Ann Finance
DOI 10.1007/s10436-010-0151-9

RESEARCH ARTICLE

Mutual fund performance: false discoveries, bias, and power

Nik Tuzov · Frederi Viens

Received: 17 July 2009 / Accepted: 17 March 2010

Abstract We analyze the performance of mutual funds from a multiple inference perspective. When the number of funds is large, random fluctuations will cause some funds falsely to appear to outperform the rest. To account for such "false discoveries," a multiple inference approach is necessary. Performance evaluation measures are unlikely to be independent across mutual funds. At the same time, the data are typically not sufficient to estimate the dependence structure of performance measures. In addition, the performance evaluation model can be misspecified. We contribute to the existing literature by applying an empirical Bayes approach that offers a possible way to take these factors into account. We also look into the question of statistical power of the performance evaluation model, which has received little attention in mutual fund studies. We find that the assumption of independence of performance evaluation measures results in significant bias, such as over-estimating the number of outperforming mutual funds. Adjusting for the mutual fund investment objective is helpful, but it still does not result in the discovery of a significant number of successful funds. A detailed analysis reveals a very low power of the study. Even if outperformers are present in the sample, they might not be recognized as such and/or too many years of data might be required to single them out.

Keywords Mutual fund · Performance evaluation · False discovery · Multiple inference · Statistical power

JEL Classification C10 · G10 · G20

N. Tuzov · F. Viens (✉)
Department of Statistics, Purdue University, 150 N. University St.,
W. Lafayette, IN 47907-2067, USA
e-mail: viens@purdue.edu

Published online: 20 April 2010

1 Introduction

The studies of performance of mutual funds go back at least 40 years (Jensen 1968), and this area is still of interest to researchers (Ammann and Verhofen 2009; Cornell et al. 2010). Although a typical mutual fund study has included a large number of funds, the issue of multiple inference (a.k.a. "simultaneous testing," "multiple testing") has received little attention.

Its importance can be illustrated as follows: suppose that we want to evaluate the performance of a large number of fund managers, a certain proportion of whom do not perform well. The performance is measured by a certain test statistic obtained from a performance evaluation model. For instance, such statistic can be a *p*-value obtained under the null hypothesis of "no outperformance". Testing each manager separately at a fixed significance level, one should expect to obtain a certain number of "false discoveries", i.e. the cases where the null hypothesis of "no outperformance" is rejected incorrectly. To distinguish between true and false discoveries, a multiple inference procedure has to be employed.

Multiple inference is straightforward when the test statistics can be assumed independent or "weakly" dependent (Sect. 2.3). This assumption is utilized in Barras et al. (2010) to evaluate the performance of about two thousand US equity mutual funds. Cuthbertson et al. (2008b) apply the same method to perform analysis of UK funds. An almost identical method is used for German data in Otamendi et al. (2008).

However, the independence assumption is unlikely to hold in practice. A typical way to handle this is to propose a parametric or non-parametric model for the dependence structure and incorporate it into the multiple inference procedure. An attempt to consider the dependence across mutual funds via non-parametric approach is made in Kosowski et al. (2006). Unfortunately, neither approach is feasible because the amount of historical mutual fund data is not sufficient to obtain a proper estimate of the dependence structure (Sect. 2.3).

In addition, the performance evaluation model used to obtain the test statistics can be misspecified, which contributes an unknown amount of bias to the inference. To the best of our knowledge, this issue has never been investigated in the mutual fund literature.

In this paper we contribute to the existing literature by using a multiple inference method that seems to offer a viable alternative that does not suffer from the abovementioned shortcomings. Recently, Efron and Tibshirani (2002), Efron (2004a, 2007a,b,c, 2008a,b) developed an empirical Bayes approach that does not rely on the independence of test statistics or the direct estimation of their dependence structure. There is evidence that in some cases, the proposed method can take into account the misspecification of performance evaluation model as well. The approach of Kosowski et al. (2006) is largely different from our method (with the exception that, like them, we use a bootstrap method for the individual estimation of test statistics). It is more appropriate to think of our method as an extension of Barras et al. (2010).

Yet another poorly explored but important question is the statistical power of the performance evaluation model. In a typical mutual fund study, no power diagnostics are provided (Daniel et al. 1997; Carhart 1997; Chen et al. 2000; Kosowski et al. 2006; Barras et al. 2010). Kothari and Warner (2001) try to shed some light on the

issue but their study does not appear exhaustive, especially given that it is not based on the data from the real mutual funds. On the other hand, Efron's method comes with comprehensive and insightful power analysis tools. In addition, it provides a rigorous and efficient way of looking into the performance of subgroups of funds, another issue of practical interest.

We apply the new approach to about two thousand US equity mutual funds observed in 1993–2007. We obtain compelling evidence that assuming independence of test statistics is inappropriate. It introduces a significant bias that results in overestimation of the number of both over- and underperforming funds. Our analysis shows that, although Efron's approach offers higher precision and power, we are still unable to find a significant number of funds that are outperforming after fees and expenses. This result is consistent with Barras et al. (2010), but is different from the findings of Kosowski et al. (2006), who report the existence of a sizable minority of skilled managers.

Finally, the power analysis shows that the study is very underpowered which leaves many outperformers unrecognized as present in the population of all funds. For the subset of funds that are recognized as outperformers, it is hard to separate them from the rest (e.g., for investment purposes). The power is especially low when we try to reduce the history to only the most recent 3–5 years of data.

Section 2 describes the data and proposed approach in detail. Section 3 presents the empirical results for US data. Section 4 concludes.

2 Methodology

2.1 Data

This study is focused on open-end, actively managed US equity mutual funds. The monthly dataset is obtained from CRSP in 03/2008 and it spans 01/1993–06/2007 (14 1/2 years). It is cleared of inappropriate types of funds, such as international, money market, index funds, etc. The minimal total net assets (TNA) in the sample is $5M, and the minimal number of observations per fund is 50. Eventually, net returns for 1911 funds are obtained. Based on the investment objective information, we define three subgroups: 886 Growth (G) funds, 398 Growth & Income (GI) funds, and 627 Aggressive Growth (AG) funds.

The pre-expense returns dataset is obtained from the first dataset by adding the known amount of expense to the net returns. Because of missing expense information, the second dataset includes 1876 funds, with 871 G, 387 GI, and 618 AG funds. The fund monthly return is computed by weighting the return of each fund's shareclass by its monthly TNA. For both datasets, the average number of observations per fund is about 129 (10 3/4 years). Detailed information about the sample construction is available upon request.

2.2 Carhart performance evaluation model

The four-factor (Carhart 1997) performance evaluation model is:

$$\begin{aligned} r_{i,t} &= \alpha_i + b_i \cdot r_{m,t} + s_i \cdot r_{smb,t} + h_i \cdot r_{hml,t} + m_i \cdot r_{mom,t} + \varepsilon_{i,t} \\ t &= 1, \ldots T \\ i &= 1, \ldots m \end{aligned} \tag{2.2.1}$$

where $r_{i,t}$ is the excess return in time period t over the risk-free rate for the fund number i; $r_{m,t}$ is the excess return on the overall equity market portfolio; $r_{smb,t}, r_{hml,t}, r_{mom,t}$ are the returns on so-called factor portfolios for size, book-to-market, and momentum factors (all obtained from CRSP). In the simplest case, the error terms $\varepsilon_t = (\varepsilon_{1,t}, \ldots, \varepsilon_{m,t})$ are assumed i.i.d. multivariate normal $N(0, \Sigma^0)$. All returns are observed and the quantities $\alpha_i, b_i, s_i, h_i, m_i$ are estimated through multiple linear regression, separately for each fund. To obtain more robust estimates in the case when (2.2.1) is misspecified, the regression is estimated via a non-parametric bootstrap procedure similar to that in Kosowski et al. (2006) and Barras et al. (2010).

The parameter α_i is measured in % per month and its value shows by how much the fund outperforms ($\alpha_i > 0$) or underperforms ($\alpha_i < 0$). For each fund i, we compute a single one-sided p-value from the test:

$$H_i^0 : \alpha_i = 0 \text{ vs. } H_i^a : \alpha_i > 0 \tag{2.2.2}$$

The obtained p-values, $\{p_i\}$, $i = 1, m$ are converted into normal z-scores:

$$z_i = \Phi^{-1}(1 - p_i) \tag{2.2.3}$$

where $\Phi^{-1}(.)$ is the inverse normal cdf. For instance, $p_i = 0.025$ corresponds to a fund that is likely to outperform and its $z_i = 1.96$; if, on the other hand, $p_i = 0.975$ (corresponds to a negative α_i) the fund is likely to underperform and its $z_i = -1.96$. From now on, the term "test statistics" will refer to either p-values, or their equivalent z-values.

The composition of our sample (except for the time span) and the performance evaluation model correspond to those in Barras et al. (2010). After 1992, there has to be a significant overlap between their data and our sample.

2.3 False discovery rate and dependent test statistics: a brief review of existing approaches

This section is dedicated to describing the various approaches to working with dependent test statistics that have been previously used in the literature. Suppose we perform the m separate tests, each with a significance level γ. Let Q be the number of rejected true null hypotheses (called "false discoveries") divided by the number of all rejections. Q is a random variable termed False Discovery Proportion, and the expected value of Q is called False Discovery Rate (FDR). The goal of a multiple inference procedure is to force FDR below a pre-specified level q, by choosing an appropriate value of γ.

Denote by P^0 a vector of m_0 p-values that correspond to true null hypotheses. When the components of P^0 are independent and stochastically less or equal to *U(0,1)*, FDR

control can be performed based on a procedure proposed by Benjamini and Hochberg (1995), Benjamini and Yekutieli (2001)).

Further, if we assume that the components of P^0 are i.i.d. $U(0,1)$, the procedure can be empowered by estimating the unknown number of true null hypotheses, m_0 (Benjamini and Hochberg 2000; Benjamini et al. 2006). The idea is to consider the subset of p-values

$$p(\lambda) = \{p_i : p_i > \lambda\}, \quad \lambda \in (0,1) \tag{2.3.1}$$

For λ large enough, $p(\lambda)$ will consist mostly of p-values corresponding to true nulls, i.e. the points in $p(\lambda)$ will approximately have $U(\lambda,\ 1)$ distribution. This is used to estimate λ: e.g., in the histogram of p-values, the plot should "level off" to the right of a certain point on the horizontal axis, that point being $\hat{\lambda}$. Then, the estimate of m_0 is:

$$\hat{m}_0 = \left[\text{number of points in } p(\hat{\lambda})\right] / \left(1 - \hat{\lambda}\right) \tag{2.3.2}$$

This approach is behind the spline estimator proposed in Storey and Tibshirani (2003) and the bootstrap estimator used in (Storey et al. 2004; Storey 2002). The latter approach is used in Barras et al. (2010).

The first practical concern about the method above is that the components of P^0 can be dependent. It is usually assumed that the distribution of P^0 can be adequately approximated by the first two moments. Therefore, we use the terms "dependence structure", "dependence", "correlation structure", "variance-covariance matrix", "joint distribution" interchangeably.

Benjamini and Yekutieli (2001) show that the FDR procedure is still adequate if the test statistics vector has so-called "positive dependency on each one from a subset" structure (PRDS). E.g., suppose that the vector of test statistics is multivariate normal $N(\mu, \Sigma)$. Then, if each null statistic has a non-negative correlation with any other statistic, the joint distribution is PRDS. Verifying the PRDS property is straightforward in some controlled experiments, where the property is implied by the experimental design. However, a mutual fund study is an observational study where, typically, we may not simplify the dependence in this manner. Claiming that each null statistic is non-negatively correlated with the rest is too restrictive to adopt as an assumption.

Another approach is to try to estimate the joint distribution of the test statistics non-parametrically. In Yekutieli and Benjamini (1999) a bootstrap procedure generates m-dimensional samples of p-values under "complete null" setting, i.e. when all m hypotheses are null. A similar resampling scheme is proposed in White (2000), Romano and Wolf (2005), Romano et al. (2007), Romano et al. (2008). Their "StepM procedure" is also akin to the approach developed for biostatistical purposes by van der Laan and Hubbard (2005). These methods assume that there are no constraints on the dependence structure.

A parametric approach can be illustrated as follows. Under the Carhart framework, the dependence structure of the test statistics is defined by Σ^0, the variance-covariance matrix of the vector $\varepsilon_t = (\varepsilon_{1,t}, \ldots, \varepsilon_{m,t})$. One can reduce the number of estimated parameters in Σ^0 by proposing a few "residual factors" that presumably account for all of the cross-sectional (across i) dependence of $\varepsilon\prime_{i,t}s$. The residual factors can be

"qualitative": e.g., one may assume that error terms coming from mutual funds with the same investment objective are correlated with the same correlation coefficient. It is also possible to derive the residual factors from the data using "dimension reduction" techniques, e.g. Principal Component Analysis (PCA). For instance, Jones and Shanken (2005) utilize a combination of "qualitative" and PCA-based residual factors.

Both parametric and non-parametric modeling techniques appear to have a fundamental problem: they only work when the utilized estimate is a "good" estimate of Σ^0, which requires too much historical data. Yekutieli and Benjamini (1999) and White (2000) show that the control of FDR is attained only asymptotically, i.e., for a fixed m and $T \to \infty$. This "size problem" is also investigated by Fan et al. (2008), who demonstrate the inadequacy of a variance-covariance matrix estimator when the data are insufficient. Efron (2007c) refers to the work of van der Laan et al. to emphasize that the corresponding results are applicable only asymptotically and are not to be used unless T is larger than m. In the context of mutual fund studies, we have m about 2000 and T between 50 and 300, which amounts to a severe "size problem". The various dimension reduction techniques allow us to "reduce the dimension" of the available data (i.e., use the available data efficiently), but they provide no solution for the size problem. In particular, PCA takes as input the estimated variance-covariance matrix of residuals from (2.2.1), and simplifies it by extracting the main principal components. While this extraction can keep most of the information, the problem is that the very input of PCA is a poor estimate of Σ^0 due to T being far smaller than m.

The "size problem" is often ignored in applications. Yekutieli and Benjamini (1999) give a weather analysis example where $m = 1977$ and $T = 39$, while using another, simulated dataset to show that FDR is controlled in which $m = 40$ and T in between 200 and 1,000. In the mutual fund performance area, the "cross-sectional bootstrap" procedure of Kosowski et al. (2006) performs a non-parametric estimate of the dependence structure with m over 2,000 and T about 300. Cuthbertson et al. (2008a) borrow this approach and apply it to UK data with $m = 900$ and T about 340. Barras et al. (2010) conduct a few Monte-Carlo experiments to see how their approach works when certain forms of dependence are introduced. However, to specify the data generating process, one has to be able to estimate the dependence adequately, which is impossible because of "size problem". For instance, one of their simulations is based on the residual correlation matrix for 898 mutual funds. Apparently, an adequate estimate would require at least 898 monthly observations, or 75 years, of data. Instead, only 60 monthly values are used and there is no way to say whether the true dependence structure is close to what is assumed in the experiment.

Another complication is that the equity market data show that Σ^0 is very time-dependent. The correlation between equity funds can go up and down depending on the state of economy. Avellaneda and Lee (2008) illustrate this effect by considering a large number of US stocks observed daily between 2002 and 2008. They show that during the "good times" of 2004–2006, the equity returns are significantly less correlated than during "bad times" (2002, 2007–2008). Therefore, in a multifactor model with a fixed number of factors, the cross-sectional dependence structure of the error terms can change drastically over time, which aggravates the already serious "size problem".

One more way to handle the dependence is the assumption of "weak dependence" (Storey et al. 2004; Storey and Tibshirani 2003; Storey 2003). When the assumption is satisfied, the p-values are treated as if independent and the (asymptotic) FDR control still takes place. There is no statistical procedure to test for weak dependence, even though one could make a qualitative argument that it holds for particular datasets: e.g., it is likely to hold when the test statistics are only dependent (if at all) within small groups with the groups being independent of each other.

Storey et al. (2004) also show that under weak dependence FDR can be controlled for any fixed value of $\hat{\lambda}$ in (2.3.2). The choice of optimal $\hat{\lambda}$ is a bias-variance tradeoff problem which they solve via bootstrapping from the m p-values. Resampling from a set of (weakly) dependent p-values is a questionable technique for which no analytical results are available. Still, some numerical examples show that the bootstrap estimation of $\hat{\lambda}$ is robust under "small group" type of weak dependence (Storey and Tibshirani 2001). The application of FDR in Barras et al. (2010) rests on the assumption of weak dependence for the purpose of both FDR control and the estimation of the optimal $\hat{\lambda}$ via bootstrap.

This assumption may not be justifiable. As stated in Barras et al. (2010) itself, mutual funds may exhibit correlated trading behaviors in large groups that can be caused, for instance, by being exposed to the same industrial sector or "herding" into particular stock(s). To address that, Barras et al. (2010) argue that the funds' test statistics are not very dependent because 15% of the fund histories in their sample do not overlap in time, and on average only 55% of return observations overlap. How much independence does the "lack of overlap" introduce? Compare this to an example of a weakly dependent structure with $m = 3,000$ and the group size of 10 in Storey et al. (2004). For a mutual fund study with $m = 2,000$, where the degree of independence is associated with the absence of overlap, we obtain the following: the entire time period should be divided into subintervals with under 10 funds observed on each subinterval. Hence, it requires at least 200 subintervals. An average fund being observed for over 10 years, it implies the study's time span has to be over 2000 years. In reality, the dataset in Barras et al. (2010) spans only 32 years, which makes the "lack of overlap" argument doubtful.

Therefore, the weak dependence property inevitably implies some rather questionable and/or hard-to-check assumptions about the data. Explicit modeling of the high-dimensional correlation structure is not feasible either. Even very restrictive assumptions may not reduce the number of model parameters to the point where the amount of data is sufficient for estimation. The next section introduces a novel approach to large-scale simultaneous inference that can offer a viable alternative.

2.4 Alternative approach: structural model and empirical null hypothesis

2.4.1 Structural model

Following Efron (2004a, 2007a,b,c, 2008a,b), we propose a model for the density of z-values that are obtained from (2.2.3):

$$\begin{aligned} &\alpha \sim g(\alpha) \\ &z|\alpha \sim N(\alpha, \sigma_0^2) \\ &f(z) = g(\alpha) * \varphi(z|0, \sigma_0^2) \end{aligned} \tag{2.4.1}$$

where $\varphi(z|\mu, \sigma^2)$ denotes the density of a normal with mean μ and variance σ^2. The α value itself is a random variable with an arbitrary (not necessarily continuous) distribution. Without loss of generality, denote by $g(\alpha)$ the density of α, and by P_g the corresponding probability measure. Each observed z-value is normal with mean α and variance σ_0^2. As a result, the density $f(z)$ is a mixture of normals with random means, which can also be expressed as a convolution (denoted by "*") of $g(\alpha)$ and $\varphi(z|0, \sigma_0^2)$.

Although the αvalues in (2.4.1) are not the same as α_i in (2.2.2), they are denoted by the same symbol because their signs coincide: a positive (zero, negative) sign for α_i in (2.2.2) corresponds to a positive (zero, negative) sign for α in (2.4.1). The observed test statistic z_i can be seen as a noisy signal from which one has to "back out" the sign of α_i.

Our interest is in testing some hypothesis about α, and the support of $g(\alpha)$ can be arbitrarily split into two disjoint "null" and "alternative" sets, depending on which case αfalls under:

$$g(\alpha) = p_0 g_0(\alpha) + p_1 g_1(\alpha) \tag{2.4.2}$$

where

$g_0(\alpha)$ is the "null" component (i.e. it is the density of the values of α given that they fall under the null hypothesis)

$g_1(\alpha)$ is the "alternative" component (i.e. it is the density of the values of α given that they fall under the alternative hypothesis)

$p_0 = P_g\ \{\alpha \text{ is in null set}\}$, the probability that α is null

$p_1 = P_g\ \{\alpha \text{ is in alternative set}\}$, the probability that α is alternative

$p_0 + p_1 = 1$

In terms of corresponding z-values this will result in a splitting of the density f of observed z values:

$$\begin{aligned} f_0(z) &= g_0 * \varphi\left(z|0, \sigma_0^2\right) && \text{is the null density of } z\text{'s} \\ f_1(z) &= g_1 * \varphi\left(z|0, \sigma_0^2\right) && \text{is the alternative density of } z\text{'s} \\ f(z) &= p_0 f_0(z) + p_1 f_1(z) && \text{is therefore a mixture resulting in the density of } z\text{'s} \end{aligned} \tag{2.4.3}$$

For instance, if we decided to test $H_0 : \alpha_i = 0$ vs. $H_a \alpha_i \neq 0$, the "null" set would consist of one point $\{\alpha = 0\}$, the "alternative" set would be $\{\alpha \neq 0\}$. If all null

p-values are i.i.d. $U(0,1)$, the corresponding density of null z-values would then be $f_0(z) = \varphi(z|0, 1)$.

We can also consider a non-atomic null set, e.g. $\{\alpha \leq 0\}$, and set the following notations:

$$g(\alpha) = p_0 g_0(\alpha) + p_1^+ g_1^+(\alpha) \tag{2.4.4}$$

where

$p_0 = P_g\{\alpha \leq 0\}$ is the probability that α is null,
$p_1^+ = P_g\{\alpha > 0\}$ is the probability that α is alternative
$g_0(\alpha)$ is the density of the composite null, with support on $\{\alpha \leq 0\}$
$g_1^+(\alpha)$ is the density of the alternative, with support on $\{\alpha > 0\}$.

Then the mixture density

$$f(z) = p_0 f_0(z) + p_1^+ f_1(z) \quad \text{is the density of } z\text{s}$$

where we defined

$$\begin{aligned} f_0(z) &= g_0 * \phi\left(z|0, \sigma_0^2\right) && \text{for the conditional density of null } z\text{'s} \\ f_1^+(z) &= g_1^+ * \phi\left(z|0, \sigma_0^2\right) && \text{for the conditional density of alternative } z\text{'s} \end{aligned}$$

and it must hold that $p_0 + p_1^+ = 1$

Since in this case the distribution of null z-values, $f_0(z)$, is a result of convolution over a non-atomic set of α values, we are going to call such $f_0(z)$ a "composite null distribution".

A slight extension of (2.4.4) would be a three-component model, which would correspond to separating the funds into three groups: underperformers, zero-alpha funds, and outperformers. This leads one to split the support of $g(\alpha)$ into three subsets, where the two alternatives are separated. Similarly to (2.4.4), we therefore set the following notation:

$$g(\alpha) = p_0 g_0(0) + p_1^+ g_1^+(\alpha) + p_1^- g_1^-(\alpha) \tag{2.4.5}$$

where

$p_0 = P_g\{\alpha = 0\}, \quad p_1^+ = P_g\{\alpha > 0\}, \quad p_1^- = P_g\{\alpha < 0\}$
$g_0(\alpha)$—"zero" density equal to delta function
$g_1^+(\alpha)$—"positive" density with support on $\{\alpha > 0\}$
$g_1^-(\alpha)$—"negative" density with support on $\{\alpha < 0\}$

$$\begin{aligned} f(z) &= p_0 f_0(z) + p_1 f_1(z) \quad \text{mixture density of } z\text{s} \\ p_1 f_1(z) &= p_1^+ f_1^+(z) + p_1^- f_1^-(z) \end{aligned}$$

where

$f_0(z) = \phi\left(z|0, \sigma_0^2\right)$ "zero" density of zs

$f_1^+(z) = g_1^+ * \phi(z|0, \sigma_0^2)$ "positive" density of z's
$f_1^-(z) = g_1^- * \phi\left(z|0, \sigma_0^2\right)$ "negative" density of z's
$p_1^+ + p_1^- = p_1, p_0 + p_1 = 1$

As mentioned in the introduction, in order to analyze the Bayesian concept of false discovery, we adopt the notion of "local false discovery rate" (fdr), which can be interpreted as a "local" version of Benjamini and Hochberg's FDR. For the two-component model, (2.4.3), it is defined as follows:

$$fdr(z) = P\{\text{case } i \text{ is null } |z_i = z\} = \frac{p_0 f_0(z)}{f(z)} \tag{2.4.6}$$

This local fdr, $fdr(z)$, is the posterior probability that the test with corresponding z-score came from the null distribution. Likewise, for the three-component model, we may define fdr+ as a posterior probability that the corresponding z-score came from the fund that is not outperforming.

$$fdr^+(z) = \left[p_0 f_0(z) + p_1^- f_1^-(z)\right]/f(z) \tag{2.4.7}$$

For identification of underperformance, $fdr^-(z)$is defined in a similar manner. When talking about outperforming (underperforming) mutual funds, the term "false discoveries" will refer to the funds that are not true outperformers (true underperformers). For the sake of simplicity, the superscripts will be omitted. The Appendix provides more details on model identification and estimation.

One may also consider rates for tails:

$$\begin{aligned} Fdr(z) &= FdrLeft(z) = P\{\text{case } i \text{ is null} \mid z_i \le z\} = \frac{p_0 F_0(z)}{F(z)} = E_f[fdr(t)|t \le z] \\ FdrRight(z) &= P\{\text{case } i \text{ is null} \mid z_i \ge z\} = E_f[fdr(t)|t \ge z] \end{aligned} \tag{2.4.8}$$

where F_0 and F are cdf's corresponding to f_0 and f. FDR and Fdr (also denoted FdrLeft) are closely related measures that reflect the average false discovery rate within a tail region. On the other hand, fdr has a local nature and provides more precision in interpreting z's on an individual basis. Another advantage of this approach is that neither (2.4.6) nor (2.4.8) assume any particular dependence structure of z's such as PRDS or the weak dependence.

2.4.2 Empirical null hypothesis

The local fdr approach is of the "empirical Bayes" kind: in all the models presented in the last section, we do not pre-specify the mixture density $f(z)$ and p_0 because, unlike in the "classical Bayes" setting, $f(z)$ and p_0 are estimated from the data. The null density $f_0(z)$ is usually pre-specified, e.g. as $\varphi(z|0, 1)$ (called "theoretical null"). In certain cases it makes sense to estimate $f_0(z)$ from the data also. (Efron 2001, 2004a, 2007c,b) introduced the concept of "empirical null" where $f_0(z)$ is approximated by $\varphi\left(z|\delta_0, \sigma_0^2\right)$ and the parameters δ_0 and σ_0^2 are estimated.

To understand why this may be necessary, consider the following hierarchical model. It is used to see how the presence of correlation affects the FDR control if we treat z's as if independent (Efron 2007c). Suppose, each pair (z_i, z_j) is bivariate normal with a distinct correlation coefficient ρ_{ij} drawn randomly from a certain normal distribution $N(0, \tau^2)$. In marginal terms, 95% of z-values come from $N(0, 1)$ and 5% come from $N(2.5, 1.25)$. For each of 1,000 generated datasets, the Benjamini-Hochberg procedure is applied with FDR control level of 0.1 to produce a certain cutoff γ (Sect. 2.3). The realization of False Discovery Proportion, Q, is computed for each dataset and the average of 1,000 values is compared to the target of 0.1.

It turns out that the average of Q is under control despite the dependence. This is intuitively clear because, since ρ_{ij} are zero on average, their impact should disappear after averaging across all datasets. However, the variance of False Discovery Proportion is very large, i.e., there are many datasets where the true proportion is far above or below its target value of 0.1. Efron shows that it is possible to reduce the variance by conditioning of the realization of a certain "dispersion variate", A, which can be estimated from the central portion of the histogram of z-values.

It can also be shown that the ensemble of null z-values will behave closely to an ensemble of i.i.d. $N(0, \sigma_0^2)$ where $\sigma_0^2 = 1 + \sqrt{2}A$. The positive realizations of A produce $\sigma_0^2 > 1$ ("overdispersion") and, as a result, too many null cases will be declared significant ("over-rejection"). A negative value of A results in $\sigma_0^2 < 1$ ("underdispersion"), and too few alternative cases are detected. To adjust for A, we have to use the empirical null $f_0(z) = \varphi(z|\delta_0, \sigma_0^2)$. To see how this can affect the inference, we introduce another version of (2.4.8):

$$\tilde{F}dr(x|A) = P\{z_i \text{ null} \mid z_i \geq x, A\} \tag{2.4.9}$$

Suppose we are interested in detecting the outperforming funds, so set $x = 2.5$. The case $A = 0$ corresponds to the theoretical null. Our calculations show that, if A takes on the value of, say, 0.16, the proportion of null z's in the tail region $\{z > 2.5\}$ is about 1.8 times as great as it is under $A = 0$. Suppose $\tilde{F}dr(2.5|0)$ is 0.2, then $\tilde{F}dr(2.5|0.16)$ is 0.36. If 100 of z's fall above 2.5, 80 of them are "true discoveries" under the theoretical null, but under $A = 0.16$ the number of true discoveries is only 64. In Sect. 3 we are going to estimate σ_0^2 and report the corresponding value of A based on the relation $A = (\sigma_0^2 - 1)/\sqrt{2}$ in order to show whether overdispersion is practically significant.

If we slightly extend the experiment above to estimate the proportion of null cases, the estimate will be biased downward for $A > 0$ and upwards for $A < 0$, but on average, across all datasets and realizations of A, it will be close to the true value of 95%. For a particular dataset, however, the results can be biased because the corresponding value of A can be sizably away from zero, and adjustment for A (i.e., switching to empirical null) may be necessary.

Barras et al. (2010, Appendix B.2, Tables AI, AII) perform a Monte-Carlo experiment that is similar to the one above: the cross-sectional correlations of error terms in (2.2.1) are distributed over a narrow interval centered approximately at zero, which means the correlations of z-values are also centered at about zero. For each of 1,000 simulated datasets, the point estimates of p_0, p_1^+, p_1^- are obtained. The confidence intervals for p_0, p_1^+, p_1^- are obtained by taking the corresponding quantiles of 1,000

values. As expected, the confidence intervals are centered at true values regardless of whether or not dependence is introduced. However, as a result of dependence, the width of confidence intervals goes up to the point where it becomes of practical significance. Consider p_1^+, whose true value is 2%. Under independence, the right boundary of 90% confidence interval for p_1^+ is 3.8%, but dependence inflates it to 6.5%. That is, due to under- or overdispersion, the point estimate of p_1^+ can be possibly off by an extra $6.5 - 3.8 = 2.7\%$. For the real dataset, Barras et al. (2010) report the point estimate of p_1^+ as 0.6% with a standard error of 0.8%. If underdispersion is present, the point estimate might be revised to $2.7 + 0.6 = 3.3\%$, which will become both practically and statistically significant. Therefore, these simulation results indicate that dependence adjustment can be necessary even when the true data-generating process is identical to that in the Monte-Carlo experiment. Kosowski et al. (2006, Appendix C), performed a similar experiment where the dependence structure was simulated non-parametrically. They do not report significant differences from the independent case. Like in the experiment above, we interpret this as the absence of bias when averaging across a large number of hypothetical datasets.

Another complication is that the marginal distribution of the null z-values can, indeed, be closer to $N(0, \sigma_0^2 \neq 1)$ than to $N(0, 1)$. According to Efron (2007b), it can happen when the model used to obtain the individual test statistics (in our case, it is (2.2.1)) is misspecified. Possible sources of misspecification are: unconsidered serial correlation or heteroskedasticity of error terms, application of asymptotically valid estimation when T is not large enough, and so on. In that case, the above mentioned example (Efron 2007c) shows that even independent test statistics can behave as if dependent. Therefore, it is hard to say whether the observed deviation from the theoretical null is caused by dependence or model misspecification.

In practice, both cross-sectional dependence and the misspecification of marginal distribution can be present. While one can try to ignore the former via justifying the independence/weak dependence assumption, the contribution of the latter is impossible to assess a priori: if one had known how the model was misspecified, one would have corrected the misspecification in the first place. Monte-Carlo simulations would not expose under- or overdispersion caused by model misspecification.

Efron (2007b) shows that, in the above example, not only the point estimate of $fdr(z)$ but also its estimated standard error, $s.e.(\hat{fdr}(z))$, are conditioned on the ancillary statistic A, and, in that sense, are conditioned on the dependence structure of z's. Likewise, the standard errors of $\hat{p}_0, \hat{\delta}_0, \hat{\sigma}_0$ are also conditioned on the dependence structure. In this example, using the empirical null is essentially a way to adjust the inference for the dependence structure of $z's$ without having to model it explicitly. In addition, this takes into account the possible misspecification of the marginal distribution of null test statistics.

The advantage of Efron's approach can be summarized as follows: to perform multiple inference, we need not the dependence structure per se, but the estimates of $f(z)$ and $p_0 f_0(z)$. When the "size problem" is present, we know very little about the true dependence structure. Also, it can be hard to verify the weak dependence/independence assumption for test statistics. Therefore, estimating the structural model directly from the observed z-scores is a viable shortcut one may choose when there is enough data. Importantly, for Efron's model, "enough data" means that m (as opposed

to T in (2.2.1)) has to be large, which implies that the "size problem" turns to our advantage.

Using the theoretical null is equivalent to assuming that the null p-values are i.i.d. $U(0,1)$ and the null z-scores are i.i.d. $N(0,1)$, which corresponds to assumptions in Barras et al. (2010). Given that FDR is very closely related to Fdr (Efron and Tibshirani 2002), we can treat our theoretical null-based procedure as a close match to the approach of Barras et al. (2010), with directly comparable results. While the theoretical null is always the first option to try, the findings of Efron suggest that it is also worth checking whether there is strong evidence against the theoretical null. If that is the case, switching to the empirical null can be a justifiable option.

2.4.3 Parameter estimation

The numerical results in this study are obtained based on the *R* package *locfdr*, which includes both theoretical and empirical null options. Regardless of whether the empirical or theoretical null is used, the estimation of the parameters of null component, $p_0 f_0(z)$, is based on the "zero assumption": it is assumed that only the null component is supported on a certain "zero interval" $(z_-; z_+)$. For the theoretical null and a fixed zero interval, the point estimate of p_0 is the same in Barras et al. (2010) (formula (2.3.2)) and Efron's approach. A small technical difference is that Barras et al. (2010) work with two-sided p-values obtained from the test $H_i^0 : \alpha_i = 0$ vs. $H_i^a : \alpha_i \neq 0$ while in this study we utilize one-sided z-values (2.2.3). In particular, the interval $U(\lambda,\ 1)$ from Sect. 2.3 corresponds to a symmetrical zero interval on z-axis: e.g., $U(0.05; 1)$ is identical to the zero interval $(-1.96; 1.96)$. The following formula shows the relation between λ, z_- and z_+:

$$\lambda = \Phi(z_-) + (1 - \Phi(z_+)) \tag{2.4.10}$$

The choice of the zero interval itself is a bias-variance tradeoff problem where the value of λ or, equivalently, the boundaries of $(z_-; z_+)$ are the smoothing parameters. Barras et al. (2010) minimize $MSE(\hat{p}_0)$ using λ as a smoothing parameter. For a fixed λ, $MSE(\hat{p}_0)$ is calculated based on bootstrap technique (Sect. 2.3) which we prefer to avoid. Instead, we use a bias-variance tradeoff estimation method similar to that in Turnbull (2007). Therefore, our method can be seen as an extension of that in Barras et al. (2010) in how we choose the zero interval, $(z_-; z_+)$, and model the null distribution, $f_0(z)$. Appendix provides more details on model estimation.

2.4.4 Power statistics

A high power means that $fdr(z)$ is small on the support of $f_1(z)$, which can be described by an overall (post hoc) power measure:

$$Efdr = \int fdr(z) f_1(z) dz / \int f_1(z) dz = E_{f_1}[fdr(z)] \tag{2.4.11}$$

It can be adapted to measure the power in the right tail:

$$EfdrRight = \int_0^{+\infty} fdr(z) f_1(z) dz / \int_0^{+\infty} f_1(z) dz = E_{f_1}[fdr(z)|z > 0] \quad (2.4.12)$$

EfdrLeft is defined similarly for the left tail ($z < 0$).

To put a cap on the proportion of false discoveries, Efron (2007b) recommends picking z-values with $fdr(z) \leq 0.2$. We also adapt (a more liberal) rule: declare the fund under- (outperforming) as long as its FdrLeft (FdrRight) are under 0.2. It means that we shall tolerate up to 20% of false discoveries when we wish to construct an outperforming fund portfolio. Similarly, we say that the study has decent power when Efdr is under 0.2.

An interesting question is whether one could improve the study by increasing the number of observations per fund, T. Assuming that the standard error of $\hat{\alpha}_i$ in (2.2.1) is proportional to $1/\sqrt{T}$, the package *locfdr* allows us to gauge how much power would be gained by increasing the value of T. This is done under the assumption that all parameters in the model, except T, are fixed at the original estimated values. Taking one step back, one may ask whether those estimates, such as $\widehat{p_1^+}$, are adequate to begin with. This issue is investigated in Sect. 3.2.

2.4.5 Performance versus investment objective

It would be interesting to look into the net performance versus investment objective. The findings of Barras et al. (2010) suggest that one may be able to increase the power by using investment objective as a control factor.

Barras et al. (2010) compare the fund categories by running their bootstrap-based procedure for each category separately. We can perform an fdr-based analysis which will not suffer from the possible misspecifications described in Sect. 2.4.2, insofar as we have the option of using the empirical null.

Efron (2008b) proposes the following method. Suppose that all z-values are divided into two classes, A and B. In mutual fund context, class A corresponds to the investment category of interest (e.g. Aggressive Growth), and class B corresponds to the rest of funds. Then the mixture density and fdr can be decomposed as follows:

$$f(z) = \pi_A \cdot f_A(z) + \pi_B \cdot f_B(z) \quad (2.4.13)$$

π_A, π_B—a priori probabilities of class A and B

$f_A(z) = p_{A0} f_{A0}(z) + p_{A1} f_{A1}(z)$—class A mixture density
$f_{A0}(z)$, $f_{A1}(z)$—null and alternative densities for class A
$fdr_A(z) = p_{A0} f_{A0}(z)/f_A(z)$—class A fdr
$f_B(z) = p_{B0} f_{B0}(z) + p_{B1} f_{B1}(z)$—class B mixture density
$f_{B0}(z)$, $f_{B1}(z)$—null and alternative densities for class B
$fdr_B(z) = p_{B0} f_{B0}(z)/f_B(z)$—class B fdr

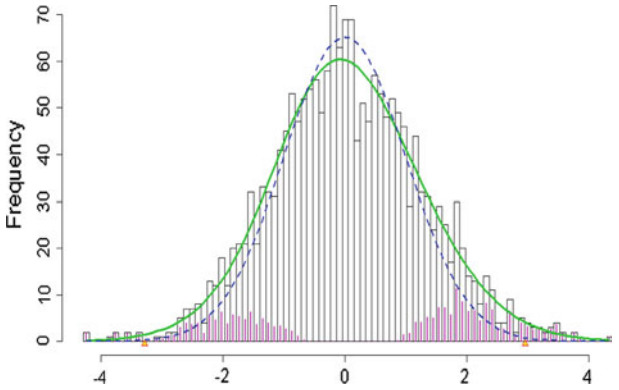

Fig. 1 Estimated densities for pre-expense returns and theoretical null. The model (2.2.1) is applied to pre-expense returns of 1876 funds. The histogram represents z-values obtained according to (2.2.2) and (2.2.3). The *solid* (*green*) *curve* is $\hat{f}(z)$, an estimate of mixture density in the structural model (2.4.5). The *dashed* (*blue*) *curve* is $\hat{p}_0 \cdot \varphi(z|0, 1)$, an estimate of the null component of mixture density. The *thin* (*pink*) *dashes* in non-central bins are "thinned counts". Each thinned count is equal to the full count in the corresponding bin times $\hat{p}_1 \hat{f}_1(z)$, an estimate of alternative component in the structural model. Thinned counts reflect the proportion of alternative cases in each bin

The main hypothesis of interest is:

$$H_0 : fdr_A(z) = fdr(z) \tag{2.4.14}$$

We do not have to run a separate fdr analysis for each group as long as the assumption

$$f_{A0}(z) = f_{B0}(z) \tag{2.4.15}$$

holds. This is another advantage of Efron's approach because it allows us to avoid redundant parameters. A certain logistic regression procedure is utilized to test the assumption (2.4.15), test the main hypothesis (2.4.14), and obtain an estimate of $fdr_A(z)$ (details are available upon request from the authors).

3 Empirical results

3.1 Pre-expense returns, theoretical null

We estimate the structural model (2.4.5) and obtain the following results. Figure 1 shows the histogram of z-scores (y axis indicates the counts of z-scores in each of 90 bins), the estimate of mixture density, $\hat{f}(z)$, showed by solid (green) curve, and the estimated null component, $\hat{p}_0 \cdot \varphi(z|0, 1)$, showed by dashed (blue) curve.

The thin (pink) dashes in non-central bins are so-called "thinned counts" that are equal to full z counts times the estimated alternative component, $\hat{p}_1 \hat{f}_1(z)$. Thinned counts reflect the proportion of alternative cases in each bin.

In Barras et al. (2010), the corresponding estimates of p_0, p_1^+, p_1^- are 85.9 (2.7), 9.6 (1.5), 4.5 (1.0). For our sample, the confidence intervals for p_0, p_1^+, p_1^- in Table 1

Table 1 Estimation results for pre-expense returns and theoretical null

	$\hat{p}_0$, %	$\hat{p}_1^+$, %	$\hat{p}_1^-$, %
	89.42 (0.75)	6.30 (0.53)	4.28 (0.53)
95% CI	(87.95; 90.89)	(5.26; 7.33)	(3.24; 5.31)
Number of funds	1678	118	80
	Zero interval	$\hat{\lambda}$	
	$(-1.5; 1.5)$	0.1336	

The table shows the estimated proportions of zero-alpha, outperforming, and underperforming mutual funds ($\hat{p}_0$, $\hat{p}_1^+$, $\hat{p}_1^-$), based on pre-expense fund returns and theoretical null distribution. The corresponding structural model is (2.4.5). Standard errors are provided in parentheses, along with 95% confidence intervals in the 3rd line. The 4th line shows the estimated number of zero-alpha, outperforming, and underperforming funds in the population of 1,876 funds. To provide more details about the estimation procedure, we report the boundaries of zero interval and the value of $\hat{\lambda}$ (see Sect. 2.4.3) in the bottom line

have a lot of intersection with the corresponding intervals in Barras et al. (2010). The estimate of positive proportion drops from 9.6 to 6.3%, which is consistent with post-1992 deterioration of mutual fund performance discovered in Barras et al. (2010). Still, the proportion of positive performers is both practically and statistically significant.

The results of Table 1 suggest that some 118 funds out of 1,876 are outperforming on pre-expense basis. Knowing that some 118 funds are worth looking into is not the same as knowing those 118 skilled funds by name. The small triangles under the horizontal axis in Fig. 1 mark the cutoffs where $fdr(z) = 0.2$. The funds to the right (left) of the right (left) cutoff can be identified as outperforming (underperforming). The majority of thinned counts fall in between the cutoffs, so the study appears underpowered. The power statistics confirm the suspicion: Efdr = 0.56, EfdrRight = 0.5, and EfdrLeft = 0.64.

It practice it means that, under FdrRight = 0.2, we can identify only 29% (34 out of 118) of outperformers. The only way to increase the proportion of identifiable under/outperformers for this sample is to tolerate a higher percentage of false discoveries, i.e. to move the left and right cutoffs closer to zero (Table 2). To select 47% (55 funds) out of total 118 outperformers, one has to tolerate FdrRight of about 0.3 meaning that 24 false discoveries have to be selected also: 24/(24 + 55) = 0.3. To select 95% (112 funds) of outperformers, one has to include about 168 false discoveries. Because of low power and small proportion of outperformers, the quality of "top lists" of fund managers is not good: e.g., "Top 79 performers" (79 = 55 + 24) will have 24 indistinguishable false discoveries, and the list of "Top 43 performers" will have some 9 false entries in it.

How would the result change if we had more years of data? In the original sample, we have on average 10 3/4 years of observations per fund; we can loosely think of this as having 10 3/4 years of data for each fund in the sample.

For instance, having 32 years of observations for each fund could help identify 90% (106 out of 118) of outperformers with FdrRight = 0.2 (Table 3). This corresponds to EfdrRight = 0.2, confirming that using 0.2 as a rule of thumb for good power is reasonable.

Table 2 Identified outperformers and false discoveries versus FdrRight for pre-expense returns and theoretical null

FdrRight	Proportion of identified outperformers, %	Number of identified outperformers (rounded)	Number of false discoveries (rounded)
0.1185	15	18 out of 118	2
0.2	29	34 out of 118	9
0.3	47	55 out of 118	24
0.4	65	78 out of 118	51
0.5	83	98 out of 118	98
0.6	95	112 out of 118	168
0.7	100	118 out of 118	275

This table describes one's ability to detect outperformers. For each level of FdrRight (defined by 2.4.8) in the first column, the second and third columns specify the proportion and number of outperformers (out of total 118) whose z-values fall above the right-hand-side cutoff determined by FdrRight. The last column shows how many useless funds (false discoveries) have to be tolerated because they also fall above the cutoff

Table 3 Increase in power versus years of available data for pre-expense returns and theoretical null

Sample size, years	Efdr	EfdrRight	Outperformers identifiable with FdrRight = 0.2
$10^{3/4}$	0.56	0.50	34 out of 118
15	0.44	0.38	70 out of 118
20	0.36	0.30	90 out of 118
25	0.30	0.25	98 out of 118
32	0.25	0.20	106 out of 118

The table reflects one's increased ability to identify outperformers as a result of hypothetical increase in the amount of historical data per fund. The first column indicates the amount of historical data, in years per fund. Second and third column report the power statistics Efdr and EfdrRight, defined in Sect. 2.4.4. The last column shows how many out of total 118 outperformers are captured given that the proportion of false discoveries is capped at 20%

In accordance to EfdrLeft = 0.64, the tables similar to Tables 2 and 3 (not shown) indicate that power in the left tail is much worse: e.g., even with 40 years per each fund only about 81% (65 out of 80) of underperformers are identified with FdrLeft = 0.2.

3.2 Pre-expense performance, empirical null

Given that the 95% confidence interval for p_0 in Table 1 is (87.95; 90.89), it is possible to assume that $p_0 \geq 0.9$. This assumption is necessary if we want to check whether the theoretical null is adequate for the data (see Appendix). We are going to add two more free parameters, i.e. assume that $f_0(z) = \varphi(z|\delta_0, \sigma_0^2)$. If the theoretical null is appropriate, the estimated empirical parameters δ_0 and σ_0 should not be significantly different from 0 and 1, respectively.

Table 4 Estimation results for pre-expense returns and empirical null

Zero interval	$\hat{\lambda}$	$\hat{p}_0$, %	EfdrRight
(−1.7; 1.7)	0.0891	98.11 (0.99) (96.17; 100.05)	0.712
$\hat{\delta}_0$	$\hat{\sigma}_0$	t-value for $H_0 : \sigma_0 = 1$	A
0.0039 (0.0353)	1.179 (0.034)	5.29	0.276

The table shows the estimated proportion of zero-alpha funds, $\hat{p}_0$, along with its 95% confidence interval. The corresponding structural model is (2.4.5). The boundaries of the zero interval, the value of $\hat{\lambda}$ (Sect. 2.4.3), and the power statistic EfdrRight (Sect. 2.4.4) are also reported. For the empirical null, $f_0(z) = \varphi(z|\delta_0, \sigma_0^2)$, we report the estimates of δ_0, σ_0, and the value of dispersion variate A that corresponds to $\hat{\sigma}_0$ (Sect. 2.4.2). The reported t-value for $H_0 : \sigma_0 = 1$ measures the statistical significance of overdispersion while the value of A gauges its practical significance. Figures in parentheses denote the standard errors of the different estimators

As we see from Table 4, $\hat{\sigma}_0$ is significantly greater than 1 with the corresponding t-value of 5.29. In other words, the z-values exhibit statistically significant overdispersion.

Comparing Figs. 1 and 2, we see that the empirical null has a much better fit to $\hat{f}(z)$ in the central part of the histogram, i.e., the bias of the null distribution is reduced. In theory, the dashed (blue) curve, $\hat{p}_0 \hat{f}_0(z)$, must always be under the solid (green) curve, $\hat{f}(z)$. This is clearly violated on Fig. 1, indicating high bias. With more free parameters, the empirical null has lower bias and higher variance. If we compare the measures of variance and bias (see Appendix) on the same zero interval (−1.7; 1.7), it turns out that the empirical null produces a variance that is 2.2 times as large and a bias that is 34.5 times as small, an obviously more favorable bias-variance tradeoff for the empirical null.

In terms of practical significance, one may think of such z-values as being marginally $N(0,1)$ and pairwise correlated with the correlation density $\rho \sim N(0, \tau^2)$, (see Sect. 2.4.2). The estimate of the dispersion variate A in Table 4 is 0.276, which is far greater than the example of $A = 0.16$ discussed in Sect. 2.4.2. Using $\tilde{F}dr(x|A)$ from (2.4.9), if we assume that $\tilde{F}dr(2.5|0) = 0.2$, then $\tilde{F}dr(2.5|0.276) = 0.47$. It means that if 100 z's fall above 2.5, 80 of them are true discoveries if the theoretical null is used, but with the empirical null that number drops to 53.

Therefore, we have both statistically and practically significant evidence against the theoretical null. The theoretical null-based inference overestimates the number of both skilled and unskilled funds in the population. The 95% confidence interval for p_0 changes from (87.95; 90.89) under the theoretical null to (96.17; 100.05) under the empirical null. The latter means that it *is possible that both underperformers and outperformers are not present in the population at all.* The estimated number of outperformers drops from 118 to 35 ($p_1^+ = 1.85\%$) and the estimated number of underperformers drops from 80 to 1. Neither 35 nor 1 are significant statistically or practically.

Since this study's sample has a significant overlap with that of Barras et al. (2010), it is very likely that the overdispersion effect of similar magnitude is present in their sample. It means that Barras et al. (2010) overestimated the percentage of skilled and unskilled funds in the population as well. Under the empirical null, the percentage of

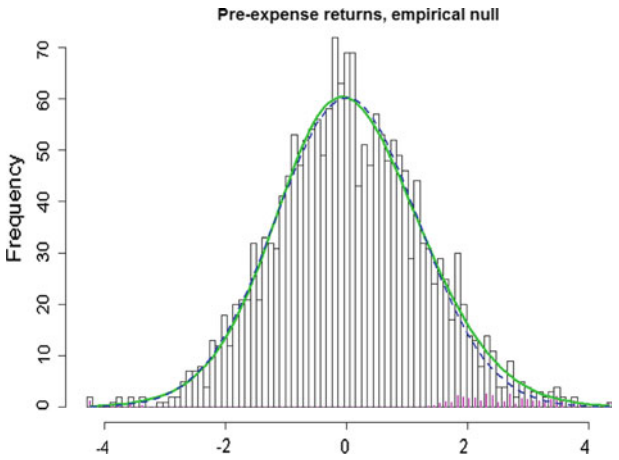

Fig. 2 Estimated densities for pre-expense returns and empirical null. The model (2.2.1) is applied to pre-expense returns of 1876 funds. The histogram represents z-values obtained according to (2.2.2) and (2.2.3). The *solid* (*green*) *curve* is $\hat{f}(z)$, an estimate of mixture density in the structural model (2.4.5). The *dashed* (*blue*) *curve* is $\hat{p}_0 \cdot \hat{f}_0(z)$, an estimate of the null component of mixture density. Since the empirical null is used, $f_0(z) = \varphi(z|\delta_0, \sigma_0^2)$where parameters δ_0 and σ_0^2 are estimated (see Table 4). The *thin* (*pink*) *dashes* in non-central bins are "thinned counts". Each thinned count is equal to the full count in the corresponding bin times $\hat{p}_1 \hat{f}_1(z)$, an estimate of alternative component in the structural model. Thinned counts reflect the proportion of alternative cases in each bin

outperformers in Barras et al. (2010) will probably be greater than 1.85%, but only because of better mutual fund performance prior to 1993.

In addition, the power is quite poor: fdr is above 0.2 everywhere, and EfdrRight = 0.712. Even if we assume that the 35 outperformers are indeed present in the population, to select 50% of outperformers (about 17 out of 35), one has to tolerate FdrRight of 0.6 by selecting about 26 false discoveries as well. The "Top 43" list of funds will have 26 false entries, $43 = 17 + 26$. To obtain decent power (EfdrRight = 0.2), it would take an unrealistic 43 years of data per fund.

The power measures used above assume that the proportion of outperformers in the population, p_1^+, is fixed at its estimated value, and it is only the number of observations per fund, T, that is subject to change. One may ask whether the estimate of p_1^+ itself is adequate. Kothari and Warner (2001) provide some basic results as to that. They create a large number of artificial mutual funds whose alphas are zero by construction. Then a certain level of alpha (between 1 and 15% per year) is introduced into all funds and measured via a number of performance evaluation models, including Carhart. The test rejection rate serves as a rough measure of test quality. For instance, given 3% alpha and Carhart model, 80% of tests reject the null under 5% test level.

Our approach uses a similar idea: from the original pre-expense data we select 122 funds with z-values in the interval $[-0.9; 0.9]$. According to both theoretical and empirical null-based results above, these funds have zero alphas. We introduce outperformance of α% per annum by adding $\alpha/12$ to each monthly return. Then we re-estimate p_1^+ based on the empirical null. The estimated number of outperformers

Table 5 Estimated number of outperformers in the population versus annual alpha

Annual alpha (%)	Estimated number of outperformers less 35	EfdrRight
1	−17 out of 122	0.601
3	−12 out of 122	0.886
5	42 out of 122	0.530
7.5	85 out of 122	0.308
10	105 out of 122	0.248
15	124 out of 122	0.171

This table reflects one's ability to estimate the proportion of outperformers in the population. According to the empirical null-based analysis, the original pre-expense dataset contains 35 outperforming funds. We select another 122 funds that have z-values close to zero and add a positive value of alpha (shown in the first column) to their returns. Then we repeat the empirical null-based estimation according to the structural model (2.4.5). The estimated number of outperformers less 35 is reported in the second column and compared to the target value of 122. The last column shows EfdrRight, a power statistic (Sect. 2.4.4)

minus 35 (the estimated number of outperformers in the original sample) will give us the idea of how many of 122 outperformers are recognized as present in the population. The advantage of this approach over that in Kothari and Warner (2001) is twofold. First, we can work with real funds as opposed to artificial ones. Second, we make a direct comparison of the true and estimated values of p_1^+, whereas the test rejection rate does not provide that information.

The results in Table 5 are very discouraging. First, for alpha between 1 and 3%, the estimated number of outperformers is even less than in the original sample. This counterintuitive result occurs because the corresponding 122 z-values shift to the right, but still remain relatively close to zero. The central portion of the histogram widens which increases $\hat{p}_0$ and reduces $\hat{p}_1^+$. It is clear that the majority of outperformers with economically significant alphas up to 5% p.a. are not included into $\hat{p}_1^+$ at all. Those funds who do manage to make it into $\hat{p}_1^+$ appear to be hardly separable from the rest: EfdrRight is well above 0.2 for all values of alpha except for an unrealistic 15% per annum.

3.3 Net performance, theoretical null

In Barras et al. (2010), the corresponding estimates of p_0, p_1^+, p_1^- are 75.4 (2.5), 0.6 (0.8), 24.0 (2.3). Again, there is a good amount of intersection of corresponding confidence intervals for p_0, p_1^+, p_1^- (Table 6, also see Fig. 3). The estimated number of outperformers (9 funds out of 1911) is not statistically or practically significant. This is different from the findings Kosowski et al. (2006) who find a sizable minority of funds that generate a significant amount of wealth per year. It is not possible to perform a more quantitative comparison because Kosowski et al. (2006) do not estimate the proportion of outperformers explicitly. Besides, their dataset includes a number of funds that have "Balanced and Income" as investment objectives. Such funds are not included neither in our dataset nor in Barras et al. (2010).

EfdrLeft = 0.35 (still well above 0.2), and the power is not good. In particular, 54% of underperformers (295 out of 547) are identified with FdrLeft = 0.2 (Table 7).

Table 6 Estimation results for net returns and theoretical null

	$\hat{p}_0$, %	$\hat{p}_1^+$, %	$\hat{p}_1^-$, %
	70.91 (1.22)	0.45 (0.86)	28.64 (0.86)
95% CI	(68.52; 73.30)	(−1.24; 2.14)	(26.95; 30.33)
Number of funds	1,355	9	547
Zero interval	$\hat{\lambda}$		
(−1.4; 1.4)	0.1615		
Efdr	EfdrRight	EfdrLeft	
0.35	0.49	0.35	

The table shows the estimated proportions of zero-alpha, outperforming, and underperforming mutual funds ($\hat{p}_0$, $\hat{p}_1^+$, $\hat{p}_1^-$). The corresponding structural model is (2.4.5). Standard errors are provided in parentheses, along with 95% confidence intervals in the 3rd line. The 4th line shows the estimated number of zero-alpha, outperforming, and underperforming funds in the population of 1911 funds. To provide more details about the estimation procedure, we report the boundaries of zero interval and the value of $\hat{\lambda}$ (see Sect. 2.4.3). The power statistics, Efdr, EfdrRight, EfdrLeft (Sect. 2.4.4) are in the bottom line

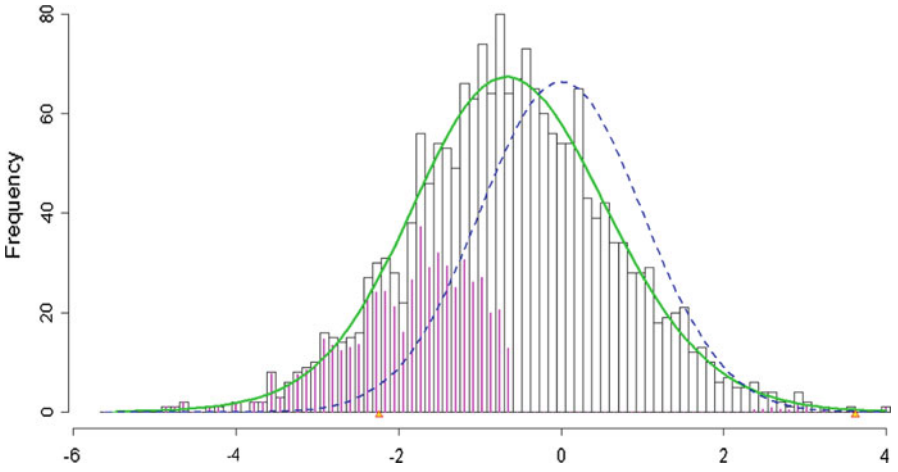

Fig. 3 Estimated densities for net returns and theoretical null. The model (2.2.1) is applied to net returns of 1911 funds. The histogram represents z-values obtained according to (2.2.2) and (2.2.3). The *solid* (*green*) *curve* is $\hat{f}(z)$, an estimate of mixture density in the structural model (2.4.5). The *dashed* (*blue*) *curve* is $\hat{p}_0 \cdot \varphi(z|0,1)$, an estimate of the null component of mixture density. The *thin* (*pink*) dashes in non-central bins are "thinned counts". Each thinned count is equal to the full count in the corresponding bin times $\hat{p}_1 \hat{f}_1(z)$, an estimate of alternative component in the structural model. Thinned counts reflect the proportion of alternative cases in each bin

Despite the low power, a high proportion of underperformers makes it much easier to construct sizable "Bottom lists" of decent quality: e.g., the "Bottom 181" list has FdrLeft of 0.11 which corresponds to about 20 false discoveries in the list.

Increasing T to 15 years per fund reduces EfdrLeft from 0.35 to 0.29, and only the unrealistic 26 years of data per fund brings EfdrLeft to 0.2. Still, if it is possible to extend back the sample and obtain 15 years of data per fund, it pays off because the identifiable (under FdrLeft = 0.2) proportion of underperformers increases from 54 to

Table 7 Identified underperformers and false discoveries versus FdrLeft for net returns and theoretical null

FdrLeft FdrLeft	Proportion of identified underperformers, %	Number of identified underperformers	Number of false discoveries
0.11	29	161 out of 547	20
0.2	54	295 out of 547	75
0.3	80	438 out of 547	188
0.4	96	525 out of 547	350
0.5	100	547 out of 547	547

This table describes one's ability to detect underperformers. For each level of FdrLeft (defined by 2.4.8) in the first column, the second and third columns specify the proportion and number of underperformers (out of total 547) whose z-values fall below the left-hand-side cutoff determined by FdrLeft. The last column shows how many useless funds (false discoveries) have to be tolerated because they also fall below the cutoff

72% (394 funds out of 547). For 26-year sample, that proportion is 90% (492 funds out of 547).

3.4 Net performance, composite empirical null

For net returns data, it is not possible to fit the simple empirical null directly as in Sect. 3.2 because p_0, the proportion of funds with zero performance, is far below 0.9 (see Appendix). Therefore, we recourse to the composite empirical null (2.4.4).

From the results in the previous section, we would expect the optimal z_+ to be at least 1.4. Efron (2004b) suggests a non-symmetrical parametric null, such as split-normal $f_0(z) = SN(\delta_0, \sigma_1^2, \sigma_2^2)$, in order to avoid the influence of the left-tail z's on the inference in the right tail. However, fitting a split-normal distribution along with normal $f_0(z) = \varphi(z|\delta_0, \eta_0^2)$ for $z_- = -4$ and $z_+ \in [1.4; 2.2]$ showed that the corresponding null components $\hat{p}_0 \hat{f}_0(z)$ are virtually identical and $\varphi(z|\delta_0, \eta_0^2)$ is quite adequate for modeling the composite null.

Here we expect a much higher power to identify outperformers than in Sect. 3.2. First, the mean of null density is shifted to the left by a sizable value of 0.624. Secondly, inclusion of z-values in $[-4; -1.4]$ reduced the standard error of $\hat{p}_0$ by 0.38% without causing any increase in the bias in the right tail. Inclusion of z-values in [1.4; 1.6] reduces $s.e.(\hat{p}_0)$ by another 0.14% and overall it drops from 1.22% (Table 6) to 0.7% (Table 8, also see Fig. 4).

In spite of this, the estimated number of outperformers grows from 9 to only 15 (still practically insignificant) and is not statistically different from zero. The only explanation is that the estimated null distribution $\hat{f}_0(z) = \varphi(z|\hat{\delta}_0, \hat{\eta}_0^2)$ reflects the fact that σ_0^2 in (2.4.4) is much greater than 1. Taking that overdispersion into account drastically reduces the final estimated number of outperformers. It eliminates all the benefits we hoped to get from the composite empirical null. In addition, EfdrRight is above 0.725 and the power is abysmal. In particular, the list of "Top 15" performers has FdrRight = 0.58 that amounts to about 9 false discoveries in the list.

Table 8 Estimation results for net returns and composite empirical null

	$\hat{p}_0$, %	$\hat{p}_1^+$, %
	99.21 (0.7)	0.79 (0.7)
95% CI	(97.84; 100.58)	(−0.58; 2.16)
Number of funds	1896	15
Zero interval	$\hat{\lambda}$	
(−4; 1.6)	0.055	
$\hat{\delta}_0$	$\hat{\eta}_0$	EfdrRight
−0.624 (0.033)	1.229 (0.028)	0.725

The table shows the estimated proportions of non-outperforming and outperforming funds ($\hat{p}_0$, $\hat{p}_1^+$), along with 95% confidence intervals in the 3rd line. The corresponding structural model is (2.4.4). The 4th line shows the estimated number of non-outperforming and outperforming funds in the population of 1911 funds. To provide more details about the estimation procedure, we report the boundaries of zero interval and the value of $\hat{\lambda}$ (Sect. 2.4.3). Since the composite empirical null, $f_0(z) = \varphi(z|\delta_0, \eta_0^2)$, is utilized, we report the estimates of δ_0 and η_0^2. The power statistic, EfdrRight, (Sect. 2.4.4), is also shown. Figures in parentheses denote the standard deviation of the different estimators

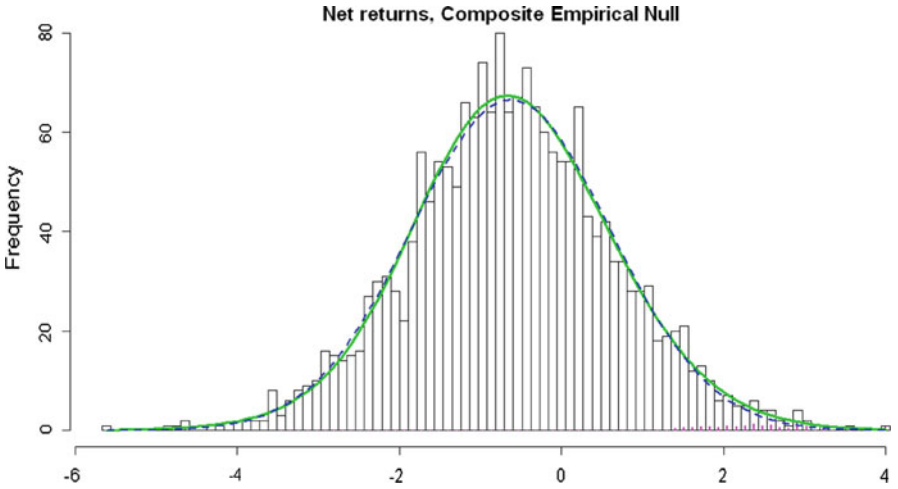

Fig. 4 Estimated densities for net returns and composite empirical null. The model (2.2.1) is applied to net returns of 1911 funds. The histogram represents z-values obtained according to (2.2.2) and (2.2.3). The *solid* (*green*) *curve* is $\hat{f}(z)$, an estimate of mixture density in the structural model (2.4.4). The *dashed* (*blue*) *curve* is $\hat{p}_0 \cdot \hat{f}_0(z)$, an estimate of the null component of mixture density. Since the empirical null is used, $f_0(z) = \varphi(z|\delta_0, \eta_0^2)$ where parameters δ_0 and η_0^2 are estimated (see Table 8). The *thin* (*pink*) dashes in right hand side bins are "thinned counts". Each thinned count is equal to the full count in the corresponding bin times $\hat{p}_1 \hat{f}_1(z)$, an estimate of alternative component in the structural model. Thinned counts reflect the proportion of alternative cases in each bin

3.5 Net outperformance versus mutual fund investment objective

Using the method outlined in Sect. 2.4.5, let us take a look at how the net outperformance depends on the investment objective category. We apply the structural model (2.4.4) like we did in the last section.

Table 9 Net outperformance versus investment objective under composite empirical null

Category	Number of funds	p-value for H_0: $f_{A0}(z) = f_{B0}(z)$	p-value for $H_0: fdr_A(z) = fdr(z)$	Number of outperformers	Proportion
G	886	0.7083	0.5606	7	0.79%
GI	398	0.9698	0.0006	0	0%
AG	627	0.6997	0.0079	19	3%
Population	1911	n/a	n/a	15	0.79%

The table shows the number of outperformers in each of three investment objective categories estimated using the methodology outlined in Sect. 2.4.5. The 2nd column shows the number of "Growth" (G), "Growth and Income" (GI), and "Aggressive Growth" (AG) funds. The 3rd column reports the p-value for the hypothesis that the null distribution in the corresponding investment objective category is the same as in the entire population of 1911 funds. The 4th column shows the p-value for the hypothesis that the false discovery rate in the investment objective category is the same as in the population. Finally, the number and proportion of outperformers are reported in the last two columns

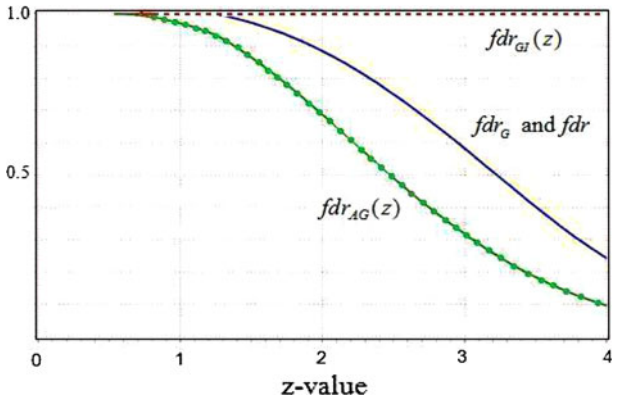

Fig. 5 False discovery rate versus mutual fund investment objective. To investigate how net outperformance depends on the mutual fund investment objective, we use the structural model (2.4.4) and the method outlined in Sect. 2.4.5. The *dashed* (*red*) *line* shows $fdr_{GI}(z)$ for Growth & Income funds, which is equal to 1 for all values of z. The *solid* (*blue*) *line* shows $fdr_G(z)$ for Growth funds, which is no different from $fdr(z)$ for the entire population of funds. The *beaded* (*green*) *curve* reflects $fdr_{AG}(z)$ for Aggressive Growth funds

Column 3 of Table 9 shows that the hypothesis $f_{A0}(z) = f_{B0}(z)$ is not rejected for any category, and it means that we do not have to re-run the entire analysis for each category separately. Column 4 suggests that $fdr_{AG}(z) \neq fdr(z)$ and $fdr_{GI}(z) \neq fdr(z)$, but we fail to reject $fdr_G(z) = fdr(z)$.

Figure 5 shows the curves corresponding to $\hat{fdr}(z)$ (which coincides with $\hat{fdr}_G(z)$), $\hat{fdr}_{GI}(z)$, and $\hat{fdr}_{AG}(z)$. Apparently, there are no skilled managers in GI group because for any z $\hat{fdr}_{GI}(z) = 1$. Using the estimate $\hat{fdr}_{AG}(z)$ and $\hat{fdr}_G(z)$, we conclude that there are 19 outperformers among 627 AG funds and 7 outperformers among 886 G funds. Therefore, while the percentage of outperformers is 0.79% in the population (15 out of 1,911), it is about 3% in AG group, 0.79% in G group and 0% in GI group.

While $\hat{fdr}(z)$ is always above 0.24, $\hat{fdr}_{AG}(z)$ is under 0.2 for $z > 3.56$. Unfortunately, only one AG fund has $z \geq 3.56$ and can be identified as outperformer. Even if we raise the fdr cutoff from 0.2 to 0.4, only 4 out of 19 AG outperformers are identified. Even a relatively superior AG group is unable to produce a number of identifiable outperformers that would be significant for investment purposes.

The results in Barras et al. (2010) for the same three groups are quite similar: AG funds are the best, G funds are similar to the entire population, and GI funds are the worst. The estimated proportions of outperformers are 3.9, 0, 0%, correspondingly, although they do not investigate whether the difference is statistically significant. We do find a statistically significant difference, which could have an interesting implication. We know that the AG group has the highest proportion of "growth" stocks and the lowest proportion of "value" stocks. For the GI group, it is the other way round, and the G group is somewhere in between AG and GI. In model (2.2.1) the book-to-market factor, h_i, is also known as "growth versus value" factor. Hence, there should not be any difference in risk-adjusted performance w.r.t. "growth versus value" dimension. Our results indicate otherwise, which, at least in terms of statistical significance, could imply that Carhart's model does not quite explain the cross-sectional variation of excess returns.

3.6 Short-term net performance

The long-term results of mutual fund net performance are quite disappointing because the number of outperformers is never practically significant: 12 in Barras et al. (2010), and the best result for this study is 26 (7 G and 19 AG funds discovered in Sect. 3.5).

However, the short-term performance may be better, as suggested by Barras et al. (2010). They partition the data into six non-overlapping subperiods of 5 years each, from 1977 to 1981 to 2002–2006. If a fund has 60 observations on a subperiod, it is treated as a separate "fund" with 5-year history. They thus increase the number of estimated alphas from 2,076 to 3,311 and $\hat{p}_1^+$ goes up from 0.6 (0.8)% to a statistically significant 2.4 (0.7)%, correspondingly. This is interpreted as evidence for superior "short-term" performance that exists for a while and gradually disappears in the long-run equilibrium. Barras et al. (2010) point out that if the equilibrium model holds, the negative performance has to disappear just as well, which is not observed in reality.

Our major concern about that analysis is that drastically reducing the number of observations per fund is very likely to increase the overdispersion of z-values. That alone could explain a higher estimated percentage of short-term outperformers and, therefore, the utilization of empirical null is even more justified here.

Similarly to Barras et al. (2010), we partition our dataset into three non-overlapping 58-month subperiods. If a fund has 50 or more observations on a subperiod, it is treated as a separate "short-term fund". In the end, there are 3,636 of such "funds". Applying the theoretical null results in $p_1^+ = 0.81(0.52)\%$ (29 outperformers), both statistically and practically insignificant.

Applying instead the composite empirical null, as in Sect. 3.4, we hope that more positive cases will be identified. However, as predicted above, the overdispersion is so severe that the estimated number of outperformers not only fails to go up but actu-

Table 10 Estimation results for "short-term" net performance under composite empirical null

	p_0, %	p_1^+, %
	99.63 (0.69)	0.37 (0.69)
95% CI	(98.28; 100.98)	(−0.98; 1.72)
Number of funds	3623	13
Zero interval	λ	
(−3.5; 1.6)	0.055	
δ_0	η_0	EfdrRight
−0.467 (0.026)	1.254 (0.024)	0.877

The table shows the estimated proportions of non-outperforming and outperforming funds ($\hat{p}_0$, $\hat{p}_1^+$), along with 95% confidence intervals in the 3rd line. The corresponding structural model is (2.4.4). The 4th line shows the estimated number of non-outperforming and outperforming funds in the population of 3,636 "short-term" funds. To provide more details about the estimation procedure, we report the boundaries of zero interval and the value of λ (Sect. 2.4.3). For the composite empirical null, $f_0(z) = \varphi(z|\delta_0, \eta_0^2)$, we report the estimates of δ_0 and η_0^2. The power statistic, EfdrRight, (Sect. 2.4.4), is also shown. Figures in parentheses denote the standard deviation of the different estimators

ally drops from 29 to 13 funds (Table 10). Therefore, we conclude that there is no compelling evidence of short-term outperformance in 1993–2007.

3.7 Discussion

In this section, we summarize and discuss the results obtained in this study.

As indicated in Sect. 2.4.2, the results obtained from our approach under the theoretical null are directly comparable to the output of Barras et al. (2010). It is reassuring that despite the difference in the employed datasets, when we use the theoretical null (Sects. 3.1, 3.3, 3.6), our findings are consistent with the Barras et al. (2010).

The switch to the empirical null is well grounded. The findings in Sect. 3.2 provide compelling evidence that the theoretical null is biased, because the test statistics exhibit both statistically and practically significant overdispersion. When the overdispersion is taken into account, the inference changes dramatically: over 10% of funds are either skilled or unskilled on pre-expense basis under the theoretical null, but under the empirical null that proportion is not distinguishable from zero.

As noted in Sect. 3.2, the empirical null results in lower bias and higher variance than the theoretical null. For practical purposes, it is convenient to monitor the standard error $s.e.(\hat{p}_0)$. Our results suggest that, all other things being equal, the structural model of Efron has more precision than the bootstrap-based approach of Barras et al. (2010). For instance, for pre-expense returns, Barras et al. (2010) report $s.e.(\hat{p}_0) = 2.7\%$, whereas our result in Sect. 3.1 is $s.e.(\hat{p}_0) = 0.75\%$, a difference of factor of 3.6 (the number of funds is about the same). Moreover, when we switch to the empirical null, we get $s.e.(\hat{p}_0) = 0.99\%$ which, contrary to expectations, is still significantly less than 2.7%. Because the proportion of outperformers is always small (well under 10%), such gain in precision is practically significant.

The empirical Bayes method also allows us to test the net performance under a more powerful composite empirical null. Because overdispersion is taken into account also,

even under that powerful setting the number of outperformers proves neither statistically nor practically significant (Sect. 3.4). Hence, the evidence for the absence of net outperformance in mutual fund industry in 1993–2007 is substantially reinforced.

The investment objective analysis in Sect. 3.5 tries to add even more power to the composite empirical null by using the investment objective category as a control variate. Thanks to the rigorous and efficient approach of Efron, we can test whether the distinct null distributions are necessary for each investment category, and finding that this is not the case, we avoid redundant parameters. Qualitatively, our findings are consistent with Barras et al. (2010): AG funds are the best and GI funds are the worst. In addition, we obtain statistical evidence for the difference in performance across investment objectives. This result seems to be at odds with the "growth versus value" factor being present in the Carhart model. Despite the slight increase in power caused by adjusting for the investment objective, the estimated number of net outperformers is still practically insignificant (26 out of 1,911).

The results for "short-term" net performance in Sect. 3.6 are also weak. Even when overdispersion is not taken into account (under theoretical null), there is no evidence of short-term outperformance in 1993–2007. The fact that the number of outperformers under the composite empirical null is even less is suggestive of severe overdispersion of "short-term" test statistics. Therefore, a significant part of "superior short-term performance" effect reported in Barras et al. (2010) must have come from the bias of the theoretical null.

If we are interested in practical applications of mutual fund performance evaluation, high power is desirable. A sharp tool to separate good individual funds from the rest can be useful to an individual investor who is not likely to be able to invest in more than one fund. Similarly, high power is useful to determine whether an individual fund manager should be rewarded or punished. One can argue that, once there is an opportunity to invest in a large number of funds, it only matters that a decent proportion of outperformers is present in the portfolio, and then even a low power study will be useful. There is some indication, from the FDR-based fund portfolio of in Barras et al. observed from 1980 to 2006 (alpha of 1.45% p.a. with p-value of 0.04), that such a strategy could be hindered today, given the significant decline, after 1990, in mutual fund industry performance.

The detailed power analysis showed that regardless of whether the utilized null is theoretical or empirical, and whether we are interested in picking winners or losers, our ability to do so is very limited. In particular, the "Top N performers" lists (for both pre-expense and net returns) have low quality: they are likely to contain a large proportion of funds that are not true outperformers. It is a result of both low proportion of outperformers and low power. However, thanks to a high proportion of net underperformers, we can construct sizable lists of "Bottom N net performers" with decent quality.

Extending the sample back (e.g., Barras et al. (2010) sample with 32-year span) can increase the number of funds but is not likely to produce many more observations per fund. For this study, the span is 14 1/2 years with the mean of l0 3/4 years per fund. Although 10% of the funds span the entire 14 1/2 years, it is still unlikely to obtain a dataset with, say, more than 15 years of observations per fund on average, regardless of how far back it is extended. Therefore, power statistics obtained when there are 15

years of observations for each fund can be considered the upper bounds for the power. Unfortunately, even having 15 years per fund does bring Efdr to 0.2 (the best result is EfdrLeft = 0.29 in Sect. 3.3), and the power is still poor. Therefore, an unsatisfactory power is inherent to both the current and Barras et al. (2010) despite a much larger time span of the latter.

Besides, a long-lived fund is likely to be managed by a few successive portfolio managers. According to John Bogle, founder of The Vanguard Group, "…the tenure of the average portfolio manager is just five years…". Kothari and Warner (2001) also indicate that the investor is likely to consider only from 3 to 5 years of fund history. Practically speaking, there are reservations as to whether the 6–15 year-old data are relevant for investors. At the same time, reducing the history is bound to reduce the power to a level where the study is absolutely uninformative. For instance, in Sect. 3.4, with with $T = 10\ 3/4$ years, EfdrRight = 0.725 (very poor). After T is reduced to about 5 years (Sect. 3.6), the power gets even worse: EfdrRight = 0.877.

The power issues above have the following implication: assuming that the proportion of outperformers is estimated correctly, it is still very hard to separate outperformers from the rest. Section 3.2 indicates that there is a big problem with estimating the proportion of outperformers itself. Consider a good fund manager whose true alpha is 5% p.a. According to the results in Sect. 3.2, such manager is not even likely to be included in $\hat{p}_1^+$, that is, its z-value falls into a bin where $f\hat{d}r(z) = 1$. Even when the good manager is recognized as present in the population, a high value of EfdrRight will not allow the investor to separate him from the rest.

It appears that any mutual fund study that is based on monthly data and a similar multifactor performance evaluation model (e.g., CAPM, Fama and French) is likely to be underpowered. While such models provide a theoretically grounded way to adjust the performance for risk, their finite sample properties cause difficulties in their implementation. In particular, identification of market efficiency from the lack of evidence of risk-adjusted outperformance, means implicitly that a fairly precise tool has been used to search for such evidence. In contrast, our results are not inconsistent with the undetected presence of many good funds.

One notes that performance measures of the above types are avoided in mutual fund prospectus. According to Kothari and Warner (2001), a possible way out is to use trade-based performance measures. However, the information on fund trades can be restricted, even for institutional investors. Another way to obtain an edge is presented in Mamaysky et al. (2007), who argue that it is unlikely for a single performance evaluation model to be equally good for all funds. They show that using a few competing models, combined with back-testing, can significantly improve the performance of mutual fund portfolios. Using such an approach coupled with a multiple inference procedure can be an interesting topic for future research.

4 Conclusion

When evaluating the performance of a large number of mutual funds simultaneously, one has to weed out false discoveries. This task is fairly straightforward when the performance test statistics are independent across funds. However, independence is

unlikely to hold for real data. On the other hand, there are not enough years of data to estimate the dependence structure of test statistics directly. In addition, a misspecified performance evaluation model can bias the results. Is there a way around these problems? The state-of-the-art approach of Efron offers a viable alternative. It also helps us investigate the usually neglected issue of statistical power in a mutual fund study.

In this paper, we analyze the performance of about 2,000 US equity mutual funds over a period of 14 1/2 years. In contrast to existing studies, we neither assume independence of test statistics across funds, nor do we try to estimate the dependence structure based on the data that are clearly insufficient for that purpose. In addition, certain features of Efron's approach make it more powerful and precise, as well as being able to perform a rigorous and efficient analysis of subgroups of funds.

Our analysis suggests that it is not appropriate to treat the test statistics as mutually independent with pre-specified null distribution. The data indicate that doing so leads to both statistically and practically significant bias, when the proportions of both under- and outperformers are overestimated. Despite the advantages of Efron's approach, we fail to identify a practically or statistically significant proportion of net outperformers. The power analysis shows that, due to the nature of data and the performance evaluation model (monthly dataset, a multifactor model), the study has a very low power. That is, we are hardly able to detect and single out the true out- or underperformers. It would require an unrealistically large history of data and/or level of outperformance to increase the power to a decent level.

Appendix

For the empirical null, $f_0(z)$ can be approximated by a parametric distribution, such as symmetrical normal $N(\delta_0, \sigma_0^2)$ or skewed split-normal $SN(\delta_0, \sigma_1^2, \sigma_2^2)$. For a given zero interval, the parameters of interest are estimated with the method of moments (denoted CME in *locfdr*).

An additional restriction $p_0 \geq 0.9$ has to hold when we use the empirical null. Efron (2004a) provides theoretical and numerical results that justify the restriction: if $p_0 \geq 0.9$ and the theoretical null is valid, then the MLE or CME estimates of δ_0 and σ_0 have to be close to 0 and 1, respectively. If they are not, it implies that the theoretical null is inadequate. If $p_0 < 0.9$ then the estimates of (δ_0, σ_0) can be significantly different from (0, 1) even when the theoretical null is valid. Hence, if one wants to check whether a switch to the empirical null is necessary, first he has to make sure that $p_0 \geq 0.9$. Using an empirical null when the theoretical null is valid and $p_0 < 0.9$ has an effect of ignoring a lot of alternative cases. For instance, if we were to use model (2.4.5) for net returns, we would end up underestimating the proportion of underperformers. However, since the proportion of net outperformers is small, it is still possible to use the empirical null for that, which we did in Sect. 3.4.

The choice of the zero interval itself is a bias-variance tradeoff problem: for a large interval, the estimate of p_0 (and, if applicable, the parameters of the empirical null) have low variance but a high bias since many non-null cases are likely to fall into the wide zero interval. For a narrow zero interval, the bias is small, but the estimates of p_0

and other parameters have large variance because too few z-values fall into$(z_-; z_+)$. To solve the problem, we consider the error of $\hat{p}_0 \hat{f}_0(z)$ scaled by $1/f(z)$:

$$Error(z) = \frac{1}{f(z)} \left[p_0 f_0(z) - \hat{p}_0 \hat{f}_0(z) \right] \tag{A.1}$$

The optimal zero interval is where the integrated $MSE(Error(z))$ is at the minimum, so we have to estimate the squared bias and variance. The *locfdr* package does not provide a direct estimate of $MSE(Error(z))$, and we are going to use some proxies to obtain the shape of bias-variance tradeoff curve.

First, we use the bias on the zero interval as a proxy for overall bias. On the zero interval we have

$$\begin{aligned} p_0 f_0(z) &= f(z), \text{ and} \\ Error(z) &= \frac{1}{f(z)} \left[p_0 f_0(z) - \hat{p}_0 \hat{f}_0(z) \right] = 1 - \frac{\hat{p}_0 \hat{f}_0(z)}{f(z)} \end{aligned} \tag{A.2}$$

The mixture density $f(z)$ is unknown, but the expected error can be estimated by using an unbiased estimator of $f(z)$ which is obtained in *locfdr* via Poisson regression over the entire z axis. The *locfdr package* also produces the estimates $f\hat{d}r(z)$ and $Var\left[\log(f\hat{d}r(z))\right]$. As a result, the estimate of average squared bias is (all integrals are computed as corresponding sums):

$$B\hat{i}as^2_\lambda = \frac{1}{z_+ - z_-} \int_{z_-}^{z_+} \left(1 - f\hat{d}r(z)\right)^2 dz \tag{A.3}$$

The error variance at point z will be

$$Var[Error(z)] = Var\left[f\hat{d}r(z) \right] \tag{A.4}$$

We are going to use the available $Var\left[\log\hat{(}fdr(z))\right]$ instead and then get the estimate of overall variance as:

$$V\hat{a}r_\lambda = \int_{-\infty}^{\infty} Var\left[\log\hat{(}fdr(z))\right] dz \tag{A.5}$$

For the theoretical null, $f_0(z)$ is not estimated, $Var(\hat{f}(z))$ does not depend on λ and its magnitude is much larger than that of $Var(\hat{p}_0 \cdot \varphi(z|0, 1))$ (Efron 2005). For that reason, we are going to use $Var_\lambda(\hat{p}_0)$ instead of (A.5) for the theoretical null. For the empirical null, we are using the full version (A.5).

$V\hat{a}r_\lambda$ and $B\hat{i}as^2_\lambda$ are not on the same scale, but we can still use them to estimate the shape of MSE curve. Following (Storey and Tibshirani 2001), we divide each

estimate by its median over the range of the smoothing parameter to get the value of bias-variance tradeoff, BVT_λ:

$$BVT_\lambda = \frac{Var_\lambda}{median_{\lambda'}(Var_{\lambda'})} + \frac{Bias^2_\lambda}{median_{\lambda'}\left(Bias^2_{\lambda'}\right)} \tag{A.6}$$

BVT_λ is not equal to the integrated $MSE(Error(z))$, but it reflects the shape of MSE curve, and the optimal value of λ is determined by minimizing BVT_λ over the range of λ. In Sect. 3, we do not provide the plots of BVT_λ to save space, but, in all cases considered, the plot of BVT_λ has a familiar U-shape. The estimated parameters of empirical component are not very sensitive to changes in the limits of $(z_-; z_+)$ which is consistent with findings of Barras et al. (2010).

Efron's method and *locfdr* package are designed for a two-component model like (2.4.2) and (2.4.4), but the three-component model (2.4.5) has a problem: the *locfdr* package produces the estimate of $p_1 f_1(z)$, but its decomposition into positive $p_1^+ f_1^+(z)$ and negative $p_1^- f_1^-(z)$ components is not identified. To get around this issue, note that $f_1^-(z)$ is a (possibly continuous) mixture of normal densities $\varphi(z|\alpha, \sigma_0^2)$, $\alpha<0$. Because all of the normal densities in the mixture have negative means, $f_1^-(z)$ is non-increasing for positive z's. Typically, the estimation produces $f\hat{d}r(z) = \hat{p}_0\hat{f}_0(z)/\hat{f}(z) = 1$ in some interval $(-l; l)$, such as $(-0.4; 0.4)$. It implies that $\hat{f}_1(z)$, $\hat{f}_1^-(z)$ and $\hat{f}_1^+(z)$ are equal to zero on $(-l; l)$. Hence, $\hat{f}_1^-(z)$ cannot have support for $z > l$ and $\hat{f}_1^-(z) = 0\,\forall z > 0$. Similarly, $\hat{f}_1^+(z) = 0\,\forall z < 0$. Therefore, while in theory some $\alpha < 0$ can produce positive z's, the estimation results imply that such z's can be produced only by $\alpha \geq 0$, and negative z's can only be produced by $\alpha \leq 0$. Hence, $fdr^+(z)$defined in (2.4.7) is equal to 1 for negative z's and is equal to $f\hat{d}r(z)$ for nonnegative z's. A similar conclusion applies to $fdr^-(z)$.

Then, the value of p_1^+ is estimated as follows:

$$\hat{p}_1^+ = \frac{\int_0^\infty \left[1 - f\hat{d}r(z)\right]dz}{\int_0^\infty \frac{\hat{f}_1^+(z)}{\hat{f}(z)}dz} \tag{A.7}$$

and similarly for $\hat{p}_1^-$.

For the two-component model, $s.e.(\hat{p}_0) = s.e.\left(\hat{p}_1\right)$, but $s.e.\left(\hat{p}_1^+\right)$ and $s.e.\left(\hat{p}_1^-\right)$ for the three-component model are not available from *locfdr*. Let us assume that

$$s.e.\left(\hat{p}_1^+\right) = s.e.\left(\hat{p}_1^-\right) = \kappa \text{ and } corr\left(\hat{p}_1^+, \hat{p}_1^-\right) \leq 0 \tag{A.8}$$

then

$$Var[\hat{p}_0] \leq 2\kappa^2 \Rightarrow \kappa \geq s.e.(\hat{p}_0)/\sqrt{2} \tag{A.9}$$

The lower bound for κ is reported instead of $s.e.\left(\hat{p}_1^+\right)$ and $s.e.\left(\hat{p}_1^-\right)$ whenever the three-component model is used.

References

Ammann, M., Verhofen, M.: The impact of prior performance on the risk-taking of mutual fund managers. Ann Financ **5**, 69–90 (2009)

Avellaneda, M., Lee, J.: Statistical Arbitrage in the U.S. equities market. SSRN. http://ssrn.com/abstract=1153505 (2008). Accessed 11 July 2008

Barras, L., Scaillet, O., Wermers, R.: False discoveries in mutual fund performance: measuring luck in estimated alphas. J Financ **65**(1), 179–216 (2010)

Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc **57**(1), 289–300 (1995)

Benjamini, Y., Hochberg, Y.: On the adaptive control of the false discovery rate in multiple testing with independent statistics. J Educ Behav Stat **25**(1), 60–83 (2000)

Benjamini, Y., Yekutieli, D.: The control of the false discovery rate in multiple testing under dependency. Ann Stat **29**(4), 1165–1188 (2001)

Benjamini, Y., Krieger, A., Yekutieli, D.: Adaptive linear step-up procedures that control the false discovery rate. Biometrika **93**(3), 491–507 (2006)

Carhart, M.: On persistence of mutual fund performance. J Financ **52**(1), 57–82 (1997)

Chen, H., Jegadeesh, N., Wermers, R.: The value of active mutual fund management: an examination of the stockholdings and trades of fund managers. J Financ Quant Anal **35**, 343–368 (2000)

Cornell, B., Cvitanic, J., Goukasian, L.: Beliefs regarding fundamental value and optimal investing. Ann Financ **6**, 83–105 (2010)

Cuthbertson, K., Nitzsche, D., O'Sullivan, N.: UK mutual fund performance: skill or luck? J Empir Financ **15**, 613–634 (2008a)

Cuthbertson, K., Nitzsche, D., O'Sullivan, N.: False discoveries: winners and losers in mutual fund performance. SSRN. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1093624 (2008b)

Daniel, K., Grinblatt, M., Titman, S., Wermers, R.: Measuring mutual fund performance with characteristic-based benchmarks. J Financ **52**(3), 1035–1058 (1997)

Efron, B.: Empirical Bayes analysis of a microarray experiment. J Am Stat Assoc **96**(456), 1151–1160 (2001)

Efron, B., Tibshirani, R.: Empirical Bayes methods and false discovery rates for microarrays. Genet Epidemiol **23**, 70–86 (2002)

Efron, B.: Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. J Am Stat Assoc **99**(465), 96–104 (2004a)

Efron, B.: Selection and estimation for large-scale simultaneous inference. http://www-stat.stanford.edu/~ckirby/brad/papers/ (2004b). Accessed 25 Jan 2009

Efron, B.: Local false discovery rates. http://www-stat.stanford.edu/~ckirby/brad/papers/ (2005) Accessed 25 Jan 2009

Efron, B.: Microarrays, empirical Bayes, and the two-groups model. Stat Sci **23**(1), 1–22 (2008a)

Efron, B.: Testing the significance of sets of genes. Ann Appl Stat **1**(1), 107–129 (2007a)

Efron, B.: Size, power, and false discovery rates. Ann Stat **35**(4), 1351–1377 (2007b)

Efron, B.: Correlation and large-scale simultaneous significance testing. J Am Stat Assoc **102**(477), 93–103 (2007)

Efron, B.: Simultaneous inference: when should hypothesis testing problems be combined? Ann Appl Stat **2**(1), 197–223 (2008)

Fan, J., Fan, Y., Lv, J.: High dimensional covariance matrix estimation using a factor model. J Econom **147**(1), 186–197 (2008)

Jensen, M.: The performance of mutual funds in the period 1945–1964. J Financ **23**(2), 389–416 (1968)

Jones, C., Shanken, J.: Mutual fund performance with learning across funds. J Financ Econ **78**, 507–552 (2005)

Kosowski, R., Timmermann, A., Wermers, R., White, H.: Can mutual fund "Stars" really pick stocks? New evidence from a bootstrap analysis. J Financ **61**(6), 2551–2596 (2006)

Kothari, S., Warner, J.: Evaluating mutual fund performance. J Financ **56**(5), 1985–2010 (2001)

Mamaysky, H., Spiegel, M., Zhang, H.: Improved forecasting of mutual fund alphas and betas. Rev Financ **11**, 359–400 (2007)

Otamendi, J., Doncel, L., Grau, P., Sainz, J.: An evaluation on the true statistical relevance of Jensen's alpha trough simulation: An application for Germany. Econ Bull **7**(10), 1–9 (2008)

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

Romano, J., Shaikh, A., Wolf, M.: Control of the false discovery rate under dependence using the bootstrap and subsampling. University of Zurich, Working paper no. 337. http://ssrn.com/abstract=1025410 (2007)

Romano, J., Shaikh, A., Wolf, M.: Formalized data snooping based on generalized error rates. Econom Theor **24**(2), 404–447 (2008)

Storey, J.: A direct approach to false discovery rates. J Roy Stat Soc B **64**, 479–498 (2002)

Storey, D.: The positive false discovery rate: a Bayesian interpretation and the q-value. Ann Stat. **31**(6), 2013–2035 (2003)

Storey, D., Tibshirani, R.: Estimating false discovery rates under dependence, with applications to DNA microarrays. http://www.genomine.org/publications.html (2001). Accessed 25 Nov 2009

Storey, D., Tibshirani, R.: Statistical Significance for genomewide studies. Proc Nat Acad Sci USA **100**, 9440 (2003)

Storey, J., Taylor, J., Siegmund, D.: Strong control, conservative point estimation and simultaneous conser-vative consistency of false discovery rates: a unified approach. J Roy Stat Soc **66**, 187 (2004)

Turnbull, B.: Optimal estimation of false discovery rates. http://www.stanford.edu/~bkatzen/ (2007). Accessed 25 Jan 2009

van der Laan, M., Hubbard, A.: Quantile function based null distribution in resampling based multiple testing. Stat Appl Genet Mol Biol. **5**, article 14 (2005)

White, H.: A reality check for data snooping. Econometrica **68**, 1097–1126 (2000)

Yekutieli, D., Benjamini, Y.: Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. J Stat Plan Infer **82**, 171 (1999)