

# Mutual fund performance: false discoveries, bias, and power

Nik Tuzov and Frederi Viens

**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 to falsely 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 not enough 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 significant number of successful funds. A detailed analysis reveals a very low power of the study. Even if outperformers are present in the sample, it requires too many years of data to single them out.

**Keywords:** Mutual fund, performance evaluation, false discovery, multiple inference, statistical power

**JEL Classification:** C10 G10 G20

## 1 Introduction

The studies of performance of mutual funds (referred to herein as MF), go back at least 40 years (Jensen 1968), and this area is still of interest to researchers (Ammann and Verhofen 2009). Although a typical MF 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  $m$  MF managers, of whom  $m_0$  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 MF manager separately at the significance level  $\gamma$ , one should expect to get  $\gamma m_0$  “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.

1  
2  
3  
4  
5  
6  
7  
8 Multiple inference is straightforward when the test statistics can be assumed  
9 independent or “weakly” dependent. This assumption is utilized in a recent series of  
10 working papers by Barras, Scaillet and Wermers (later referred to as BSW) that were  
11 made public online between 2005 and 2009. They evaluate the performance of 2076  
12 actively managed US equity MF over the period 1975-2006. BSW paper will be the main  
13 reference point for our study. In (Cuthbertson et al. 2008B) the method of BSW is  
14 borrowed to perform analysis of UK funds. An almost identical method is used for  
15 German data in (Otamendi et al. 2008). However, there are certain reservations about  
16 whether the independence assumption holds for the real MF data.

17 Accommodating dependent test statistics adds a layer of complexity. A typical way to  
18 handle this is to propose a parametric or non-parametric model for the dependence  
19 structure and incorporate it into the multiple inference procedure. In the context of MF  
20 study, this approach is not feasible because the amount of historical data is not sufficient  
21 to obtain a proper estimate of the dependence structure. In addition, the performance  
22 evaluation model used to obtain the test statistics can be misspecified, which contributes  
23 an unknown amount of bias to the inference.

24 Yet another poorly explored but important question is the statistical power of the  
25 performance evaluation model. In a typical MF study, no power diagnostics are provided.  
26 The study of (Kothari and Warner 2001) tries to shed some light on the issue but does not  
27 appear exhaustive, especially given that it is not based on the real MF data.

28 In this paper we contribute to the existing literature by using a multiple inference  
29 method that seems to offer a viable alternative that does not suffer from the  
30 abovementioned shortcomings. Recently, (Efron 2001-2008) developed an empirical  
31 Bayes approach that does not rely on the independence of test statistics or the direct  
32 estimation of their dependence structure. There is evidence that in some cases, the  
33 proposed method can take into account the misspecification of performance evaluation  
34 model as well. The method comes with comprehensive and insightful power analysis  
35 tools. In addition, it provides a rigorous and efficient way of looking into the performance  
36 of subgroups of MF, another issue of practical interest. The proposed approach was  
37 originally developed for gene microarray studies, and this is its first application in the  
38 realm of Finance.

39 We apply the new approach to about 2000 US equity MF observed in 1993-2007. We  
40 obtain compelling evidence that assuming independence of test statistics is inappropriate.  
41 It introduces a significant bias that results in overestimation of the number of both over-  
42 and underperforming MF. Our analysis shows that, although Efron’s approach offers  
43 higher precision and power, we are still unable to find a significant number of MF that  
44 are outperforming after fees and expenses. Finally, the power analysis shows that the  
45 study is very underpowered. The power is especially low when we try to reduce the  
46 history to only the most recent 3-5 years of data.

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

## 50 51 **2 Methodology**

### 52 53 **2.1 Data**

54 This study is focused on open-end, actively managed US equity MF. The monthly  
55 dataset is obtained from CRSP in 03/2008 and it spans 01/1993–06/2007 (14 ½ years). It  
56 is cleared of inappropriate types of funds, such as international, money market, index  
57  
58  
59  
60  
61  
62  
63  
64  
65

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 MF shareclass by its monthly TNA. For both datasets, the average number of observations per fund is about 129 (10 ¾ 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:

$$r_{i,t} = \alpha_i + b_i \cdot r_{m,t} + s_i \cdot r_{smb,t} + h_i \cdot r_{hml,t} + m_i \cdot r_{mom,t} + \varepsilon_{i,t} \quad (2.2.1)$$

$$t = 1, \dots, T$$

$$i = 1, \dots, m$$

where  $r_{i,t}$  is the excess return in time period  $t$  over the risk-free rate for the MF 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  $\varepsilon_i = (\varepsilon_{i,t}, \dots, \varepsilon_{i,T})$  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, the regression is estimated via a non-parametric bootstrap procedure similar to that in (Kosowski et al. 2006).

The parameter  $\alpha_i$  is measured in % per month and its value shows by how much the fund outperforms ( $\alpha_i > 0$ ) or underperforms ( $\alpha_i < 0$ ). We compute  $m$  one-sided p-values for the tests:

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

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

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

where  $\Phi^{-1}(\cdot)$  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  $\alpha_i$ ) 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 BSW. After 1992, there has to be a significant overlap between the BSW data and our sample.

### 2.3 False Discovery Rate and dependent test statistics

Suppose we perform the  $m$  separate tests (2.2.2), each with a significance level  $\gamma$ . Let  $Q$  be the number of rejected true null hypotheses (called “false discoveries”) divided by the number of all rejections. When talking about outperformers (underperformers), the term “false discoveries” refers to the funds that are not true outperformers (true underperformers). 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  $\gamma$ . For instance, to construct a portfolio of outperforming MF, it is meaningful to put a cap on the expected proportion of the funds that are not truly outperforming via the constraint  $FDR \leq q$ .

Denote by  $P^0$  a vector of  $m_0$  p-values that correspond to true null hypotheses in (2.2.2). 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

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

For  $\lambda$  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  $\lambda$ : 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:

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

This approach is behind the spline estimator of (Storey and Tibshirani 2003) and the bootstrap estimator of (Storey et al. 2004). The latter approach is used in BSW.

The main 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, MF study is an observational study where, typically, we may not simplify the dependence in this manner. Even if we are willing to assume that the joint distribution of test statistics is multivariate normal, the property that each null statistic is non-negatively correlated with the rest  $(m - 1)$  statistics appears too restrictive.

Another approach is to try to estimate the joint distribution of the test statistics directly. 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

1  
2  
3  
4  
5  
6  
7 and Wolf 2005; Romano et al. 2007, 2008). Their “StepM procedure” is also akin to the  
8 approach developed for biostatistical purposes by (van der Laan and Hubbard 2005).  
9 These methods assume that there are no constraints on the dependence structure.

10 Under the Carhart framework, the dependence structure of the test statistics is  
11 defined by  $\Sigma^0$ , the variance-covariance matrix of the vector  $\varepsilon_i = (\varepsilon_{i,1}, \dots, \varepsilon_{i,d})$ . One can  
12 reduce the number of estimated parameters in  $\Sigma^0$  by proposing a few “residual factors”  
13 that presumably account for all of the cross-sectional (across  $i$ ) dependence of  $\varepsilon_{i,d}$ 's. The  
14 residual factors can be “qualitative”: e.g., one may assume that error terms coming from  
15 MF with the same investment objective are correlated with the same correlation  
16 coefficient. It is also possible to derive the residual factors from the data using  
17 “dimension reduction” techniques, e.g. Principal Component Analysis (PCA).  
18

19 Suppose that, under Carhart’s framework, the matrix  $X$  of size  $m \times T$  contains the  
20 regression residuals. Then, the unconstrained estimate of  $\Sigma^0$  is  $\Sigma^1 = XX' / T$ . Next, the  
21 goal of PCA is to identify a relatively small number,  $p < m$ , of linear combinations of  
22 rows of  $X$ , and use these combinations to approximate  $\Sigma^0$ . The  $p$  most useful linear  
23 combinations correspond to the eigenvectors of  $\Sigma^1$  that have the  $p$  largest eigenvalues.  
24 These combinations form the  $p$  “residual factors” that define  $\Sigma^2$ , a constrained estimate of  
25  $\Sigma^0$ . (Jones and Shanken 2005) utilize a combination of “qualitative” and PCA-based  
26 residual factors.  
27

28 All these modeling techniques appear to have a fundamental problem: they only  
29 work when the utilized estimate, be it  $\Sigma^1$  or  $\Sigma^2$ , is a “good” estimate of  $\Sigma^0$ , which  
30 requires too much historical data. The results of (Yekutieli and Benjamini 1999; White  
31 2000) state that the control of FDR is attained only asymptotically, i.e., for a fixed  $m$  and  
32  $T \rightarrow \infty$ . This “size problem” is also investigated by Fan et al. (2008), who demonstrate  
33 the inadequacy of a variance-covariance matrix estimator when the data are insufficient.  
34 (Efron 2007C) refers to the work of van der Laan et al. to emphasize that the  
35 corresponding results are applicable only asymptotically and are not to be used unless  $T$   
36 is larger than  $m$ .  
37

38 In the context of MF studies, we have  $m$  about 2000 and  $T$  between 100 and 300,  
39 which amounts to a severe “size problem”. Note that while the rank of  $\Sigma^0$  may be  
40 anywhere between 1 and 2000, the rank of  $\Sigma^1$  is always under 300. That is, we know so  
41 little about  $\Sigma^0$  that we cannot even provide a reasonable estimate of its rank. The various  
42 dimension reduction techniques allow us to “reduce the dimension” of the available data  
43 (i.e., use the available data efficiently), but they do not solve the “size problem”.

44 At the same time, the “size problem” is often ignored in applications. (Yekutieli and  
45 Benjamini 1999) give a weather analysis example where  $m = 1977$  and  $T = 39$ , while  
46 using another, simulated dataset to show that FDR is controlled in which  $m = 40$  and  $T$  in  
47 between 200 and 1000. In the MF performance area, (Kosowski et al. 2006) use a non-  
48 restricted estimate of the dependence structure with  $m > 2000$  and  $T$  about 300. The study  
49 of (Cuthbertson et al. 2008A) borrows this approach and applies it to UK data with  $m =$   
50 900 and  $T$  about 340.

51 Another complication is that the equity market data show that  $\Sigma^0$  is very time-  
52 dependent. The correlation between two otherwise weakly correlated equity MF goes up  
53 during the so-called “flight-to-quality” periods. (Avellaneda and Lee 2008) illustrate this  
54 effect by considering a large number of US stocks observed daily between 10/2002 and  
55 02/2008. The return correlation matrix is computed based on a one-year rolling window.  
56 They perform PCA and estimate the number of principal components,  $p$ , that are  
57 necessary to explain 55% of the variance in the system. It turns out that during the “good  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4  
5  
6  
7  
8 times” of 2004-2006,  $p$  can be as large as 25, whereas during the “bad times” (2002,  
9 2007-2008)  $p$  is between 7 and 15. This suggests that, in a multifactor model with a fixed  
10 number of factors, the cross-sectional dependence structure of the residuals can change  
11 drastically over time, which aggravates the already serious “size problem”.

12 Even if we assume that  $\Sigma^0$  is time-independent and subject to restrictions, the  
13 number of parameters in the model is still large compared to the number of time points.  
14 For instance, if we introduce the residual factors of (Jones and Shanken 2005) or  
15 (Avellaneda and Lee 2008) we will end up having to estimate from 13 to 25 regression  
16 coefficients with an average number of observations in our sample equal to 129.  $\Sigma^2$  may  
17 be an unbiased estimator of  $\Sigma^0$ , but it will still have high variance.

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

25 (Storey et al. 2004) also show that under weak dependence FDR can be controlled  
26 for any fixed value of  $\hat{\lambda}$  in (2.3.2). The choice of optimal  $\hat{\lambda}$  is a bias-variance tradeoff  
27 problem which they solve via bootstrapping from the  $m$  p-values. Resampling from a set  
28 of (weakly) dependent p-values is a questionable technique for which no analytical  
29 results are available. Still, some numerical examples show that the bootstrap estimation  
30 of  $\hat{\lambda}$  is robust under “small group” type of weak dependence (Storey and Tibshirani  
31 2001). The application of FDR in BSW study rests on the assumption of weak  
32 dependence for the purpose of both FDR control and the estimation of the optimal  $\hat{\lambda}$  via  
33 bootstrap.  
34

35 This assumption may not be justifiable. As stated in BSW study itself, MF may  
36 exhibit correlated trading behaviors in large groups that can be caused, for instance, by  
37 being exposed to the same industrial sector or “herding” into particular stock(s). To  
38 address that, BSW argue that the funds’ test statistics are not very dependent because  
39 15% of the fund histories in their sample do not overlap in time, and on average only  
40 55% of return observations overlap. How much independence does the “lack of overlap”  
41 introduce? Compare this to an example of a weakly dependent structure with  $m=3000$  and  
42 the group size of 10 in (Storey et al. 2004). For a MF study with  $m=2000$ , where the  
43 degree of independence is associated with the absence of overlap, we obtain the  
44 following: the entire time period should be divided into subintervals with under 10 funds  
45 observed on each subinterval. Hence, it requires at least 200 subintervals. An average  
46 fund being observed for over 10 years, it implies the study’s time span has to be over  
47 2000 years. In reality, BSW data span only 32 years, which makes the “lack of overlap”  
48 argument doubtful for both BSW data and our sample. In addition, BSW (05/2008)  
49 estimate the residual correlation matrix of size  $898 \times 898$ . They claim that since the range  
50 of pairwise correlation terms is not “too far” from zero, it justifies the weak dependence  
51 assumption. In Section 2.4.2 we show that it is not always the case.  
52

53 Therefore, the weak dependence property inevitably implies some rather  
54 questionable and/or hard-to-check assumptions about the data. Explicit modeling of the  
55 high-dimensional correlation structure is not feasible either. Even very restrictive  
56 assumptions may not reduce the number of model parameters to the point where the  
57 amount of data is sufficient for estimation. The next section introduces a novel approach  
58 to large-scale simultaneous inference that can offer a viable alternative.  
59  
60  
61  
62  
63  
64  
65

## 2.4 Alternative approach: structural model and empirical null hypothesis

### 2.4.1 Structural model

In a number of papers published between 2001 and 2008, Efron proposed the following structural model that ties together  $\alpha$  and  $z$  values:

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

where  $g(\alpha)$  is an arbitrary distribution,  $\varphi(z|\mu, \sigma^2)$  denotes the density of a normal with mean  $\mu$  and variance  $\sigma^2$ , and “\*” denotes convolution. Our interest is in testing some hypothesis about  $\alpha$ , and the support of  $g(\alpha)$  can be arbitrarily split into two disjoint “null” and “alternative” sets:

$$\begin{aligned}g(\alpha) &= p_0 g_0(\alpha) + p_1 g_1(\alpha) \\ \text{where} \\ g_0(\alpha) &\text{-- "null" component} \\ g_1(\alpha) &\text{-- "alternative" component} \\ p_0 &= P_g\{\alpha \text{ is null}\} \\ p_1 &= P_g\{\alpha \text{ is alternative}\} \\ p_0 + p_1 &= 1\end{aligned}\tag{2.4.2}$$

In terms of corresponding  $z$ -values this will result in:

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

For instance, for the null  $H_0: \alpha = 0$  the “null” set consists of one point  $\{\alpha = 0\}$ , the “alternative” set is  $\{\alpha \neq 0\}$ , and  $p_0 = P_g\{\alpha = 0\} = m_0 / m$ . If all null  $p$ -values are i.i.d.  $U(0, 1)$ , the corresponding density of null  $z$ -values is  $\varphi(z|0, 1)$ .

The test will become a lot more powerful if we can calculate the  $p$ -value under the composite null:

$$H_i^0: \alpha_i \leq 0 \text{ vs. } H_i^a: \alpha_i > 0\tag{2.4.4}$$

Usually, the distribution of test statistics under the composite null is unknown, so the simple null is used instead. For this study, we can use the data itself to estimate the null, simple or composite (see next section). In terms of the structural model, we have

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

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

where

$$p_0 = P_g\{\alpha \leq 0\}, \quad p_1^+ = P_g\{\alpha > 0\}$$

$g_0(\alpha)$  – null density with support on  $\{\alpha \leq 0\}$

$g_1^+(\alpha)$  – alternative density with support on  $\{\alpha > 0\}$

and

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

where

$$f_0(z) = g_0 * \varphi(z | 0, \sigma_0^2) \quad \text{- density of null z's}$$

$$f_1^+(z) = g_1^+ * \varphi(z | 0, \sigma_0^2) \quad \text{- density of alternative z's}$$

$$p_0 + p_1^+ = 1$$

To distinguish between significant z-values that are positive and significant z-values that are negative, we can split the support of  $g(\alpha)$  into three subsets:

$$g(\alpha) = p_0 g_0(\alpha) + p_1^+ g_1^+(\alpha) + p_1^- g_1^-(\alpha) \tag{2.4.6}$$

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  $\{\alpha < 0\}$

$$f(z) = p_0 f_0(z) + p_1 f_1(z) \quad \text{- mixture density of z's}$$

$$p_1 f_1(z) = p_1^+ f_1^+(z) + p_1^- f_1^-(z)$$

where

$$f_0(z) = \varphi(z | 0, \sigma_0^2) \quad \text{- "zero" density of z's}$$

$$f_1^+(z) = g_1^+ * \varphi(z | 0, \sigma_0^2) \quad \text{- "positive" density of z's}$$

$$f_1^-(z) = g_1^- * \varphi(z | 0, \sigma_0^2) \quad \text{- "negative" density of z's}$$

$$p_1^+ + p_1^- = p_1, \quad p_0 + p_1 = 1$$

The Bayesian concept of "local false discovery rate" (fdr) can be interpreted as a "local" version of Benjamini and Hochberg's FDR. It is defined as follows:

$$fdr(z) = P\{\text{case } i \text{ is null} | z_i = z\} = \frac{p_0 f_0(z)}{f(z)} \tag{2.4.7}$$

Local fdr,  $fdr(z)$ , is the posterior probability that the test with corresponding z-score came from the null distribution. When we use (2.4.6) to identify outperformers, we actually have two "null components" and fdr has to be modified as follows:

$$fdr^+(z) = [p_0 f_0(z) + p_1 f_1^-(z)] / f(z) \tag{2.4.8}$$

For identification of underperformance,  $fdr^-(z)$  is defined in a similar manner. For the sake of simplicity, the superscripts will be omitted. Consider also

$$Fdr(z) = FdrLeft(z) = P\{\text{case } i \text{ is null} | z_i \leq z\} = \frac{p_0 F_0(z)}{F(z)} = E_f[fdr(t) | t \leq z] \tag{2.4.9}$$

$$FdrRight(z) = P\{\text{case } i \text{ is null} | z_i \geq z\} = E_f[fdr(t) | t \geq z]$$

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

1  
2  
3  
4  
5  
6  
7  
8 (2.4.7) nor (2.4.9) assume any particular dependence structure of  $z$ 's such as PRDS or the  
9 weak dependence.

#### 10 **2.4.2 Empirical null hypothesis**

11  
12 The local fdr approach is of the “empirical Bayes” kind: in (2.4.3) we do not pre-specify  
13 the mixture density  $f(z)$  and  $p_0$  because, unlike in the “classical Bayes” setting,  $f(z)$   
14 and  $p_0$  are estimated from the data. Under standard FDR approach from Section 2.3, the  
15 null density  $f_0(z)$  is pre-specified as  $\varphi(z|0,1)$  (called “theoretical null”). In certain cases  
16 it makes sense to estimate  $f_0(z)$  from the data also. (Efron 2003; 2004A; 2007C; 2007B)  
17 introduced the concept of “empirical null” where  $f_0(z)$  is approximated by  $\varphi(z|\delta_0, \sigma_0^2)$   
18 and the parameters  $\delta_0$  and  $\sigma_0^2$  are estimated.

19  
20 To understand why this may be necessary, consider the following example (Efron  
21 2007C). Suppose that all  $z$ 's are marginally  $N(0,1)$ , that is, all  $z$ 's are null. Each pair  $(z_i, z_j)$   
22 is bivariate normal with a distinct correlation coefficient  $\rho_{ij}$  drawn randomly from a  
23 certain normal distribution  $N(0, \tau^2)$ . Further, let  $A$  be a single independent realization  
24 (called “dispersion variate”) from  $N(0, \tau^2)$ . It can be shown that the ensemble of all  $z$ -  
25 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  
26 positive realizations of  $A$  produce  $\sigma_0^2 > 1$  (“overdispersion”) and, as a result, too many  
27 null cases will be declared significant (“over-rejection”), unless we use the empirical  
28 null  $f_0(z) = \varphi(z|\delta_0, \sigma_0^2)$ .

29  
30 Note that the marginal distribution of the null  $z$ -values can, indeed, be  $N(0, \sigma_0^2)$   
31 instead of  $N(0,1)$ . According to (Efron 2007B), it can happen when the model used to  
32 obtain the individual test statistics (in our case, it is (2.2.1)) is misspecified. Possible  
33 sources of misspecification are: unconsidered serial correlation or heteroskedasticity of  
34 error terms, application of asymptotically valid estimation when  $T$  is not large enough,  
35 and so on. In that case, the example above shows that even independent test statistics can  
36 behave as if dependent.

37  
38 In practice, both “genuine” dependence and the misspecification of marginal  
39 distribution are likely to be present. While one can try to ignore the former via justifying  
40 the independence / weak dependence assumption, the contribution of the latter is  
41 impossible to assess a priori: if one had known how the model was misspecified, one  
42 would have corrected the misspecification in the first place.

43  
44 (Efron 2007B) shows that, in the above example, not only the point estimate of  
45  $fdr(z)$  but also its estimated standard error,  $s.e.(fdr(z))$ , are conditioned on the ancillary  
46 statistic  $A$ , and, in that sense, are conditioned on the dependence structure of  $z$ 's.  
47 Likewise, the standard errors of  $\hat{p}_0, \hat{\delta}_0, \hat{\sigma}_0$  are also conditioned on the dependence  
48 structure. In this example, using the empirical null is essentially a way to adjust the  
49 inference for the dependence structure of  $z$ 's without having to model it explicitly. In  
50 addition, this takes into account the possible misspecification of the marginal distribution  
51 of null test statistics.

52  
53 Let us look into the weak dependence assumption in the context of the example  
54 above. Based on model (2.2.1), one could roughly estimate the density of  $\rho$  based on the  
55 empirical distribution of pairwise correlations in the residual correlation matrix. (BSW  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

05/2008) estimate the  $898 \times 898$  correlation matrix based on 898 funds observed for 60 months each. They find the estimated 25%, 50% and 75% quantiles for  $\rho$  are equal to  $-0.09; 0.05; 0.19$ , correspondingly. For the sake of argument, suppose that the three quantiles of true  $\rho$  are  $-0.09; 0.0; 0.09$  and  $\rho$  is normal, which implies  $\rho \sim N(0, \hat{\rho}^2 = 0.133^2)$ . To see how this can affect the inference, we introduce another version of (2.4.9):

$$\tilde{Fdr}(x|A) = P\{z_i \text{ null} \mid z_i \geq x, A\} \quad (2.4.10)$$

The case  $A = 0$  corresponds to the theoretical null. Suppose we are interested in detecting the outperforming MF, so set  $x = 2.5$ . Our calculations show that, if  $A$  takes on the value of  $0.16$  (just 1.2 standard deviations from the mean of zero), 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  $\tilde{Fdr}(2.5|0)$  is  $0.2$ , then  $\tilde{Fdr}(2.5|0.16)$  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. This illustrates that even such a seemingly close-to-zero range of  $\rho$  does not imply that we can treat the test statistics as independent.

The advantage of the 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  $T$ ) 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 BSW. Given that FDR is very closely related to Fdr (Efron 2002), we can treat our theoretical null-based procedure as a close match to BSW approach, with directly comparable results. While the theoretical null is always the first option to try, the abovementioned 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 the "zero assumption": it is assumed that only the null component is supported on a certain "zero interval"  $(z_-; z_+)$ . The parameters of interest are estimated with either the MLE or the method of moments. The interval  $U(\lambda, 1)$  from Section 2.3 corresponds to a symmetrical zero interval: e.g.,  $U(0.05; 1)$  corresponds to the zero interval  $(-1.96; 1.96)$ . The following formula shows the relation between  $\lambda$  from (2.3.1),  $z_-$  and  $z_+$ :

$$\lambda = \Phi(z_-) + (1 - \Phi(z_+)) \quad (2.4.11)$$

For the theoretical null and a fixed zero interval, the point estimate of  $m_0$  is the same in the BSW method (formula (2.3.2)) and Efron's approach. If the empirical null is chosen,  $f_0(z)$  can be approximated by a parametric distribution, such as normal  $N(\delta_0, \sigma_0^2)$  or

1  
2  
3  
4  
5  
6  
7 skewed split-normal  $SN(\delta_0, \sigma_1^2, \sigma_2^2)$ . An additional technical restriction,  $p_0 \geq 0.9$ , has to  
8 hold when we use the empirical null (Efron 2003).

9  
10 The choice of the zero interval itself is a bias-variance tradeoff problem where the  
11 value of  $\lambda$  or the boundaries of  $(z_-; z_+)$  are the smoothing parameters. BSW minimize  
12  $MSE(\hat{p}_0)$  using  $\lambda$  as a smoothing parameter. For a fixed  $\lambda$ ,  $MSE(\hat{p}_0)$  is calculated based  
13 on bootstrap technique (see Section 2.3) which we prefer to avoid for our study. Instead,  
14 we use a bias-variance tradeoff estimation method similar to that in (Turnbull 2007).  
15

#### 16 **2.4.4 Power statistics**

17  
18 A high power means that  $fdr(z)$  is small on the support of  $f_1(z)$ , which can be described  
19 by an overall (post hoc) power measure:

$$20 \quad E_{f_1} fdr = \int fdr(z) f_1(z) dz / \int f_1(z) dz = E_{f_1}[fdr(z)] \quad (2.4.12)$$

21  
22 It can be adapted to measure the power in the right tail:

$$23 \quad E_{f_1} fdr_{Right} = \int_0^{+\infty} fdr(z) f_1(z) dz / \int_0^{+\infty} f_1(z) dz = E_{f_1}[fdr(z) | z > 0] \quad (2.4.13)$$

24  
25 E $fdr_{Left}$  is defined similarly for the left tail ( $z < 0$ ).

26  
27 To put a cap on the proportion of false discoveries, (Efron 2007B) recommends  
28 picking z-values with  $fdr(z) \leq 0.2$ . We also adapt (a more liberal) rule: declare the fund  
29 under- (outperforming) as long as its FdrLeft (FdrRight) are under 0.2. It means that we  
30 shall tolerate up to 20% of false discoveries when we wish to construct an outperforming  
31 portfolio of MF. Similarly, we say that the study has decent power when E $fdr$  is under  
32 0.2.

33 An interesting question is whether one could improve the study by increasing the  
34 number of observations per fund,  $T$ . Assuming that the standard error of  $\alpha_i$  in (2.2.1) is  
35 proportional to  $1/\sqrt{T}$ , the package *locfdr* allows us to gauge how much power would be  
36 gained by increasing the value of  $T$ .  
37

#### 38 **2.4.5 Performance vs. investment objective**

39  
40 It would be interesting to look into the net performance versus investment objective.  
41 Statistically speaking, the findings of BSW suggest that one may be able to increase the  
42 power by using investment objective as a control factor.

43 BSW compare the fund categories by running their bootstrap-based procedure for  
44 each category separately. We can perform an fdr-based analysis which will not suffer  
45 from the possible misspecifications described in Section 2.4.2, insofar as we have the  
46 option of using the empirical null.

47 (Efron 2008B) proposes the following method. Suppose that all z-values are divided  
48 into two classes, A and B. In MF context, class A corresponds to the investment category  
49 of interest (e.g. Aggressive Growth), and class B corresponds to the rest of funds. Then  
50 the mixture density and fdr can be decomposed as follows:  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

$$f(z) = \pi_A \cdot f_A(z) + \pi_B \cdot f_B(z) \tag{2.4.14}$$

$\pi_A, \pi_B$  - a priori probabilities of class A and B

$$\begin{aligned} f_A(z) &= p_{A0}f_{A0}(z) + p_{A1}f_{A1}(z) && \text{- class A mixture density} \\ f_{A0}(z), f_{A1}(z) &&& \text{- null and alternative densities for class A} \\ fdr_A(z) &= p_{A0}f_{A0}(z) / f_A(z) && \text{- class A fdr} \end{aligned}$$

$$\begin{aligned} f_B(z) &= p_{B0}f_{B0}(z) + p_{B1}f_{B1}(z) && \text{- class B mixture density} \\ f_{B0}(z), f_{B1}(z) &&& \text{- null and alternative densities for class B} \\ fdr_B(z) &= p_{B0}f_{B0}(z) / f_B(z) && \text{- class B fdr} \end{aligned}$$

The main hypothesis of interest is:

$$H_0: fdr_A(z) = fdr(z) \tag{2.4.15}$$

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

$$f_{A0}(z) = f_{B0}(z) \tag{2.4.16}$$

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 (2.4.16), test the main hypothesis (2.4.15), and obtain an estimate of  $fdr_A(z)$ .

Details are available upon request.

### 3. Empirical Results

#### 3.1 Pre-expense returns, theoretical null

We estimate the structural model (2.4.6) and obtain the following results. Figure 3.1.1 shows the histogram of z-scores (y axis indicates the counts of z-scores in each of 90 bins), the estimate of mixture density,  $\hat{f}(z)$ , (green curve) and the estimated null component,  $\hat{p}_0 \cdot \varphi(z|0,1)$ , (blue dashed curve).

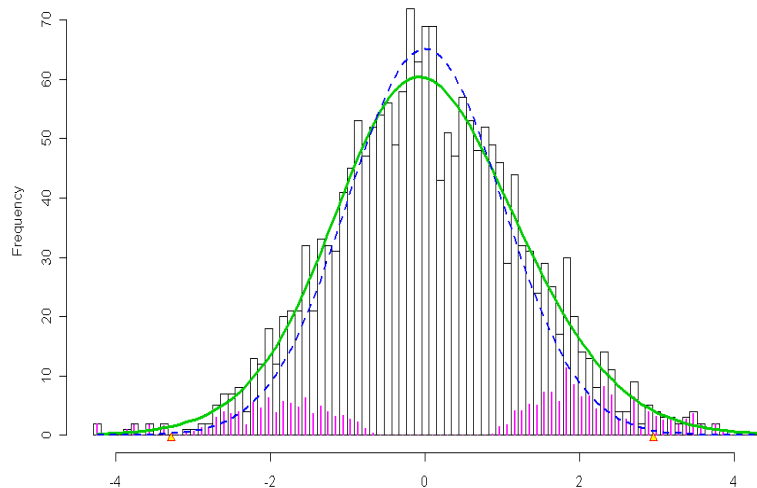


Figure 3.1.1 Estimated mixture density (green) and its null component (blue dashed) for pre-expense returns and theoretical null

The pink dashes are so-called “thinned counts” that are equal to observed z counts times the estimated alternative component,  $\hat{p}_i \hat{f}_i(z)$ .

Table 3.1.1 Estimation results for pre-expense returns and theoretical null

	$p_0$ , %	$p_1^+$ , %	$p_1^-$ , %
	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	$\lambda$	
	(-1.5; 1.5)	0.1336	

The corresponding BSW (05/2008) 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 3.1.1 have a lot of intersection with the corresponding intervals in BSW. The estimate of positive proportion drops from 9.6% to 6.3%, which is consistent with post-1992 deterioration of MF performance discovered in BSW. Still, the proportion of positive performers is both practically and statistically significant.

The results of Table 3.1.1 suggest that some 118 MF out of 1876 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 yellow triangles in Figure 3.1.1 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  $EfdrLeft = 0.64$ .

Table 3.1.2 Identified outperformers and false discoveries vs. 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	14.94	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

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 3.1.2). 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 MF 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  $10 \frac{3}{4}$  years of observations per fund; we can loosely think of this as having  $10 \frac{3}{4}$  years of data for each fund in the sample.

Table 3.1.3 Increase in power vs. years of available data for pre-expense returns and theoretical null

Sample size, Years	Efdr	EfdrRight	Outperformers identifiable with FdrRight=0.2
10 $\frac{3}{4}$	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

For instance, having 32 years of observations for each fund could help identify 90% of outperformers with FdrRight=0.2 (Table 3.1.3). The fact that it corresponds to EfdrRight = 0.2 confirms that using 0.2 as a rule of thumb for good power is reasonable.

In accordance to EfdrLeft = 0.64, the tables similar to Table 3.1.2 and 3.1.3 (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 3.1.1 is (87.95; 90.89), it is possible to assume that  $p_0 \geq 0.9$  in order to check whether the theoretical null is adequate for the data. 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  $\delta_0$  and  $\sigma_0$  should not be significantly different from 0 and 1, respectively.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

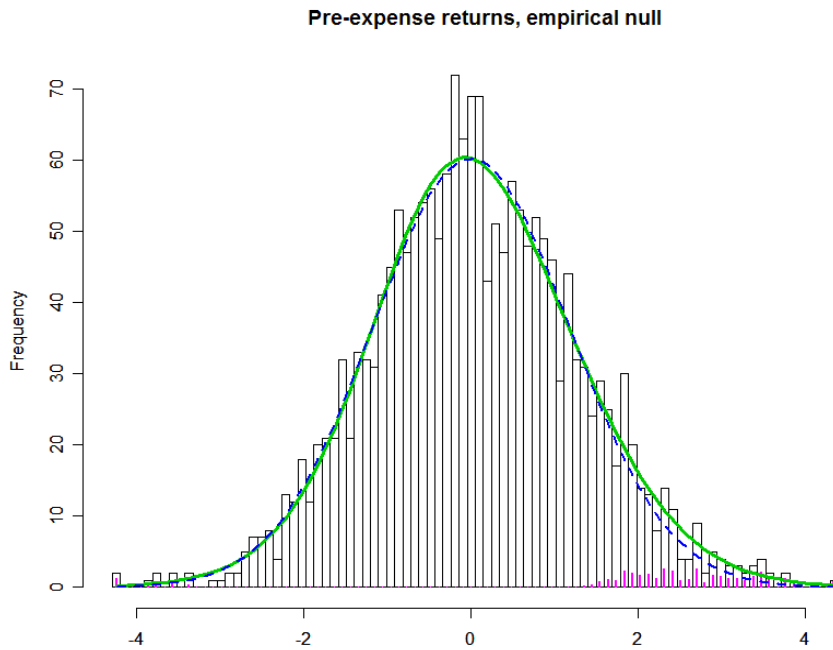


Figure 3.2.1 Estimated mixture density (green) and its null component (blue dashed) for pre-expense returns and empirical null

Table 3.2.1 Estimation results for pre-expense returns and empirical null

Zero interval	$\lambda$	$p_0, \%$	EfdrRight
(-1.7; 1.7)	0.0891	98.11 (0.99) (96.17; 100.05)	0.712
$\delta_0$	$\sigma_0$	t-value for $H_0: \sigma_0 = 1$	$A$
0.0039 (0.0353)	1.179 (0.034)	5.29	0.276

As we see from Table 3.2.1,  $\sigma_0$  is significantly greater than 1 with the corresponding t-value of 5.29. In other words, the z-values exhibit statistically significant overdispersion.

Comparing Figures 3.1.1 and 3.2.1, 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 blue dashed curve,  $\hat{p}_0 \hat{f}_0(z)$ , must always be under the green curve,  $\hat{f}(z)$ . This is clearly violated on Figure 3.1.1, 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 mentioned in Section 2.4.3 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

1  
2  
3  
4  
5  
6  
7 bias that is 34.5 times as small, an obviously more favorable bias-variance tradeoff for  
8 the empirical null.  
9

10 In terms of practical significance, one may think of such z-values as being  
11 marginally  $N(0, 1)$  and pairwise correlated with the correlation density  $\rho \sim N(0, \tau^2)$ , as in  
12 the example of Section 2.4.2. The estimate of the dispersion variate  $A$  in Table 3.2.1 is  
13 0.276, which is even greater than the preliminary guess of  $A = 0.16$  discussed in Section  
14 2.4.2. Returning to the example based on  $\tilde{F}dr(x|A)$  from (2.4.10), if we assume that  
15  $\tilde{F}dr(2.5|0) = 0.2$ , then  $\tilde{F}dr(2.5|0.276) = 0.47$ . It means that if 100 z's fall above 2.5, 80 of  
16 them are true discoveries if the theoretical null is used, but with the empirical null that  
17 number drops to 53.  
18

19 Therefore, we have both statistically and practically significant evidence against the  
20 theoretical null. The theoretical null-based inference overestimates the number of both  
21 skilled and unskilled funds in the population. The 95% confidence interval for  $p_0$   
22 changes from (87.95; 90.89) under the theoretical null to (96.17; 100.05) under the  
23 empirical null. The latter means that it is possible that both underperformers and  
24 outperformers are not present in the population at all. The estimated number of  
25 outperformers drops from 118 to 35 ( $p_1^+ = 1.85\%$ ) and the estimated number of  
26 underperformers drops from 80 to 1. Neither 35 nor 1 are significant statistically or  
27 practically.  
28

29 In addition, the power is quite poor: fdr is above 0.2 everywhere, and  $EfdrRight =$   
30 0.712. Even if we assume that the 35 outperformers are indeed present in the population,  
31 to select 50% of outperformers (about 17 out of 35), one has to tolerate  $FdrRight$  of 0.6  
32 by selecting about 26 false discoveries as well. The "Top 43" list of funds will have 26  
33 false entries,  $43 = 17 + 26$ . To obtain decent power ( $EfdrRight = 0.2$ ), it would take an  
34 unrealistic 43 years of data per fund.  
35

36 Since this study's sample has a significant overlap with that of BSW it is very likely  
37 that the overdispersion effect of similar magnitude is present in their sample. It means  
38 that BSW study overestimated the percentage of skilled and unskilled funds in the  
39 population just as well. Under the empirical null, the percentage of outperformers in  
40 BSW sample will probably be greater than 1.85%, but only because of better MF  
41 performance prior to 1993.  
42

### 43 3.3 Net performance, Theoretical Null

44 The corresponding BSW (05/2008) estimates of  $p_0, p_1^+, p_1^-$  are 75.4 (2.5), 0.6 (0.8), 24.0  
45 (2.3). Again, there is a good amount of intersection of corresponding confidence intervals  
46 for  $p_0, p_1^+, p_1^-$  (Table 3.3.1). It appears that the estimated number of outperformers (9  
47 funds out of 1911) is not statistically or practically significant.  $EfdrLeft = 0.35$  (still well  
48 above 0.2), and the power is not good. In particular, 54% of underperformers (295 out of  
49 547) are identified with  $FdrLeft = 0.2$  (Table 3.3.2). Despite the low power, a high  
50 proportion of underperformers makes it much easier to construct sizable "Bottom lists" of  
51 decent quality: e.g., the "Bottom 181" list has  $FdrLeft$  of 0.11 which corresponds to about  
52 20 false discoveries in the list.  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

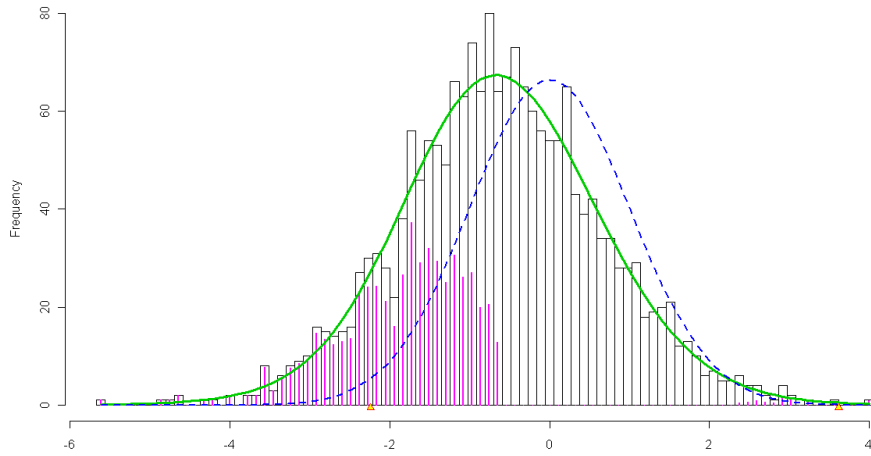


Figure 3.3.1 Estimated mixture density (green) and its null component (blue dashed) for net returns and theoretical null

Table 3.3.1 Estimation results for net returns and theoretical null

	$p_0, \%$	$p_1^+, \%$	$p_1^-, \%$
	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	1355	9	547
Zero interval	$\lambda$		
(-1.4; 1.4)	0.1615		
Efdr	EfdrRight	EfdrLeft	
0.35	0.49	0.35	

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

Table 3.3.2 Identified underperformers and false discoveries vs. FdrLeft for net returns and theoretical null

	Proportion of identified underperformers, %	Number of identified underperformers (rounded)	Number of false discoveries (rounded)
0.11	29.4	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

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 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 Section 3.2 because  $p_0$ , the proportion of funds with zero performance, is far below 0.9. However, we can do better: let us fit the more powerful 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.

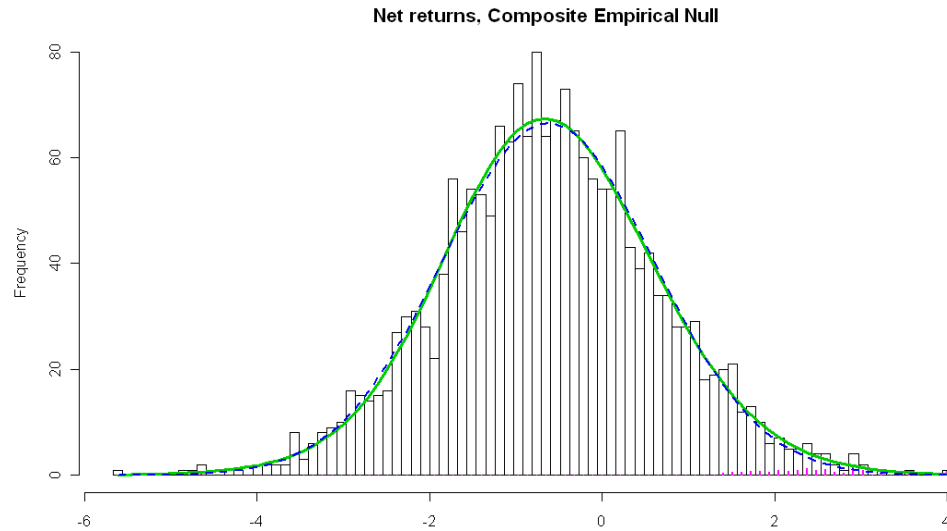


Figure 3.4.1 Estimated mixture density (green) and its null component (blue dashed) for net returns and composite empirical null.

Table 3.4.1 Estimation results for net returns and composite empirical null

	$p_0, \%$	$p_1^+, \%$
	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	$\lambda$	
(-4; 1.6)	0.055	
$\hat{\delta}_0$	$\eta_0$	EfdrRight
-0.624 (0.033)	1.229 (0.028)	0.725

Here we expect a much higher power to identify outperformers than in Section 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% to 0.7%.

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  $\sigma_0^2$  in (2.4.5) 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.

### 3.5 Net outperformance vs. Mutual Fund Investment Objective

Using the method outlined in Section 2.4.5, let us take a look at how the MF net outperformance depends on the investment objective category.

Table 3.5.1 Net outperformance vs. investment objective, 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%

Column 3 of Table 3.5.1 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  $fdr_G(z) = fdr(z)$ .

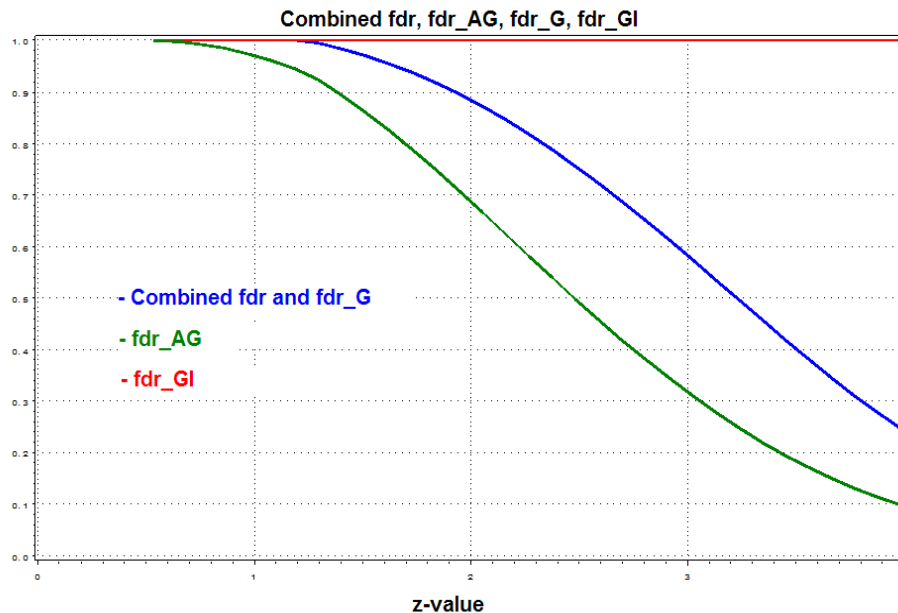


Figure 3.5.1 Combined fdr and Growth fdr (blue), Aggressive Growth fdr (green), Growth & Income fdr (red)

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

Figure 3.5.1 shows the curves corresponding to  $\hat{f}dr(z)$  (which coincides with  $\hat{f}dr_G(z)$ ),  $\hat{f}dr_{GI}(z)$ , and  $\hat{f}dr_{AG}(z)$ . Apparently, there are no skilled managers in GI group because for any  $z$   $\hat{f}dr_{GI}(z)=1$ . Using the estimate  $\hat{f}dr_{AG}(z)$  and  $\hat{f}dr_G(z)$ , 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 1911), it is about 3% in AG group, 0.79% in G group and 0% in GI group.

While  $\hat{f}dr(z)$  is always above 0.24,  $\hat{f}dr_{AG}(z)$  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 a quite aggressive level of 0.4, only 4 out of 19 AG outperformers are identified. Even a relatively superior AG group is unable to produce a practically significant number of identifiable outperformers.

The results of BSW for the same three groups (G, GI, AG) are difficult to interpret. In their 05/2007 version (based on 1464 funds, 1975-2002) they claim that GI funds have the lowest proportion of skilled managers (0%) and the AG funds are the best (8.0%). In BSW of 05/2008 (2076 funds, 1975-2006) they state that “results for the three investment-objective subgroups... are similar” but do not provide the numbers, and look into the “short-term performance” (see Section 3.6) to find that AG is the best (4% of outperformers) and GI is the worst (0%).

### 3.6 Short-term Net Performance

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

However, the short-term performance may be better, as suggested by BSW. They partition the data into six non-overlapping subperiods of 5 years each, from 1977-1981 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 2076 to 3311 and  $\hat{p}_i^+$  goes up from 0.6 (0.8)% to a statistically significant 2.4 (0.7)%, correspondingly. In BSW this is interpreted as evidence for superior “short-term” performance that exists for a while and gradually disappears in the long-run equilibrium. BSW 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. In the end, the “short-term”  $z$ -values will probably be more overdispersed than the original  $z$ -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 BSW, 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 3636 of such “funds”. Applying the theoretical null results in  $p_i^+ = 0.81$  (0.52)% (29 outperformers), both statistically and practically insignificant.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

Table 3.6.1 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	$\lambda$	
(-3.5; 1.6)	0.055	
$\delta_0$	$\eta_0$	EfdrRight
-0.467 (0.026)	1.254 (0.024)	0.877

Applying instead the composite empirical null, as in Section 3.4, allows us to 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 actually drops from 29 to 13 funds (Table 3.6.1). 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 Section 2.4.2, the results obtained from our approach under the theoretical null are directly comparable to the output of BSW method. It is reassuring that despite the difference in the employed datasets, when we use the theoretical null (Sections 3.1, 3.3, 3.6), our findings are consistent with the BSW results.

The switch to the empirical null is well grounded. First of all, in Section 2.4.2 we find that the BSW data itself do not support the weak dependence assumption. Based on our own data, the findings in Section 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 Section 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 BSW. For instance, for pre-expense returns BSW method producers  $s.e.(\hat{p}_0)=2.7\%$ , whereas our result in Section 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 the more powerful composite null  $H_i^0 : \alpha_i \leq 0$  as opposed to the simple null  $H_i^0 : \alpha_i = 0$  used in BSW and most MF studies. When we estimate the composite empirical null from the data, overdispersion is taken into account also. Since even under that powerful setting the

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

number of outperformers proves neither statistically nor practically significant (Section 3.4), the evidence for the absence of net outperformance in MF industry in 1993-2007 is substantially reinforced.

The investment objective analysis in Section 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. Also, BSW do not directly test for the difference between the investment categories, while our test (2.4.15) does exactly that. Qualitatively, our findings are consistent with BSW results: AG funds are the best and GI funds are the worst. 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 1911).

The results for “short-term” net performance in Section 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 BSW must have come from the bias of the theoretical null.

If we are interested in practical applications of MF performance evaluation, the study has to have a high power. 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. Unfortunately, advice on how not to invest may be less appreciated.

A positive finding is that Section 3.3 shows that about 71% of funds have zero net performance under the theoretical null. When overdispersion is taken into account, that number can only go up. It is safe to say that over 70% of funds in the sample have net return alphas that are not distinguishable from zero. That proportion will probably remain large even after some unconsidered fees (such as loads) are taken into account. Zero alpha funds are of value because they essentially provide a free (on average) access to the US equity market. For a risk-neutral investor, zero-alpha funds are superior to index funds whose net alphas are negative, although close to zero. To estimate the total gain, one may use the study of (Elton et al. 2004) who look into fifty-two S&P500 index funds over 1996-2001 and find that their average alpha is minus 0.41% p.a.

Extending the sample back (e.g., BSW 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  $14 \frac{1}{2}$  years with the mean of  $10 \frac{3}{4}$  years per fund. Although 10% of the funds span the entire  $14 \frac{1}{2}$  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 Section 3.3), and the power is still poor. Therefore, an unsatisfactory power is inherent to both the current and BSW study despite a much larger time span of the latter.

1  
2  
3  
4  
5  
6  
7  
8 To make matters worse, a long-lived MF is likely to be managed by a few  
9 successive portfolio managers. According to John Bogle, founder of The Vanguard  
10 Group, "...the tenure of the average portfolio manager is just five years...". (Kothari and  
11 Warner 2001) also indicate that the investor is likely to consider only from 3 to 5 years of  
12 MF history. So, practically speaking, there are reservations as to whether the 6-15 year-  
13 old data are relevant for investors. At the same time, reducing the history is bound to  
14 reduce the power to a level where the study is absolutely uninformative. For instance, in  
15 Section 3.4, with with  $T = 10 \frac{3}{4}$  years,  $EfdrRight = 0.725$  (very poor). After  $T$  is reduced  
16 to about 5 years (Section 3.6), the power gets even worse:  $EfdrRight = 0.877$ . It appears  
17 that any MF study that is based on monthly data and a similar multifactor performance  
18 evaluation model is bound to be very underpowered, especially if the purpose is future  
19 investment. A possible way out is to use holdings-based performance measures.  
20 According to (Kothari and Warner 2001) such measures are more powerful. However, the  
21 holdings data are not always available.

22 We finish this discussion by noting that any multiple inference procedure works  
23 with "input list" of test statistics and the "quality" of this list is at least as important as an  
24 appropriate multiple inference method. In particular, in Section 2.4.2 it is suggested that  
25 the empirical null-based procedure "takes into account" the asset pricing model  
26 misspecification. Suppose, for simplicity, that all  $z$ 's are independent and the only source  
27 of overdispersion ( $\sigma_0 > 1$ ) is misspecification of model (2.2.1), e.g. caused by a value of  $T$   
28 that is too small. Essentially,  $\sigma_0 > 1$  tells us that there is some extra noise in  $z$ -values  
29 which we have to take into account by using  $f_0(z) = \varphi(z | \delta_0, \sigma_0^2)$  instead of  $f_0(z) = \varphi(z | 0, 1)$ .  
30 Taking that into account will prevent us from making false discoveries, but it will not  
31 make the extra noise disappear. If the level of noise is very high, the procedure will  
32 simply declare that all or almost all cases are null. Therefore, one cannot just rely on a  
33 multiple inference method to improve the identification of outperformers. In particular,  
34 (Mamaysky et al. 2007) argue that it is unlikely for a single performance evaluation  
35 model to be equally good for each fund in the sample. They show that using a few  
36 competing models, combined with back-testing, can significantly improve the  
37 performance of portfolios of MF. Using such an approach coupled with a multiple  
38 inference procedure can be an interesting topic for future research.  
39  
40

#### 41 **4 Conclusion**

42  
43 When evaluating the performance of a large number of MF simultaneously, one has  
44 to weed out false discoveries. This task is fairly straightforward when the performance  
45 test statistics are independent across funds. However, independence is unlikely to hold for  
46 real data. On the other hand, there are not enough years of data to estimate the  
47 dependence structure of test statistics directly. In addition, a misspecified performance  
48 evaluation model can bias the results. Is there a way around these problems? The state-of-  
49 the-art approach of Efron offers a viable alternative. It also helps us investigate the  
50 usually neglected issue of statistical power in a MF study.

51 In this paper, we analyze the performance of about 2000 US equity MF over a  
52 period of 14  $\frac{1}{2}$  years. In contrast to existing studies, we neither assume independence of  
53 test statistics across MF, nor do we try to estimate the dependence structure based on the  
54 data that are clearly insufficient for that purpose. In addition, certain features of Efron's  
55 approach make it more powerful and precise, as well as being able to perform a rigorous  
56 and efficient analysis of subgroups of MF.  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4  
5  
6  
7  
8 Our analysis suggests that it is not appropriate to treat the test statistics as mutually  
9 independent with pre-specified null distribution. The data indicate that doing so leads to  
10 both statistically and practically significant bias, when the proportions of both under- and  
11 outperformers are overestimated. Despite the advantages of Efron's approach, we fail to  
12 identify a practically or statistically significant proportion of net outperformers. The  
13 power analysis shows that, due to the nature of data and the performance evaluation  
14 model (monthly dataset, a multifactor model), the study has a very low power. That is,  
15 we are hardly able to single out the true out- or underperformers. It would require an  
16 unrealistically large number of years of data to increase the power to a decent level.

## 17 18 **References**

- 19 Ammann, M., Verhofen, M.: The impact of prior performance on the risk-taking of mutual fund managers. *Ann*  
20 *Financ* 5:69-90 (2009)
- 21 Avellaneda, M., Lee, J.: Statistical Arbitrage in the U.S. Equities Market. SSRN.  
22 <http://ssrn.com/abstract=1153505> (2008). Accessed 11 July 2008
- 23 Barras, L., Scaillet, O., Wermers, R.: False Discoveries in Mutual Fund Performance: Measuring Luck in  
24 Estimated Alphas. To appear in *J Financ*
- 25 Benjamini, Y., Hochberg, Y.: Controlling the False Discovery Rate: A Practical and Powerful Approach to  
26 Multiple Testing. *J Roy Stat Soc Vol* 57, No.1, pp. 289-300 (1995)
- 27 Benjamini, Y., Hochberg, Y.: On the Adaptive Control of the False Discovery Rate in Multiple Testing with  
28 Independent Statistics. *J Educ Behav Stat Vol* 25, No. 1, 60-83 (2000).
- 29 Benjamini, Y., Yekutieli, D.: The Control of the False Discovery Rate in Multiple Testing under Dependency.  
30 *Ann Stat Vol* 29, Number 4, 1165-1188 (2001)
- 31 Benjamini, Y., Krieger, A., Yekutieli, D.: Adaptive linear step-up procedures that control the false discovery  
32 rate. *Biometrika* 93(3), 491-507 (2006)
- 33 Carhart, M.: On persistence of mutual fund performance. *J Financ Vol* 52, No. 1, (1997)
- 34 Chen, H., Jegadeesh, N., Wermers, R.: The value of active mutual fund management: An examination of the  
35 stockholdings and trades of fund managers. *J Financ Quant Anal* 35, 343-368 (2000)
- 36 Cuthbertson, K., Nitzsche, D., O'Sullivan, N.: UK mutual fund performance: Skill or luck? *J Empir Financ* 15,  
37 613-634 (2008A)
- 38 Cuthbertson, K., Nitzsche, D., O'Sullivan, N.: False discoveries: winners and losers in mutual fund  
39 performance. SSRN. [http://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=1093624](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1093624) (2008B).  
40 Accessed 20 Feb 2008
- 41 Daniel, K., Grinblatt, M., Titman, S., Wermers, R.: Measuring mutual fund performance with characteristic-  
42 based benchmarks. *J Financ Vol* 52, Issue 3 (1997)
- 43 Efron, B.: Empirical Bayes Analysis of a Microarray Experiment. *J Am Stat Assoc Vol* 96, No 456, 1151-1160  
44 (2001)
- 45 Efron, B., Tibshirani, R.: Empirical Bayes Methods and False Discovery Rates for Microarrays. *Genet*  
46 *Epidemiol* 23:70-86 (2002)
- 47 Efron, B.: Large-Scale Simultaneous Hypothesis Testing: the choice of a null hypothesis. *J Am Stat Assoc*  
48 99(465): 96-104 (2004A)
- 49 Efron, B.: Selection and Estimation for Large-Scale Simultaneous Inference.  
50 <http://www-stat.stanford.edu/~ckirby/brad/papers/> (2004B). Accessed 25 Jan 2009
- 51 Efron, B.: Local False Discovery Rates.  
52 <http://www-stat.stanford.edu/~ckirby/brad/papers/> (2005). Accessed 25 Jan 2009
- 53 Efron, B.: Microarrays, Empirical Bayes, and the Two-Groups Model. *Stat Sci Vol* 23, Number 1, 1-22.  
54 (2008A)
- 55 Efron, B.: Testing the significance of sets of genes. *Ann Appl Stat Vol* 1, Number 1, 107-129 (2007A)
- 56 Efron, B.: Size, Power, and False Discovery Rates. *Ann Stat Vol* 35, No. 4, 1351-1377 (2007B)
- 57 Efron, B.: Correlation and Large-Scale Simultaneous Significance Testing. *J Am Stat Assoc* 102(477): 93-103  
58 (2007C)
- 59 Efron, B.: Simultaneous inference: when should hypothesis testing problems be combined? *Ann Appl Stat Vol*  
60 2, Number 1, 197-223 (2008B)
- 61 Elton, E., Gruber, M., Busse, J.: Are Investors Rational? Choices Among Index Funds. *J Financ* 59, 261-288  
62 (2004)
- 63 Fan, J., Fan, Y., Lv, J.: High dimensional covariance matrix estimation using a factor model. *J Econometrics*  
64 147, 186-197 (2008)
- 65 Jensen, M.: The Performance of Mutual Funds in the Period 1945-1964. *J Financ Vol* 23, No. 2 (1968)

- 1  
2  
3  
4  
5  
6  
7 Jones, C., Shanken, J.: Mutual fund performance with learning across funds. *J Financ Econ* 78, 507-552 (2005)  
8 Kosowski, R., Timmermann, A., Wermers, R., White, H.: Can Mutual Fund “Stars” Really Pick Stocks? New  
9 Evidence from a Bootstrap Analysis. *J Financ* Vol 61 (2006)  
10 Kothari, S., Warner, J.: Evaluating Mutual Fund Performance. *J Financ* Vol 56, No. 5, 1985-2010 (2001)  
11 Mamaysky, H., Spiegel, M., Zhang, H.: Improved Forecasting of Mutual Fund Alphas and Betas. *Rev Financ*  
12 11 (2007)  
13 Otamendi, J., Doncel, L., Grau, P., Sainz, J.: An evaluation on the true statistical relevance of Jensen's alpha  
14 through simulation: An application for Germany. *Econ Bull* Vol 7, No. 10 pp. 1-9 (2008)  
15 Romano, J., Wolf, M.: Stepwise Multiple Testing as Formalized Data Snooping. *Econometrica* 73, 1237-1282  
16 (2005)  
17 Romano, J., Shaikh, A., Wolf, M.: Control of the False Discovery Rate Under Dependence Using the Bootstrap  
18 and Subsampling. University of Zurich Working Paper No. 337.  
19 <http://ssrn.com/abstract=1025410> (2007). Accessed 25 Jan 2009  
20 Romano, J., Shaikh, A., Wolf, M.: Formalized Data Snooping Based on Generalized Error Rates. *Economet*  
21 *Theor* 24 (2008)  
22 Storey, D., Tibshirani, R., 2001. Estimating False Discovery Rates Under Dependence, with Applications to  
23 DNA Microarrays. <http://stat.stanford.edu/reports/papers2001.html> (2001). Accessed 25 Jan 2009  
24 Storey, J.: A Direct Approach to False Discovery Rates. *J Roy Stat Soc B* 64, 479-498 (2002)  
25 Storey, D., Tibshirani, R.: Statistical Significance for Genomewide Studies. *Proc Nat Acad Sci USA* Vol 100, p.  
26 9440 (2003)  
27 Storey, D.: The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann Stat* Vol 31,  
28 Number 6 (2003)  
29 Storey, J., Taylor, J., Siegmund, D.: Strong control, conservative point estimation and simultaneous  
30 conservative consistency of false discovery rates: a unified approach. *J Roy Stat Soc* Vol 66, p. 187  
31 (2004)  
32 Turnbull, B.: Optimal Estimation of False Discovery Rates.  
33 <http://www.stanford.edu/~bkatzen/> (2007). Accessed 25 Jan 2009  
34 van der Laan, M., Hubbard, A.: Quantile function based null distribution in resampling based multiple testing.  
35 *Stat Appl Genet Mol Biol* 5, article 14 (2005)  
36 White, H.: A Reality Check for Data Snooping. *Econometrica* 68, 1097-1126 (2000)  
37 Yekutieli, D., Benjamini, Y.: Resampling-based false discovery rate controlling multiple test procedures for  
38 correlated test statistics. *J Stat Plan Infer* Vol 82, p. 171 (1999)  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65