{pFDR}\left(\Gamma_{r(p,z)}\right). \tag{7}$$

Given a target $\tau \in [0,1]$, a procedure that rejects a hypothesis if and only if its q-value $\leq \tau$ guarantees that pFDR is controlled at τ.

Empirically, let $\hat{\pi}_0(z)$ and $\hat{f}(p,z)$ be the estimated functions $\pi_0(z)$ and $f(p,z)$, respectively.[8] We calculate $\hat{r}(p,z) = \hat{\pi}_0(z)/\hat{f}(p,z)$ and estimate the q-value function as

[6]In Sections IV.A and IV.B of the Supplementary Material, we show that this requirement can be eased for a typically cross-sectional dependence in mutual fund data. We also note that the FDR framework of Storey (2002) and the FDR$^+$ of BSW also work outside the independent and identically distributed framework (see Storey, Taylor, and Siegmund (2004), Bajgrowicz and Scaillet (2012).

[7]For more details about the role of $Z \sim \text{Uniform}(0,1)$ and the derivation of equation (3), see CRS.

[8]See the Appendix for more details.

$$\hat{q}(p_i, z_i) = \frac{1}{\#S_i} \sum_{k \in S_i} \hat{r}(p_k, z_k), \tag{8}$$

where $S_i = \{j | \hat{r}(p_j, z_j) \leq \hat{r}(p_i, z_i)\}$ and p_i is the p-value of test i and $\#S_i$ returns the number of elements in the set S_i.[9] Then, given a target pFDR level $\tau \in [0, 1]$, the null hypothesis $H_{0,i}$ is rejected if and only if $\hat{q}(p_i, z_i) \leq \tau$. CRS call this procedure functional FDR (fFDR).

B. The **fFDR**$^+$: Application in Selecting Outperforming Funds

By applying the fFDR methodology to mutual funds at a given target pFDR level τ, we obtain a set that includes both significantly outperforming and underperforming funds. To further improve mutual fund selection, we propose a fFDR-based method that selects a group of significantly outperforming funds with control of luck. In the following section, we introduce our fFDR$^+$ and discuss its application in a mutual fund context.

Consider a selection of R^+ outperforming funds including V^+ wrongly selected zero-alpha or underperforming funds. We define the pFDR in those significantly outperforming funds as

$$\text{pFDR}^+ = \mathbb{E}\left(\frac{V^+}{R^+} \middle| R^+ > 0\right). \tag{9}$$

For m tests, let A^+ be the set of hypotheses with positive estimated alpha (i.e., $A^+ = \{i | \hat{\alpha}_i > 0\}$), where $\hat{\alpha}_i$ is the estimated alpha of fund i. At a given target τ of pFDR$^+$, by implementing the fFDR procedure to control pFDR at the target τ on the funds in set A^+, we obtain all the funds with positive estimated alphas (referred to as significant alphas).[10] Hence, the fFDR selects positive-alpha funds with control of pFDR at the given target; we call this procedure the functional FDR "plus" (fFDR$^+$).

Different from our approach, BSW propose a procedure to estimate the FDR in detecting outperforming mutual funds, namely the FDR$^+$, which utilizes only p-value and alpha of funds. For the sake of space, we present details of the FDR$^+$ and a comparison between it and our fFDR$^+$ in Section I of the Supplementary Material.

As shown in AP, the FDR$^+$ relies on an overconservative estimate of the null proportion and utilizes only p-values and the estimated alphas. On the other hand, our fFDR$^+$ additionally uses a fund characteristic and expresses the null proportion as a function of it, while accounting for the joint distribution of the p-value and the fund characteristic. As documented in CRS, this results in a better estimate of FDR,

[9] The $\#S_i$ is the number of discoveries given $\theta = \hat{r}(p_j, z_j)$, while the numerator is the expected number of false discoveries. This estimation is proposed by Newton, Noueiry, Sarkar, and Ahlquist (2004) and Storey, Akey, and Kruglyak (2005) and subsequently adopted in CRS.

[10] In doing so, we assume that the number of funds that are outperforming but exhibit a negative estimated alpha is negligible. This is sensible as in practice we will not select those funds anyway. In BSW, as discussed in Section I of the Supplementary Material, having a positive estimated alpha is a necessary condition for a fund to be selected as outperformer.

in terms of both bias and variance, and an increased power in detecting outperforming funds. We are illustrating the prominent power of the fFDR$^+$ via a set of simulation studies in the following sections. In the empirical section, we will show the actual profitability that the fund characteristics can bring to investors while controlling for luck.

III. Data

We use monthly mutual fund data from Jan. 1975 to Dec. 2022 collected from the CRSP database. As CRSP reports funds at the share class level, we use MFLINKS to acquire fund data at the portfolio level. For a fund at a given point in time with multiple share classes, we average the share classes' net return weighted by the total net asset (TNA) value at the beginning of the month.[11] The TNA at the fund level is estimated by the sum of the share classes' TNA. We omit the following month return after a missed return observation as CRSP fills this with the accumulated returns since the last nonmissing month. To obtain the holdings data of the funds, which will be used to calculate our covariates, we merge the CRSP and Thomson/CDA databases by utilizing MFLINKS. The holdings database provides us with stock identifiers, which we use to link the funds' position with the CRSP equity files. From this equity database, we obtain information such as the price and number of shares outstanding of the stocks that the funds hold on their reported portfolio date. We use these to calculate the return gap and the active weight, which are described in more detail later.

We consider only funds with an investment objective belonging to the categories growth, aggressive growth, and growth and income. Both CRSP and CDA provide this information; CDA is more consistent over time; hence, we choose that. As the funds' investment objective can change, we collect all the funds in these categories. To obtain a precise 4-factor alpha estimate, we select only funds with at least 60 monthly observations. Overall, we gather a sample of 2,291 funds, which provides the empirical metrics for our simulation study.

In the empirical part, when calculating the related covariates, we additionally require each fund to hold at least 10 stocks; this is consistent with Kacperczyk et al. (2008) and Doshi et al. (2015) and is needed here as we use the return gap and active weight from their studies as 2 of our covariates. The number of funds used when constructing our covariate-based portfolios varies over years and will be reported in detail in the empirical section.

IV. Simulation Setup

In this section, we present the details of our simulation design consisting of the choice of the model, the distributions of the alpha population, the data-generating process, and the metrics that we will use to gauge the performance of the methods.

[11]Since 1991, we use the monthly TNA of the fund's share classes. Before 1991, most of the funds report their TNA on a quarterly basis. For this, we follow Amihud and Goyenko (2013) to fill in the missing TNA of each fund (at the share class level) by its most recently available one.

A. The Model

Following the majority of the existing literature on mutual fund performance, we use the 4-factor model of Carhart (1997) to compute a fund's performance:

$$r_{i,t} = \alpha_i + b_i r_{m,t} + s_i r_{smb,t} + h_i r_{hml,t} + m_i r_{mom,t} + \varepsilon_{i,t}, \quad i = 1, \ldots, m, \tag{10}$$

where $r_{i,t}$ is the excess net return of fund i over the risk-free rate (i.e., the 1-month treasury bill rate), $r_{m,t}$ is the market's excess return on the CRSP New York Stock Exchange (NYSE)/American Express (AMEX)/National Association of Securities Dealers Automated Quotations (NASDAQ) value-weighted market portfolio, $r_{smb,t}$ is the Fama–French small minus big factor, $r_{hml,t}$ is the high minus low factor, $r_{mom,t}$ is the momentum factor, and $\varepsilon_{i,t}$ is the noise of fund i at time t. All factors and the 1-month treasury bill rate are obtained from French's website.

Our simulations are designed similarly to BSW and AP in terms of the data-generating process accounting, in addition, for an informative covariate and considering more distribution types of the fund alpha population. Whereas BSW and AP focus on the estimated proportions of the outperforming, underperforming, and zero-alpha funds, we consider the performance of the FDR^+ and fFDR^+. More specifically, for a given fund alpha distribution, we first generate in each iteration the true fund alpha population and a covariate that conveys information about the alpha of each fund. Second, we simulate the Fama–French factors (factor loadings) by drawing from a normal distribution with parameters equal to their sample counterparts (obtained from estimations of model (10)). Next, the noise is generated under both cross-sectional independence and dependence. In the first case, the noise is drawn cross-sectionally independent from a normal distribution (i.e., $\varepsilon_{it} \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right)$), where, as in Barras, Scaillet, and Wermers (2019), σ_ε is set equal to the median of its real-data counterpart (i.e., approximately 0.0183 for our sample). The results under this assumption are reported in the Section IV.B. In the dependent case, the noise is generated as in BSW and the simulation results are deferred to Section IV.A of the Supplementary Material. The simulated data are then used to generate the net return for each fund.[12] Subsequently, by carrying out regression (10) of the generated net return on the simulated Fama–French factors, we estimate the alpha and calculate the related p-values for the tests for equation (2). Finally, based on these estimated alphas, p-values, and the covariate, we implement the fFDR^+ and FDR^+, for a given FDR target, to identify the significantly outperforming funds. We estimate the actual FDR of the fFDR^+ and check whether it meets the given target. We then compare the two methods in terms of power, defined as the expected ratio of the number of true-positive-alpha funds detected to the total number of true-positive-alpha funds in the population.

B. The Distribution of Fund Alphas

We consider 3 different types for the distribution of fund alphas: a discrete, a discrete–continuous mixture, and a continuous. A covariate Z conveys information about the alpha of each fund in the population; more specifically, a fund with $Z = z$ has

[12] We consider both balanced and unbalanced panel data. For the interest of space, the simulation results of the unbalanced panel data case are deferred to Section IV.B of the Supplementary Material.

a probability $\pi_0(z)$ of being zero alpha. Also, without loss of generality, we assume that, for nonzero-alpha funds, their covariates and alphas are positively correlated.[13]

First, in the discrete type, we draw alphas from 3 mass points $-\alpha^* < 0$, 0, and $\alpha^* > 0$ with probabilities π^-, π_0, and π^+. Thus,

$$\alpha \sim \pi^- \delta_{\alpha=-\alpha^*} + \pi_0 \delta_{\alpha=0} + \pi^+ \delta_{\alpha=\alpha^*}. \tag{11}$$

We consider 5 values for $\alpha^* \in \{1.5, 2, 2.5, 3, 3.5\}$ (the values are annualized and in %) together with 6 combinations of the proportions (π^+, π_0, π^-) based on $\pi^+ \in \{0.1, 0.13\}$, $\pi^-/\pi^+ \in \{1.5, 3, 6\}$, and $\pi_0 = 1 - \pi^- - \pi^+$ (i.e., a total of 30 cases).[14]

In the mixed discrete–continuous distribution, we draw alphas from 2 components including the mass point 0 and the normal distribution $\mathcal{N}(0, \sigma^2)$ with, respectively, probabilities $\pi_0 \in (0,1)$ and $1 - \pi_0$. We have, therefore, that

$$\alpha \sim \pi_0 \delta_{\alpha=0} + (1 - \pi_0)\mathcal{N}\left(0, \sigma^2\right). \tag{12}$$

We consider 5 values for $\sigma \in \{1, 2, 3, 4, 5\}$ (the values are annualized and in %) and the same 6 π_0 values as in the discrete distribution earlier.

Finally, in the continuous case, we draw alphas from a mixture of two normal distributions $\mathcal{N}\left(\mu_1, \sigma_1^2\right)$ and $\mathcal{N}\left(\mu_2, \sigma_2^2\right)$ with, respectively, probabilities $\pi_1 \in [0,1]$ and $\pi_2 = 1 - \pi_1$, i.e.,

$$\alpha \sim \pi_1 \mathcal{N}\left(\mu_1, \sigma_1^2\right) + \pi_2 \mathcal{N}\left(\mu_2, \sigma_2^2\right). \tag{13}$$

When π_1 and π_2 are positive, we have indeed a mixture; we adopt from Harvey and Liu (2018) $\pi_1 = 0.3$ and $\pi_2 = 0.7$, and to point up the performance of our method, we consider 15 combinations based on $(\mu_1, \mu_2) \in \{(-2.3, -0.7), (-2, -0.5), (-2.5, 0)\}$ and $(\sigma_1, \sigma_2) \in \{(1, 0.5), (1.5, 0.6), (2, 1), (2.5, 1.25), (3, 1.5)\}$ (the values of the pairs are annualized and in %).[15]

In equation (13), $\pi_0 = 0$, whereas in equations (11) and (12), $\pi_0 > 0$. When $\pi_0 > 0$, we study an up-and-down shape of $\pi_0(z)$. Specifically, to guarantee $\pi_0(z) \in [0,1]$ for all z, we choose

$$\pi_0(z) = \min\{1, \max(f(z), 0)\} \in [0,1],$$

where

$$f(z) = 3.5(z - 0.5)^3 - 0.5(z - 0.5) + c \tag{14}$$

and c is chosen to satisfy $\int_0^1 \pi_0(z)dz = \pi_0$. This way we are able to investigate the effect of π_0 on the power of the methods by varying c while keeping the shape of $\pi_0(z)$ roughly unchanged.[16]

[13] If the correlation is negative, we use $-Z$ instead.

[14] The chosen π^+ values are close to those used in the recent literature: $\pi^+ = 10.6\%$ (see Harvey and Liu (2018)) and $\pi^+ = 13\%$ (see Andrikogiannopoulou and Papakonstantinou (2016)). The ratio $\pi^-/\pi^+ = 6$ is studied in AP. Aiming to extend the range of our study, we consider also the ratios 1.5 and 3.

[15] Our choices are intended to be wide enough to encompass the cases of Harvey and Liu (2018): $(\pi_1, \pi_2) = (0.283, 0.717)$, $(\mu_1, \mu_2) = (-2.277, -0.685)$, and $(\sigma_1, \sigma_2) = (1.513, 0.586)$. In Section IV.C of the Supplementary Material, we additionally present results of the case $\pi_2 = 0$, i.e., when the mixture becomes a single normal distribution.

[16] In Section IV.D of the Supplementary Material, we show that the alternative choices of a decreasing function $\pi_0(z)$ with $f(z) = -1.5(z - 0.5)^3 + c$, an increasing function $\pi_0(z)$ with

Suppose the distribution of alpha and the form of $\pi_0(z)$ are determined. We enerate the covariate vector $(z_1, z_2, \ldots, z_m)$ with each element drawn from the uniform distribution $[0,1]$ and assign them to the funds satisfying the descriptions mentioned at the beginning of this section. The noise in equation (10) is generated cross-sectionally independent or dependent. In the former case, it is drawn from a normal distribution $\mathcal{N}(0, \sigma_\varepsilon^2)$, where, as in Barras et al. (2019), σ_ε is set equal to the median of its real-data counterpart (i.e., approximately 0.0182 for our sample). For each replication, we implement the fFDR$^+$ and FDR$^+$ and compute the rate of falsely selected funds among those classified as outperformers and the rate of truly outperforming funds detected. The 2 metrics are averaged across 1,000 replications to obtain estimates for the actual FDR and the power of each procedure.[17]

V. Analysis of False Discovery Rate and Power

We set the number of funds for simulations at 2,000, which is close to our sample of 2,291 funds. We demonstrate the ability of the fFDR$^+$ to control FDR for balanced panel data, where the number of observations per fund is equal to 284, under cross-sectional independence. For the interest of space, we refer to Sections IV.A and IV.B of the Supplementary Material for the results under cross-sectional dependence as well as the unbalanced panel data cases. We then compare the powers of the fFDR$^+$ and the FDR$^+$ in controlling the FDR at the 10% level; we extend to higher levels and highlight the differences between the two procedures. In each simulation study, we analyze the relationship between the powers of the two methods: i) the proportion of zero-alpha funds in the sample and ii) the magnitude and proportion of positive-alpha funds in the sample. We also study the impact of the number of funds in the sample and the number of observations per fund on the power. Finally, we examine the impact of estimation errors in the covariates, in the power of our procedure.

In general, the fFDR$^+$ controls well the FDR at any given targets. When the FDR target is set at 10%, the fFDR$^+$ detects more positive-alpha funds than the FDR$^+$ with a difference in power up to 30%, depending on cases and parameters of the distributions.[18] When we raise the FDR target to higher levels, the difference is even higher in favor of the fFDR$^+$. The results are consistent regardless of the number of funds in the sample, the structure of the panel data, and the dependence of the cross-sectional error terms.

In an empirical setting, the fund characteristics are estimated quantities. This is translated to an estimation noise that may affect the power of our procedure. Our simulations reveal that our method is robust in terms of power up to moderate-to-high estimation noise.

$f(z) = 1.5(z-0.5)^3 + c$, or a constant function $\pi_0(z) = c$ result in some discrepancies, without affecting, though, our main conclusions.

[17]We refer to Section II of the Supplementary Material for a detailed description of the simulation procedure.

[18]In Section V of the Supplementary Material, we additionally study the fFDR$^+$ with the use of an *non-informative* covariate, which is a covariate generated randomly and independently from the tests. We find that the fFDR$^+$ controls well for FDR and its power is similar to that of the FDR$^+$.

A. False Discovery Rate Control of the fFDR$^+$

For varying targets of FDR $\in \{5\%, 10\%, \ldots, 90\%\}$, we implement the simulation procedure in Section IV with balanced panel data. Figures 1, 2, and 3 exhibit our results for the generated data under cross-sectional independence.

In Figure 1, we show our results for the discrete distribution (11) for varying α^*. Graphs A–C correspond to $\pi^+ = 0.1$, whereas Graphs D–F correspond to $\pi^+ = 0.13$. From left to right, the ratio π^-/π^+ increases from 1.5 to 6 (with the null proportion π_0 decreasing accordingly). For example, Graph A exhibits the actual FDR (vertical axis) and the given targets of FDR (horizontal axis) with the alphas drawn from a discrete population of which 75%, 10%, and 15% are, respectively, zero-, positive-, and negative-alpha funds. A point on or below the 45° line indicates that the fFDR$^+$ controls FDR well for the given level; this is the case for $\alpha^* = 1.5$ at all FDR targets. For $\alpha^* = 3.5$, FDR is slightly not met for targets in the interval (0.1, 0.8). In general, we witness slight failure of the fFDR$^+$ to control for FDR when α^* is abnormally high. In the last case with smallest π_0, FDR is controlled well. In Figure 2, we study the case of the fund alpha population described by the mixed discrete–continuous distribution (12). We organize our results based on the same null proportions π_0 as in Figure 1 and present these for varying σ. We observe that the FDR target is slightly unmet only for extreme values of σ when the null proportion is very high and this effect

FIGURE 1

Performance of the **fFDR**$^+$ for Discrete Distribution of $\boldsymbol{\alpha}$

Figure 1 shows the performance of the fFDR$^+$ in terms of FDR control when alphas are drawn from a discrete distribution. The simulated data are balanced panels with cross-sectional independence.

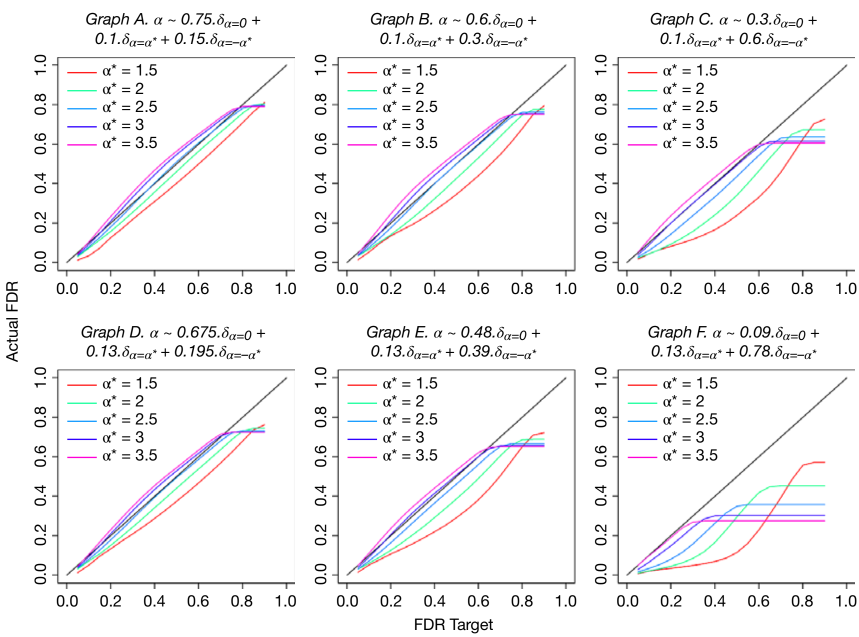


FIGURE 2

Performance of the **fFDR**$^+$ for Discrete and Normal Distribution Mixture of $\boldsymbol{\alpha}$

Figure 2 shows the performance of the fFDR$^+$ in terms of FDR control when alphas are drawn from a mixture of discrete and normal distributions. The simulated data are balanced panels with cross-sectional independence.

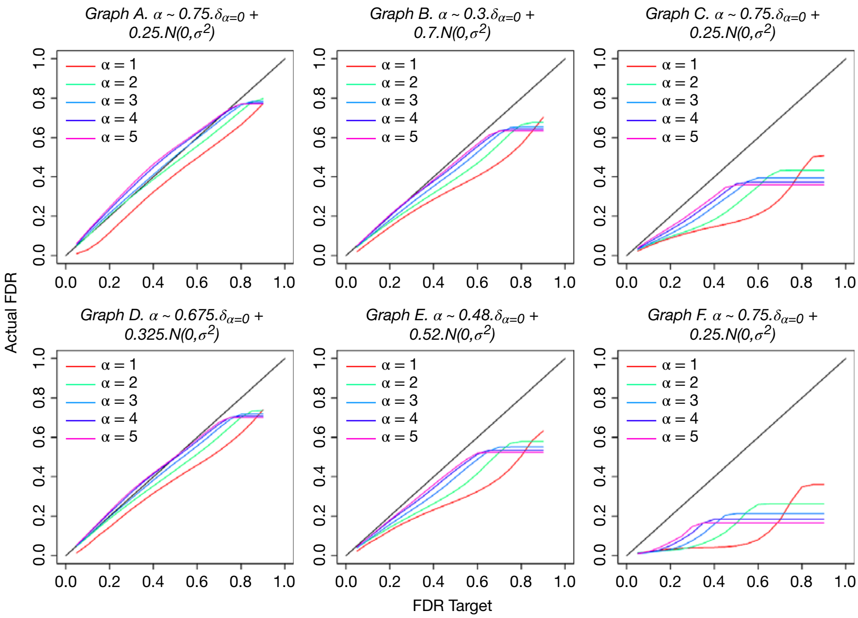
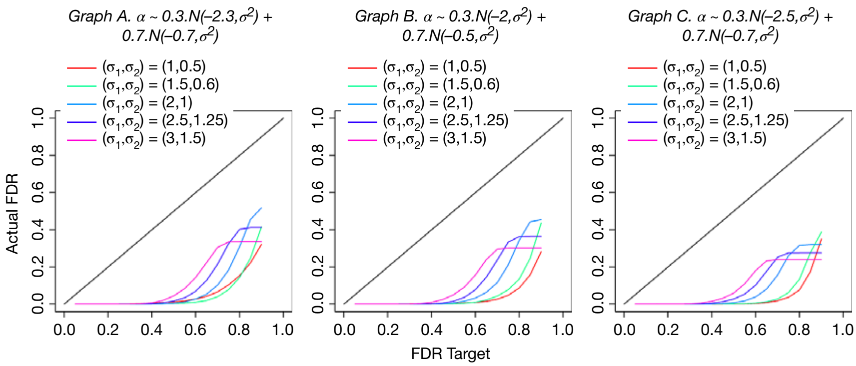


FIGURE 3

Performance of the **fFDR**$^+$ for Continuous Distribution of $\boldsymbol{\alpha}$

Figure 3 shows the performance of the fFDR$^+$ in terms of FDR control when alphas are drawn from a continuous distribution, which is a mixture of 2 normals. The simulated data are balanced panels with cross-sectional independence.



is also milder compared to the discrete distribution cases. Finally, in Figure 3, we report the results for the continuous distribution (13) for varying μ or (μ_1,μ_2) and σ or (σ_1,σ_2). We find that the fFDR$^+$ controls FDR well at all targets.

In summary, our simulations are based on proposed fund alpha distributions from the recent literature, from the least realistic cases, with all the outperforming and underperforming funds assumed to have the same mass alpha value, to more realistic ones, where the alpha is drawn from a continuous distribution, in which no fund has exact 0 but rather mostly negative alpha. Our results suggest that, for the continuous distribution, the proposed fFDR^+ approach controls well for FDR at any given target. In Section III of the Supplementary Material, we show that the variance of the estimated actual FDR of the fFDR^+ is smaller than that of the FDR^+. This means that the reported estimated actual FDR curves of the fFDR^+ are less varying than those of the FDR^+. In other words, if the estimated actual FDR of the two methods is the same and lies below or on the 45° line, there is less chance that the actual FDR of the fFDR^+ lies above the 45° line than that of the FDR^+.

In Sections IV.A and IV.B of the Supplementary Material, we repeat the exercise for balanced data under cross-sectional dependence and unbalanced data under both cross-sectional independence and dependence. Our findings remain robust.

B. Power Analysis

Next, we study the power of our fFDR^+ approach in detecting truly positive-alpha funds, calculated as described in Section IV, and compare it with the FDR^+ of BSW for FDR control at 10%. Although the magnitude of our results varies with different FDR targets, our main conclusion of the power superiority of the fFDR^+ remains.

In Panel A of Table 1, we report the discrete distribution (11). For $(\pi^+, \pi_0, \pi^-) = (10,75,15)\%$ with highest π_0 and smallest $\alpha^* = 1.5$, both the fFDR^+ and FDR^+ achieve similar powers (i.e., 1% and 0.6%, respectively). This is expected in this particular case as the number and magnitude of the true-positive alphas are small, while we are controlling for FDR at 10%.[19] The superiority of the fFDR^+ is more perceptible and stabler for larger α^*. This discrepancy depends not only on the magnitude and proportion of positive alphas but also on the proportion of zero alphas. This is because both procedures use the null proportion (π_0 in the FDR^+ and $\pi_0(z)$ in the fFDR^+) to estimate FDR. With the same magnitude and proportion of positive alphas, the small proportion of zero alphas implies the higher power of both the fFDR^+ and FDR^+. The effect of the null proportion on the gap of the fFDR^+ over the FDR^+ is stronger when the magnitude of positive alphas is not too high. The gap varies by case and may even exceed 30% (when $\pi^+ = 10\%, \pi_0 = 30\%$, and $\alpha^* = 2.5$).[20]

Panel B of Table 1 exhibits the power upshots for the case of the fund alpha population described by the distribution mixture (12). This implies the dependence

[19]As will be shown later, with a higher FDR target (such as 30%), the power of the fFDR^+ exceeds that of the FDR^+ by 6%. Considering a higher target than 10% is sensible for trading and diversification purposes as otherwise very few or no outperforming funds are selected. In the study of BSW, the estimated FDR in the empirical application is at least 41.5% on average (depending on portfolio).

[20]As shown in Section IV of the Supplementary Material, the relevant reports vary slightly when the simulated data are generated with alternative forms of $\pi_0(z)$ mentioned in footnote 16, with unbalanced panel, or with cross-sectional dependence; however, the overall picture remains the same.

TABLE 1

Performance Comparison in Terms of Power (%)

Table 1 compares the power of the $fFDR^+$ and FDR^+ at FDR target of 10% when the alphas of 2,000 funds are drawn from a discrete distribution (i.e., $\alpha \sim \pi^+\delta_{\alpha=\alpha^*} + \pi_0\delta_{\alpha=0} + \pi^-\delta_{\alpha=-\alpha^*}$ (Panel A)), a discrete–normal distribution mixture (i.e., $\alpha \sim \pi_0\delta_{\alpha=0} + (1-\pi_0)\mathcal{N}(0,\sigma^2)$ (Panel B)), and a mixture of two normal distributions (i.e., $\alpha \sim 0.3\mathcal{N}(\mu_1,\sigma_1^2) + 0.7\mathcal{N}(\mu_2,\sigma_2^2)$ (Panel C)) with various setting of parameters. The simulated data are a balanced panel with 284 observations per fund and are generated with cross-sectional independence.

Panel A. Discrete Distribution

(π^+,π_0,π^-)	Procedure	$\alpha^*=1.5$	$\alpha^*=2$	$\alpha^*=2.5$	$\alpha^*=3$	$\alpha^*=3.5$
$(10,75,15)\%$	$fFDR^+$	1	6.8	23.9	46.6	68.7
	FDR^+	0.6	2.9	13.9	33.6	55.3
$(10,60,30)\%$	$fFDR^+$	2	12.6	35.5	59.6	77.8
	FDR^+	0.5	3.4	16.2	37.7	58.5
$(10,30,60)\%$	$fFDR^+$	5.5	26	54	77.6	90.2
	FDR^+	0.6	5.3	23.3	49.9	71.3
$(13,67.5,19.5)\%$	$fFDR^+$	1.8	11.5	32.8	56.7	76.7
	FDR^+	0.7	5	19.9	41.7	62.8
$(13,48,39)\%$	$fFDR^+$	3.8	19.3	44.6	70	85.1
	FDR^+	0.7	5.5	23.5	48.5	68.3
$(13,9,78)\%$	$fFDR^+$	9.7	37.6	70.7	91.5	97.8
	FDR^+	0.9	10	41	73.4	89.8

Panel B. Discrete–Normal Distribution Mixture

π_0	Procedure	$\sigma=1$	$\sigma=2$	$\sigma=3$	$\sigma=4$	$\sigma=5$
75%	$fFDR^+$	0.6	16.8	37.3	51.8	61.3
	FDR^+	0.3	9.2	27.7	42.4	52.9
60%	$fFDR^+$	1.8	22.6	44.2	58.1	67.2
	FDR^+	0.4	12.3	32.8	47.5	57.8
30%	$fFDR^+$	5.1	32.9	54.9	68.1	75.5
	FDR^+	0.6	18.7	41.3	56.5	66.1
67.5%	$fFDR^+$	1.1	20.1	40.9	55.3	64.2
	FDR^+	0.3	11	30.4	45.3	55.7
48%	$fFDR^+$	3.2	27.9	49.1	62.8	71.6
	FDR^+	0.4	15.4	36.4	51.5	61.4
9%	$fFDR^+$	7.5	39.8	62.2	74.6	81.4
	FDR^+	0.9	23.5	48.7	63.9	73.1

Panel C. Mixture of Two Normal Distributions

(μ_1,μ_2)	(σ_1,σ_2) Procedure	$(1,0.5)$	$(1.5,0.6)$	$(2,1)$	$(2.5,1.25)$	$(3,1.5)$
$(-2.3,-0.7)$		$\pi^+=6\%$	$\pi^+=10.4\%$	$\pi^+=20.7\%$	$\pi^+=25.5\%$	$\pi^+=29.1\%$
	$fFDR^+$	0.1	0.5	5.8	14.4	24.5
	FDR^+	0	0	0.4	2.4	8.1
$(-2,-0.5)$		$\pi^+=11.8\%$	$\pi^+=16.9\%$	$\pi^+=26.4\%$	$\pi^+=30.5\%$	$\pi^+=33.4\%$
	$fFDR^+$	0.1	0.7	7	16.5	26.5
	FDR^+	0	0	0.6	3.6	10.1
$(-2.5,0)$		$\pi^+=35.2\%$	$\pi^+=36.4\%$	$\pi^+=38.2\%$	$\pi^+=39.8\%$	$\pi^+=41.1\%$
	$fFDR^+$	0.5	1.1	9.9	19.3	29.4
	FDR^+	0	0.1	1.1	5.1	12.7

of the proportion and magnitude of positive alphas on the proportion of zero-alpha funds and the σ value for nonzero alphas. We expect a higher power for both methods for a smaller zero-alpha proportion and/or a higher value of σ. We find that the $fFDR^+$ is more powerful than the FDR^+. More specifically, for the balanced data under cross-sectional independence and $\pi_0 = 75\%$, the power of the $fFDR^+$ (FDR^+) increases from 0.6% to 61.3% (0.3% to 52.9%) with increasing σ from 1 to 5. For given, say, $\sigma = 2$, the power of the $fFDR^+$ (FDR^+) increases from 16.8% to 39.8% (9.2% and 23.5%) with reducing π_0. The gap is generally evident for $\sigma > 1$ with power differences around 10%, but it can also reach up to 16%.

https://doi.org/10.1017/S0022109024000097 Published online by Cambridge University Press

Finally, in Panel C of Table 1, we study the outcome from using the mixture of normals as in equation (13), with $\pi_1 = 0.3$ and $\pi_2 = 0.7$ and nonpositive means (μ_1, μ_2) to limit the likelihood of a positive alpha. The proportion of positive alphas ranges from 6% to 41.1%. For small (σ_1, σ_2) values, the positive alphas are also small in magnitude, and consequently, the power is negligible. When (σ_1, σ_2) are higher than (2,1), the power of both methods and their discrepancy increase significantly and favorably for the fFDR$^+$ reaching up to 16%.

Our analysis has shown that, when controlling for FDR at an as low level as 10%, both fFDR$^+$ and FDR$^+$ have good power for large (in magnitude) alphas. When this happens, the gain in power of the fFDR$^+$ over the FDR$^+$ can vary depending on the underlying fund alpha distribution: 10% to 16% (continuous distribution) and 20% to 30% (discrete distribution). On the other hand, when the zero-alpha proportion is high and the proportion and magnitude of positive alphas are small, the power of both methods reduces.

Finally, we demonstrate in Section IV.A and IV.B of the Supplementary Material that our conclusions are not affected by the data structure (balanced versus unbalanced panel) or dependencies.

C. Power and FDR Trade-Off

In what follows, we study the impact on power when controlling for FDR at different (higher than 10% level) targets. Our results show clear differences between the fFDR$^+$ and FDR$^+$ and, in support of the former, even for cases of negligible power for a 10% target. Constructing mutual fund portfolios at higher FDR levels is sensible as otherwise we may end up with empty portfolios. Investors have to face a trade-off between the power in detecting outperforming funds and the FDR threshold, which we will discuss next.

We focus on cases of very low power when the FDR is controlled at 10%. For brevity, we present in Table 2 our results for only balanced data under cross-sectional independence and FDR targets up to 90%, noting that these are largely unchanged for unbalanced data. In particular, for the underlying discrete fund alpha distribution, the fFDR$^+$ gains rapidly increasing power with increasing FDR targets, peaking at 38% in excess of the FDR$^+$ when the target is set at 70%. For the

TABLE 2

Power Comparison (in %) for Varying FDR Targets (%)

Table 2 presents some selected cases of low powers of the fFDR$^+$ and FDR$^+$ at FDR target of 10%. We consider a discrete distribution: $\alpha \sim 0.75\delta_{\alpha=0} + 0.1\delta_{\alpha=1.5} + 0.15\delta_{\alpha=-1.5}$; a discrete–normal mixture: $\alpha \sim 0.75\delta_{\alpha=0} + 0.2\mathcal{N}(0, 1.5^2)$; and a 2-normal mixture: $\alpha \sim 0.3\mathcal{N}(-2.3, 1^2) + 0.7\mathcal{N}(-0.7, 0.5^2)$. The simulated data are balanced panels with cross-sectional independence.

		FDR Target								
Distribution	Procedure	10	20	30	40	50	60	70	80	90
Discrete	fFDR$^+$	0.4	2.9	9	19.6	34	50.6	66	77.6	85.9
	FDR$^+$	0.4	1.1	2.2	4.3	8.4	15.6	27.4	45.6	68.5
Mixture of discrete and normal	fFDR$^+$	0.4	1.5	3.7	7.3	12.7	20.8	32.4	47.4	64.4
	FDR$^+$	0.2	0.5	0.7	1.2	1.9	3.3	5.7	11.4	27.4
Mixture of normals	fFDR$^+$	0	0.2	0.5	1.4	3.1	6.5	12.3	21.6	35.3
	FDR$^+$	0	0	0	0.1	0.1	0.1	0.1	0.1	0.2

continuous distribution, the power of the FDR$^+$ changes very slowly and is persistently negligible (mixture of normals) even for FDR controlled at 90%. On the other hand, the fFDR$^+$ detects abundant positive-alpha funds with a power that can reach up to 35% in excess of the FDR$^+$ (mixture of two normal distributions with 90% target).

In Section VI of the Supplementary Material, we conduct a set of simulations to investigate the impact of varying number of observations T per fund and the number of funds m on the power. We see that the power of the fFDR$^+$ increases at a much faster pace, compared to the FDR$^+$, with increasing T, and slightly decreases with rising m.

In Section VII, we design simulations to study the impact of using a covariate with noise on the power of the fFDR$^+$ by adding to the original covariate a noise reflecting potential estimation biases. We see that the fFDR$^+$ controls well for FDR and the power of the fFDR$^+$ is lower than that in Table 1, but still remarkably higher than that of the FDR$^+$.[21]

Concluding this section, we recollect that the simulated power of the fFDR$^+$ in detecting outperforming funds is found to be larger than the FDR$^+$s. This persists for different fund alpha distributions, balanced and unbalanced data, and cross-sectional dependence of error terms accounted for or not. This power advantage depends on the magnitude and proportion of positive alphas as well as the proportion of zero alphas in the population, the number of funds in the sample, estimation errors in the covariates (fund characteristics), and the average number of observations per fund. Especially when the last factor is small, leading to a diminished power for both procedures, we can recover that for the fFDR$^+$ by uplifting the FDR level. In our empirical application of Section VI, we show how investors can benefit from this.

VI. Empirical Results

A. The **FDR**$^+$ and **fFDR**$^+$ Portfolios

In this section, we illustrate how the fFDR$^+$ helps identify outperforming mutual funds using a portfolio approach following BSW. We use as covariates six fund characteristics including the R^2 of Amihud and Goyenko (2013), the fund size of Harvey and Liu (2017), the return gap of Kacperczyk et al. (2008), the active weight of Doshi et al. (2015), the fund flow, and the expense ratio. For the interest of space, we refer the details of the construction of the 6 covariates to Section VIII of the Supplementary Material. In addition to the aforementioned well-known covariates, we propose 4 new covariates that are based on asset pricing models and are available for all funds in our sample. These are the Sharpe ratio, the beta and Treynor ratio obtained from the CAPM, and the idiosyncratic volatility (sigma) of the Carhart 4-factor model. The Sharpe and Treynor ratios are risk-adjusted performance measures of funds, whereas the beta and sigma reflect systematic and idiosyncratic risk, respectively. These metrics reveal aspects of the past mutual

[21] In Section V of the Supplementary Material, we further show that if a non-informative is used instead, the fFDR$^+$ controls well FDR and gains the same power as the FDR$^+$.

funds' performance and, thus, may assist in identifying outperforming and underperforming funds. Asset pricing metrics are regularly used by wealth managers and academics in the fields of trading, asset pricing, and investors' performance, but are overlooked in the mutual funds' literature.[22]

Similar to BSW, at the end of year t, we select a group of funds to invest in year $t+1$ based on historical information from the last 5 years ($t-4$ to t). In order to implement the fFDR$^+$ and FDR$^+$, we require the observed values of the covariates of each fund, the estimated alpha, and the p-value of each test. We execute, first, the Carhart 4-factor model over the 5-year period to estimate the alpha.[23]

The informative value of the return gap, active weight, fund flow, and fund size on funds' performance is persistent: The choice between using the most recent (final-year) observations for these covariates or their average values over the whole in-sample (5 years) is of less importance, as demonstrated by our robustness check in Section XI of the Supplementary Material.[24] Although the predictability of the covariates may last for a long horizon of up to 5 years, we expect their informative values to decrease with time; hence, forming portfolios based on their recent realizations is preferred to their average values of the whole last 5 years' time. Because of this, return gap, active weight, fund flow, fund size, and expense ratio are calculated based on data in the final year of the in-sample (i.e., we use the exposure of the fund flow in year t for the fund flow and the value at the end of year t for the fund size, whereas for the active weight and the return gap we use their average exposures in year t). The R^2, Sharpe ratio, beta, sigma, and Treynor ratio are based on the whole 5 years. We note that each of the covariates is converted into interval $[0,1]$ via the formula described in Section II.A.

We calculate our p-values in a similar fashion to BSW. For the funds that suffer from heteroscedasticity or autocorrelation, we calculate the t-statistics based on the heteroscedasticity and autocorrelation-consistent standard deviation estimator of Newey and West (1987).[25] For each fund, we implement 1,000 bootstrap replications to estimate the distribution of the t-statistic and subsequently calculate the bootstrapped p-value for the fund.[26]

As required by our method, the p-values of any truly zero-alpha funds, given a covariate value, should be uniformly distributed. Although it is difficult for us to validate this requirement in reality as we never know which funds are truly zero alpha, it appears intuitive for us to assume that this condition is satisfied. Consider, for example, the R^2. We expect the truly zero-alpha funds to invest randomly in the

[22]For instance, Clifford, Fulkerson, Jame, and Jordan (2021) study the relationship between idiosyncratic volatility and mutual funds' flows, but they do not focus on using this fund characteristic as a factor for fund selection.

[23]In Section IX of the Supplementary Material, we further validate the performance of the methods with the use of simulated data (i.e., the return and 10 covariates) resembling the real sample.

[24]Readers may refer to Kacperczyk et al. (2008), Doshi et al. (2015), Zheng (1999), and Harvey and Liu (2017) for the studies of the persistence of the return gap, active weight, fund flow, and fund size, respectively. It should also be noted that in our fFDR framework, all covariates are transformed to uniform with only the ranking of the covariates across the fund counting.

[25]We check heteroscedasticity, autocorrelation, and autoregressive conditional heteroscedasticity (ARCH) effect by using the White, Ljung–Box, and Engle tests, respectively. We see that half of the funds in our sample suffer from at least one of the mentioned effects.

[26]The bootstrapping procedure may result in duplicated bootstrapped p-values. For this, we use an adequate number of replications to reduce that effect and obtain good estimates of $\pi_0(z)$ and $f(p,z)$.

stock market; thus, they should possess an R^2 value of roughly equal to 1. Conditional on a specific R^2 value that a truly zero-alpha fund could have (i.e., close to 1), if the fund is truly zero alpha, then its p-value should follow a uniform distribution like any usual true null hypothesis test.[27]

Next, we describe the selection process of outperforming funds to invest in year $t+1$ given a FDR target τ in $(0,1)$. First, we recall the relevant selection process for BSW's "FDRτ" portfolio. For each γ on the grid $\{0.01, 0.02, \ldots, 0.6\}$, we calculate the $\widehat{\text{FDR}}_\gamma^+$ given by equation (4) in the Supplementary Material. Then, we find γ^* such that $\widehat{\text{FDR}}_{\gamma^*}^+$ is closest to τ; this is the significant threshold for BSW's portfolio that is, all the positively estimated alpha funds in the in-sample window with p-values $\leq \gamma^*$ will be included in the FDRτ portfolio. This guarantees the nonempty property of the portfolio but does not always meet the FDR target τ, thereby $\widehat{\text{FDR}}_{\gamma^*}^+$ may be much higher than τ.

Second, we select outperforming funds for a fFDR-based portfolio, namely "fFDRτ." To establish comparable fFDRτ and FDRτ portfolios, we implement the fFDR$^+$ (with a particular covariate) to control pFDR$^+$ at a target τ^* that reflects the FDR level controlled by the FDRτ portfolio but has to be less than 1.[28] As the FDR of the FDRτ portfolio is controlled at level $\widehat{\text{FDR}}_{\gamma^*}^+$, which may be greater than 1 or less than τ, we set $\tau^* = \tau$ if $\widehat{\text{FDR}}_{\gamma^*}^+ \leq \tau < 1$ and $\tau^* = \widehat{\text{FDR}}_{\gamma^*}^+$ if $\tau < \widehat{\text{FDR}}_{\gamma^*}^+ < 1$.[29] If $\widehat{\text{FDR}}_{\gamma^*}^+ \geq 1$, we just select all the funds in the FDRτ portfolio.

Similar to BSW, for both fFDRτ and FDRτ portfolios, we invest equally in the selected funds in the following year. If a selected fund does not survive for a month during the year, then its weights are redistributed to the remaining (surviving) funds. As aforementioned, at the beginning of each year we select funds into a portfolio by using the previous 5 consecutive years as in-sample. To be eligible for this, a fund needs to have 60 observations in the in-sample. We start constructing our portfolios from Dec. 1981.

B. Performance Comparison

In this section, we assess the portfolios' performance based on their alphas. We demonstrate the advantage of the fFDR$^+$ in picking outperforming funds and the

[27] Indeed, the p-value of each test i is defined as $p_i = 1 - F(|t_i|)$, where $F(|t_i|) = \mathbb{P}(|\mathcal{T}_i| < |t_i| \,|\, \alpha_i = 0)$ and $\mathcal{T}_i$ is the conventional t statistic of test i and t_i is its estimated value. If hypothesis $\alpha_i = 0$ is true, conditional on a specific covariate value, the p-value of test i is uniformly distributed since $\mathbb{P}(P_i < p_i) = \mathbb{P}(1 - F(|\mathcal{T}_i|) < p_i) = \mathbb{P}\big(|\mathcal{T}_i| > F^{-1}(1-p_i)\big) = 1 - \mathbb{P}\big(|\mathcal{T}_i| < F^{-1}(1-p_i)\big) = 1 - F\big(F^{-1}(1-p_i)\big) = p_i$.

[28] If we implement the fFDR$^+$ and FDR$^+$ to strictly control FDR at a target, say, $\tau = 10\%$ or $\tau = 20\%$, both result in empty portfolios for many years. With BSW's FDRτ portfolios, the problem is solved. In BSW's study, for the FDR10% portfolio, the empirical $\widehat{\text{FDR}}_{\gamma^*}^+$ is always greater than 10% with an average of 41.5%. For our data, among the 38 times of portfolio construction, with target $\tau = 20\%$ (10%) the $\widehat{\text{FDR}}_{\gamma^*}^+$ is less than τ on 8 (0) occasions and greater than 1 on 5 occasions for both targets.

[29] We could have set $\tau^* = \widehat{\text{FDR}}_{\gamma^*}^+$ for both cases. However, it seems fairer to set $\tau^* = \tau$ if $\widehat{\text{FDR}}_{\gamma^*}^+ \leq \tau$ since both portfolios initially aim to control FDR at τ.

efficient use of the information contained in fund characteristics. We estimate the alpha evolution and the average alphas of our fFDRτ portfolios based on the 10 covariates and compare them with those of the FDRτ portfolio. We also explore the performance of the fFDRτ portfolios after linearly combining the 10 covariates and using their first principal component, an OLS regression, a LASSO, a ridge regression, and an elastic net.

We focus on portfolios with small FDR targets of $\tau = 10\%$. We repeat all estimations with $\tau = 20\%$ in Section X of the Supplementary Material. Our results remain unchanged for all exercises.

1. The Alpha Evolution

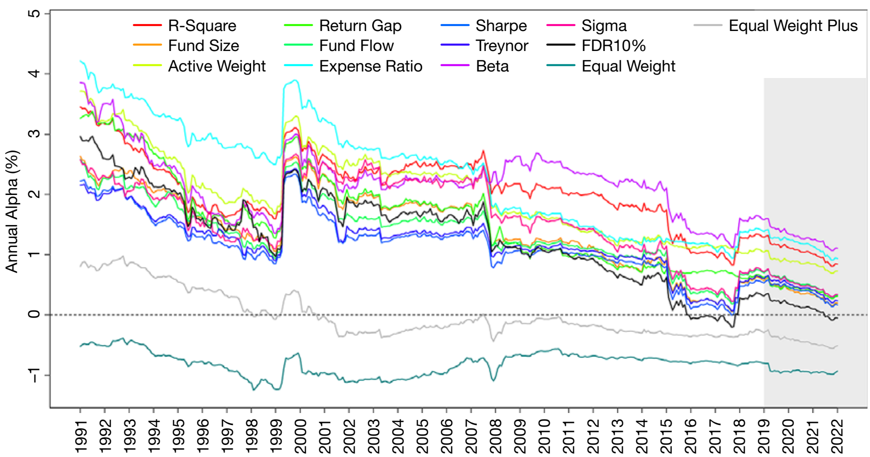
For each portfolio, we obtain its alpha evolution by calculating the Carhart 4-factor alpha using its returns from Jan. 1982 up to the end of each month from Dec. 1991 onward. In addition to the aforementioned portfolios, we construct 2 naive benchmark equal-weighted portfolios, without control for the FDR: one that simply includes all the mutual funds in the in-sample window to be invested in the following year and another that contains only those with positive estimated alphas. We name these 2 portfolios equal weight and equal weight plus.

We present all the alpha evolution in Figure 4. It is obvious that the FDR10% portfolio gains higher alphas than the equal-weighted portfolio and that all the fFDR10% portfolios outperform the FDR10%. Ultimately, at the end of 2022, the fFDR10% portfolio with the beta covariate is found to be the best with annualized alpha of about 1.1%, followed by the fFDR10% portfolios with the expense ratio, R^2, active weight, sigma, return gap, fund flow, Treynor ratio, fund size, and Sharpe ratio covariates achieving annualized alphas of at least 0.17%. By contrast, the portfolio FDR10%, without using the information of fund characteristics, winds up with a small negative alpha of −0.05%. It is noteworthy that the performance of the

FIGURE 4

Alpha Evolution of fFDR**10**% and FDR**10**% Portfolios over Time

Figure 4 presents the evolution of annualized alphas (in %) of the 10 fFDR10% portfolios corresponding to the 10 covariates, the FDR10% portfolio of BSW, and the 2 equal-weighted portfolios.



https://doi.org/10.1017/S0022109024000097 Published online by Cambridge University Press

fFDR10% and FDR10% portfolios is affected by the COVID-19 pandemic period, which is marked by shaded area. Prior to the event, we observe that all portfolios seem to rebounce and gain alphas ranging from 0.53% to 1.58% for the fFDR10% portfolios and 0.32% for the FDR10% one. For this reason, in the following, we report results of 2 samples: one ends in 2019 and one ends in 2022.

2. The Average Alpha

The alpha evolution in the Section VI.B.1 is calculated based on the portfolio returns from the start of 1982 up to a time point of interest. This may represent limited information in the case of investors with a different investment period of, say, 5 or 10 years. For this, in Panel A of Table 3, we report the average alpha that investors will gain if they invest for $n \in \{5,10,15,20,30,35,40,41\}$ consecutive years: For each portfolio, we calculate its "n-year" alpha based on the portfolio returns over a period of $12n$ consecutive months, and we repeat by shifting every time 1 month forward and eventually present the average alpha. We report the fFDR10% portfolio for each fund characteristic and the portfolio FDR10%. We note that the last case, $n = 41$, corresponds to the alphas for the whole period from Jan. 1982 to Dec. 2022, and, in this case, alphas are presented as the rightmost points in Figure 4. Panel B of Table 3 reports similar metrics but for period from 1982 to the end of 2019 and $n \in \{5,10,15,20,30,35,38\}$.

TABLE 3

Comparison of Portfolios' Alphas (in %) for Varying Time Lengths of Investing

In Table 3, we consider 11 portfolios including 10 fFDR10% portfolios corresponding to the 10 covariates and the FDR10% portfolio of BSW. We compare the average alphas of the portfolios that are kept in periods of exactly n consecutive years. For example, consider $n = 5$. For each portfolio, we calculate the alpha for the first 5 years based on the portfolios' returns from Jan. 1982 to Dec. 1986. Then, we roll forward by a month and calculate the second alpha. The process is repeated, and the last alpha is estimated based on the portfolios' returns from Jan. 2018 to Dec. 2022. The average of these alphas is presented in the first rows in Panel A of the table. Panel B reports similar metrics with the use of portfolios' return series from Jan. 1982 to Dec. 2019.

	fFDR10%										
n	R^2	Fund Size	Active Weight	Return Gap	Fund Flow	Expense	Sharpe	Treynor	Beta	Sigma	FDR10%
Panel A. Whole Sample											
5	0.95	0.10	0.78	0.57	0.19	1.37	−0.03	0.10	1.13	0.59	−0.06
10	0.88	0.01	0.72	0.41	0.30	1.07	0.08	0.18	1.25	0.49	−0.21
15	0.98	0.04	0.75	0.35	0.35	0.93	0.20	0.29	1.33	0.53	−0.23
20	1.21	0.26	0.93	0.50	0.50	0.96	0.39	0.44	1.45	0.80	−0.03
25	1.13	0.20	0.81	0.43	0.45	0.87	0.38	0.44	1.44	0.67	−0.06
30	0.88	0.10	0.62	0.33	0.36	0.74	0.29	0.36	1.30	0.41	−0.20
35	0.77	0.07	0.67	0.38	0.28	0.94	0.22	0.27	1.12	0.25	−0.23
40	0.76	0.12	0.65	0.22	0.24	0.90	0.17	0.22	1.01	0.26	−0.16
41	0.84	0.20	0.73	0.31	0.33	0.94	0.17	0.24	1.11	0.34	−0.05
Panel B. Sample Period Prior to Dec. 2019											
5	1.08	0.16	0.94	0.70	0.26	1.33	0.03	0.17	1.29	0.68	−0.01
10	1.13	0.23	0.91	0.56	0.48	1.12	0.30	0.40	1.55	0.72	0.02
15	1.38	0.43	1.07	0.67	0.64	1.18	0.51	0.59	1.58	0.97	0.16
20	1.52	0.61	1.22	0.80	0.77	1.21	0.65	0.72	1.70	1.16	0.30
25	1.27	0.41	0.91	0.56	0.60	0.95	0.51	0.59	1.62	0.85	0.12
30	1.08	0.30	0.78	0.51	0.51	0.85	0.44	0.52	1.54	0.55	−0.02
35	0.86	0.12	0.84	0.51	0.28	1.05	0.17	0.23	1.19	0.28	−0.21
38	1.29	0.57	1.05	0.62	0.70	1.39	0.53	0.60	1.58	0.72	0.32

In both panels of Table 3, we find that the fFDR10% portfolios outperform the FDR10% portfolio for all considered covariates and for all n. Although these results should be interpreted with caution (some covariates were not well known in the literature at the start of our sample, such as the active weight and the fund size, which were published in 2015 and 2017, respectively), they do indicate the stability of our approach for different investment horizons.

C. Combined Covariates

So far, we have considered the effect from the information brought in by each single covariate. In what follows, we explore the effect from combining the information from the different fund characteristics and the potential consequent performance improvement. More specifically, we create a new covariate given by the linear combination of the underlying fund characteristics. More specifically, for each fund i at time t, we have

$$\text{(15)}\quad \begin{aligned}\text{NEW\_COVARIATE}_{t,i} = {} & c_{1,t}R^2_{t,i} + c_{2,t}\text{ACTIVE\_WEIGHT}_{t,i} \\ & + c_{3,t}\text{RETURN\_GAP}_{t,i} + c_{4,t}\text{FUND\_SIZE}_{t,i} \\ & + c_{5,t}\text{FUND\_FLOW}_{t,i} + c_{6,t}\text{EXPENSE\_RATIO}_{t,i} \\ & + c_{7,t}\text{SHARPE\_RATIO}_{t,i} + c_{8,t}\text{TREYNOR\_RATIO}_{t,i} \\ & + c_{9,t}\text{SIGMA}_{t,i} + c_{10,t}\text{BETA}_{t,i}.\end{aligned}$$

We consider two approaches to estimate the coefficients $c_{1,t},\ldots,c_{10,t}$ in equation (15). First, we use as our new covariate the first principal component of all ten (standardized) fund characteristics. By transforming the fund characteristics into their principal components, their information about the performance of a fund is preserved and conveyed. We use the first principal component as it captures most of the variation of the covariates. Second, we use a linear model that regresses the fund returns for year k on the observed value of the covariates in year $k-1$, where $k \in \{t, t-1, t-2, t-3\}$. Then, we predict the return for year $t+1$ based on the estimated regression model and the covariates in year t. This is equivalent to using equation (15) with the regression's estimated coefficients as the $c_{1,t},\ldots,c_{10,t}$. Thereby, we consider 4 linear regression models including the OLS, the LASSO of Tibshirani (1996), and the ridge regression and the elastic net of Zou and Hastie (2005).[30]

Figure 5 exhibits the performance of the fFDRτ portfolios with the newly created covariates in terms of the alpha evolution.[31] We find that the portfolios based on the combined covariate obtained from the ridge and elastic net perform best among the combined covariates at $\tau = 10\%$.

In Table 4, we show the average n-year alphas of the fFDR10% portfolios from Jan. 1982 to Dec. 2022 (Panel A) and to Dec. 2019 (Panel B). The elastic net performs also better for all time lengths except the longest ones of each considering

[30]For each method (except OLS), the covariates are standardized before being used in the estimation. We use cross-validation to determine the parameters in the LASSO, ridge, and elastic net methods.

[31]There are a few years where LASSO and the elastic net shrink all the regression coefficients to 0. In these cases, the new covariate is equal to 0 for all funds, and to avoid an empty portfolio, we simply select all the funds in the FDRτ portfolio.

https://doi.org/10.1017/S0022109024000097 Published online by Cambridge University Press

FIGURE 5

Alpha Evolution of the fFDR10% Portfolios with Combined Covariates

Figure 5 shows the alpha evolution of the fFDR10% portfolios with each using a covariate obtained from either the principal component method or regression method; for the former, the covariate is the first principal component (PC 1) of the 5 covariates, whereas for the latter the new covariate is a linear combination of the 5 underlying covariates with the weights obtained based on one of the OLS, LASSO, ridge, and elastic net regressions.

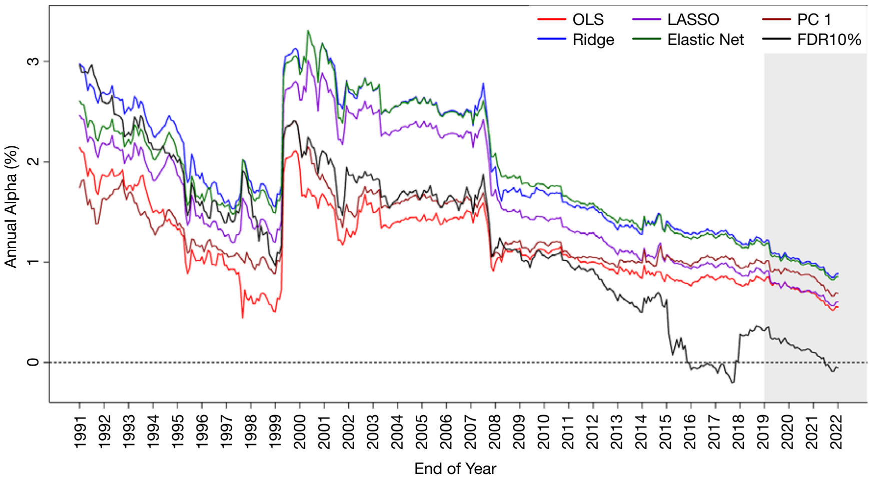


TABLE 4

Average *n*-Year Alpha of the fFDR10% Portfolios with Combined Covariates

Table 4 displays the average *n*-year alpha (annualized and in %) of the fFDR10% portfolios, which use covariates obtained by the first principal component (PC 1), the OLS, LASSO, ridge, and elastic net (see descriptions in Figure 5). The average *n*-year alpha of each portfolio is calculated as per the description in Table 3.

n	OLS	Ridge	LASSO	Elastic Net	PC 1
Panel A. Whole Sample					
5	0.49	1.07	0.89	1.14	0.89
10	0.55	1.05	0.82	1.15	0.95
15	0.68	1.03	0.80	1.18	0.96
20	0.90	1.19	0.96	1.31	1.07
25	0.83	1.10	0.86	1.24	1.02
30	0.68	0.97	0.70	1.07	0.92
35	0.69	0.99	0.69	1.02	0.97
40	0.51	0.84	0.55	0.82	0.71
41	0.55	0.89	0.61	0.86	0.69
Panel B. Sample Period Prior to Dec. 2019					
5	0.61	1.24	1.05	1.33	1.03
10	0.70	1.23	1.00	1.38	1.05
15	0.92	1.35	1.13	1.49	1.17
20	1.09	1.52	1.30	1.66	1.28
25	0.87	1.27	1.03	1.42	1.09
30	0.77	1.16	0.87	1.24	1.03
35	0.72	1.15	0.84	1.15	1.00
38	0.81	1.20	0.89	1.17	0.99

https://doi.org/10.1017/S0022109024000097 Published online by Cambridge University Press

the full sample periods. However, the best combined covariate does not beat the beta under the fFDR framework (as shown in Table 3).[32]

D. Comparison with Single- and Double-Sorting Portfolios

We also compare the performance of the portfolios formed in the fFDR framework with a traditional sorting portfolio formation. If a covariate has a highly linear relationship with the performance of mutual funds, then forming a portfolio based on sorting the funds on the covariate should be sufficient. We construct single- and double-sorting portfolios similar to Kacperczyk et al. (2008) and Doshi et al. (2015), and Amihud and Goyenko (2013), respectively. For the interest of space, we present the results in Section XV of the Supplementary Material. Thereby, the portfolios based on the fFDR gain positive alphas and beat the corresponding sorted portfolios in most cases. These results further validate the advantage of our method in exploiting the nonlinear relationship of fund characteristics, luck, and funds' performance. The inability of the traditional sorted portfolios that dominate the related literature to reflect the predictive value of the covariates under study is thus noteworthy.

In Section XVI of the Supplementary Material, we further examine a combination of the FDR^{+} procedure and covariates via constructing FDR10% portfolios in each quintile based on the covariates. We see that such combination cannot substitute our fFDR^{+} approach.

As further robustness checks, in Section XVII of the Supplementary Material, we demonstrate that our findings are robust with respect to a data subset where we require a minimum of 15 million in TNA for a fund to be considered.

In Section XVIII of the Supplementary Material, we construct a similar set of portfolios, namely $\mathrm{fFDR}^{-}\tau$, that aim to select underperforming funds. We see that these portfolios successfully pick the unprofitable funds and are consistently beaten by the equal-weighted portfolios.

VII. Concluding Discussion

In this article, we introduce the fFDR^{+}, a novel multiple hypothesis testing framework that incorporates fund characteristics to assess the conditional performance of mutual funds by controlling data snooping bias. We conduct simulation experiments to assess how well our method performs in controlling FDR and raising power compared to prior FDR methods. We then construct empirical portfolios based on our new method and use ten fund characteristics as informative covariates. We study six characteristics, which, based on earlier contributions, convey information about mutual funds' performance and propose four new ones based on asset pricing

[32]In Section XII of the Supplementary Material, we provide a detailed comparison of all the $\mathrm{fFDR}\tau$ portfolios in regard to several trading metrics, whereas in Section XIII of the same appendix the performance in terms of wealth evolution is presented. In Section XIV of the Supplementary Material, we further partition our sample into 4 nonoverlapping subperiods including the first 3 decades 1982–1991, 1992–2001, and 2002–2011 and the remainder. Overall, we see that all portfolios perform well in terms of alpha in the first 2 subperiods and then decline in the third subperiod. Two-thirds of the fFDR^{+} portfolios rebound in the remaining period up to 2019, but all portfolios are decreasing if the pandemic years are included. In terms of Sharpe ratio, however, all portfolios gain the highest reports in the final subperiod.

models. We show how the admixture of control for FDR and incorporated characteristics advances the generation of more positive and higher alphas than a portfolio that controls FDR only or a portfolio based on sorting on the covariate and the past funds' performance.

The implications of our study are both methodological and empirical. The methodological literature in the field of selecting outperforming mutual funds is rich and expanding—such as Kosowski et al. (2006), Andrikogiannopoulou and Papakonstantinou (2016), Harvey and Liu (2020), and Grønborg, Lunde, Timmermann, and Wermers (2021))—all these have their merits and present promising empirical findings. In our study, we focus on FDR, while we defer an examination of their power relative to ours to future research. Nevertheless, we ought to note 3 main distinguishing features of our method. First, it allows the use of more data in the form of fund characteristics, while the vast majority of others are limited to funds' past returns and their cross dependencies. Second, it is simple to implement and computationally less intensive than some of the most recent ones (e.g., the double bootstrap of Harvey and Liu (2020)). Third, our work can be extended to other problems in which statistical power weighs more than conservatism (i.e., the FDR threshold is higher), such as in the selection of hedge funds and bond funds or the assessment of trading strategies.

The empirical implications of our study are also of interest to academics and practitioners. We demonstrate that the 6 traditional mutual fund characteristics can offer important information. However, the relationship between these covariates, luck, and funds' performance is nonlinear. To fully exploit them, one should rely on powerful methods that control luck and noise. Our method ensures that we also introduce four new characteristics and find that their information in our context is important and surpasses that of traditional ones, a finding that is expected to be of interest to investment managers who are concerned with portfolio performance in a timely manner.

As with any methodological approach, there are caveats with our fFDR procedure. In particular, this requires large data sets and gains higher power as the FDR threshold increases (see Section V.C). This implies that our approach should not be applied in problems that require a small FDR target (i.e., when the risk of a false discovery can lead to disastrous outcomes). As in our context of mutual funds' performance, it is difficult to explore covariates that seem promising (see, e.g., the list of covariates studied in Jones and Mo (2021)) but with limited data availability.

Appendix. Estimating $\pi_0(\mathbf{z})$ and $f(p,z)$

Let $\{(p_i,z_i)\}_{i=1}^m$ be the collection of p-value and covariate realizations of the different funds under consideration, with $\{z_i\}_{i=1}^m$ transformed into uniform distribution $[0,1]$ (see Section II.A). We create fund bins $\{K_b\}_{b=1}^n$, where K_b contains a fund i if $z_i \in ((b-1)/n, b/n]$, and for each bin K_b, we estimate a common $\pi_0(z)$ for all the funds i in the bin. For some common $\lambda \in (0,1)$, we estimate the $\pi_0(z)$ in each bin b by

$$\hat{\pi}_{0,b}(\lambda) = \frac{\#\{i | p_i > \lambda,\ z_i \in K_b\}}{(1-\lambda)\#K_b}, \quad b = 1,2,\ldots,n. \tag{16}$$

We determine λ by minimizing the mean integrated square error (MISE):

(17) $$\text{MISE}(\lambda) = \text{BIAS}^2 + \text{VARIANCE} = \left(\int_0^1 \phi(z,\lambda)dz - \pi_0\right)^2 + \int_0^1 [\hat{\pi}_0(z,\lambda) - \phi(z,\lambda)]^2 dz.$$

We estimate π_0 using the smoothing spline method of Storey and Tibshirani ((2003), Remark B).[33] Similarly to CRS, we calculate $\hat{\pi}_0(z_i,\lambda) = \hat{\pi}_{0,b}(\lambda)$ for each grid value $\lambda \in \Lambda = \{0.4, 0.5, \ldots, 0.9\}$, $i = 1, \ldots, m$ and $b = 1, 2, \ldots, n$, the $\hat{\pi}_0(z_i,\lambda)$, and, subsequently, $\int_0^1 \hat{\pi}_0(z,\lambda)dz = \sum_{i=1}^m \hat{\pi}_0(z_i,\lambda)/m$. The unknown $\phi(z,\lambda)$ is estimated by $\hat{\phi}(\lambda,z) = \hat{\pi}_0(z,\Lambda_{\min}) - c_\lambda(1 - \hat{\pi}_0(z,\Lambda_{\min}))$, where c_λ is chosen such that $\int_0^1 \hat{\phi}(\lambda,z)dz = \int_0^1 \hat{\pi}_0(\lambda,z)dz$. We then obtain the optimal $\lambda^* = \arg\min_\lambda \text{MISE}(\lambda)$.

To estimate the joint density function $f(p,z)$, CRS use a local likelihood kernel density estimation (KDE) method with a probit transformation (Geenens (2014)). Specifically, let $\Phi(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^t e^{-x^2/2}dx$ and Φ^{-1} be its inverse. Using $z_i' = \Phi^{-1}(z_i)$ and $p_i' = \Phi^{-1}(p_i)$, we obtain a "pseudo-sample" $\{(p_i',z_i')\}_{i=1}^n$; i.e., we transform the variables (p,z) to (p',z'); we denote by $\tilde{f}(p',z')$ the joint density function of (p',z'), which CRS estimate using the local likelihood KDE method.[34] The bandwidth of the KDE is chosen locally via a k-nearest-neighbor approach using generalized cross-validation; this step can be implemented easily via the freely available R package locfit. The desired density function is then estimated as $\hat{f}(p,z) = \frac{\tilde{f}(p',z')}{\phi(p')\phi(z')}$, where $\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$.

Additionally, $f(p,z)$ may be nonincreasing in p for each fixed z. CRS implement one more step that modifies the $\hat{f}(p,z)$ so that monotonicity is ensured. In our simulations, we use all the aforementioned techniques. In the empirical part, the monotonicity is switched off as this property is unknown in our data. For more details, readers are referred to CRS and their R package `fFDR`, Geenens (2014), and the references therein.

Supplementary material

To view supplementary material for this article, please visit http://doi.org/10.1017/S0022109024000097.

References

Amihud, Y., and R. Goyenko. "Mutual Fund's R^2 as Predictor of Performance." *Review of Financial Studies*, 26 (2013), 667–694.

Andrikogiannopoulou, A., and F. Papakonstantinou. "Estimating mutual fund skill: A new approach." Swiss Finance Institute Research Paper No. 14-42 (2016).

[33] On rare occasions when the sample size m is small, the smoothing spline method does not work adequately. In these cases, we use the bootstrap method of Barras et al. ((2010), Appendix A.1).

[34] This approach is to overcome the erratic behavior at boundary observed when using other popular approaches. Recently, Wen and Wu (2015) propose an alternative method to the issue. We additionally conduct simulations to compare the 2 approaches and see that the local likelihood KDE approach is computationally more efficient and performs better under our framework (FDR control and power gain). For the sake of space, the results are available from the authors.

Andrikogiannopoulou, A., and F. Papakonstantinou. "Reassessing False Discoveries in Mutual Fund Performance: Skill, Luck, or Lack of Power?" *Journal of Finance*, 74 (2019), 2667–2688.

Bajgrowicz, P., and O. Scaillet. "Technical Trading Revisited: False Discoveries, Persistence Tests, and Transaction Costs." *Journal of Financial Economics*, 106 (2012), 473–491.

Barras, L.; P. Gagliardini; and O. Scaillet. "Skill, Scale, and Value Creation in the Mutual Fund Industry." *Journal of Finance*, 77 (2022), 601–638.

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

Barras, L.; O. Scaillet; and R. Wermers. "Reassessing False Discoveries in Mutual Fund Performance: Skill, Luck, or Lack of Power? A Reply." Swiss Finance Institute Research Paper No. 19-61 (2019).

Benjamini, Y., and Y. Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society: Series B (Methodological)*, 57 (1995), 289–300.

Berk, J. B., and J. H. van Binsbergen. "Measuring Skill in the Mutual Fund Industry." *Journal of Financial Economics*, 118 (2015), 1–20.

Bessembinder, H.; M. J. Cooper; and F. Zhang. "Mutual Fund Performance at Long Horizons." *Journal of Financial Economics*, 147 (2023), 132–158.

Carhart, M. M. "On Persistence in Mutual Fund Performance." *Journal of Finance*, 52 (1997), 57–82.

Chen, X.; D. G. Robinson; and J. D. Storey. "The Functional False Discovery Rate with Applications to Genomics." *Biostatistics*, 22 (2021), 68–81.

Chordia, T.; A. Goyal; and A. Saretto. "Anomalies and False Rejections." *Review of Financial Studies*, 33 (2020), 2134–2179.

Clifford, C. P.; J. A. Fulkerson; R. Jame; and B. D. Jordan. "Salience and Mutual Fund Investor Demand for Idiosyncratic Volatility." *Management Science*, 67 (2021), 5234–5254.

Doshi, H.; R. Elkamhi; and M. Simutin. "Managerial Activeness and Mutual Fund Performance." *Review of Asset Pricing Studies*, 5 (2015), 156–184.

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

Geenens, G. "Probit Transformation for Nonparametric Kernel Estimation on the Unit Interval." *Journal of the American Statistical Association*, 109 (2014), 346–358.

Grønborg, N. S.; A. Lunde; A. Timmermann; and R. Wermers. "Picking Funds with Confidence." *Journal of Financial Economics*, 139 (2021), 1–28.

Hansen, P. R. "A Test for Superior Predictive Ability." *Journal of Business & Economic Statistics*, 23 (2005), 365–380.

Harvey, C. R., and Y. Liu. "Decreasing Returns to Scale, Fund Flows, and Performance." Working Paper, Duke University, NBER, and Purdue University (2017).

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

Harvey, C. R., and Y. Liu. "False (and Missed) Discoveries in Financial Economics." *Journal of Finance*, 75 (2020), 2503–2553.

Hsu, P.-H.; Y.-C. Hsu; and C.-M. Kuan. "Testing the Predictive Ability of Technical Analysis Using a New Stepwise Test Without Data Snooping Bias." *Journal of Empirical Finance*, 17 (2010), 471–484.

Hsu, P.-H.; M. P. Taylor; and Z. Wang. "Technical Trading: Is It Still Beating the Foreign Exchange Market?" *Journal of International Economics*, 102 (2016), 188–208.

Ignatiadis, N., and W. Huber. "Covariate Powered Cross-Weighted Multiple Testing." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83 (2021), 720–751.

Jensen, M. C. "The Performance of Mutual Funds in the Period 1945–1964." *Journal of Finance*, 23 (1968), 389–416.

Jones, C. S., and H. Mo. "Out-of-Sample Performance of Mutual Fund Predictors." *Review of Financial Studies*, 34 (2021), 149–193.

Jones, C. S., and J. Shanken. "Mutual Fund Performance with Learning Across Funds." *Journal of Financial Economics*, 78 (2005), 507–552.

Kacperczyk, M.; C. Sialm; and L. Zheng. "Unobserved Actions of Mutual Funds." *Review of Financial Studies*, 21 (2008), 2379–2416.

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

Newey, W. K., and K. D. West. "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica*, 55 (1987), 703–708.

Newton, M. A.; A. Noueiry; D. Sarkar; and P. Ahlquist. "Detecting Differential Gene Expression with a Semiparametric Hierarchical Mixture Method." *Biostatistics*, 5 (2004), 155–176.

Romano, J. P., and M. Wolf. "Stepwise Multiple Testing as Formalized Data Snooping." *Econometrica*, 73 (2005), 1237–1282.

Storey, J. D. "A Direct Approach to False Discovery Rates." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64 (2002), 479–498.

Storey, J. D. "The Positive False Discovery Rate: A Bayesian Interpretation and the q-Value." *Annals of Statistics*, 31 (2003), 2013–2035.

Storey, J. D.; J. M. Akey; and L. Kruglyak. "Multiple Locus Linkage Analysis of Genomewide Expression in Yeast." *PLOS Biology*, 3 (2005), 1380–1390.

Storey, J. D.; J. E. Taylor; and D. Siegmund. "Strong Control, Conservative Point Estimation and Simultaneous Conservative Consistency of False Discovery Rates: A Unified Approach." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 66 (2004), 187–205.

Storey, J. D., and R. Tibshirani. "Statistical Significance for Genomewide Studies." *Proceedings of the National Academy of Sciences*, 100 (2003), 9440–9445.

Sullivan, R.; A. Timmermann; and H. White. "Data-Snooping, Technical Trading Rule Performance, and the Bootstrap." *Journal of Finance*, 54 (1999), 1647–1691.

Sullivan, R.; A. Timmermann; and H. White. "Dangers of Data Mining: The Case of Calendar Effects in Stock Returns." *Journal of Econometrics*, 105 (2001), 249–286.

Tibshirani, R. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)*, 58 (1996), 267–288.

Wen, K., and X. Wu. "An Improved Transformation-Based Kernel Estimator of Densities on the Unit Interval." *Journal of the American Statistical Association*, 110 (2015), 773–783.

Wermers, R. "Mutual Fund Herding and the Impact on Stock Prices." *Journal of Finance*, 54 (1999), 581–622.

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

Zheng, L. "Is Money Smart? A Study of Mutual Fund Investors' Fund Selection Ability." *Journal of Finance*, 54 (1999), 901–933.

Zou, H., and T. Hastie. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67 (2005), 301–320.