Alpha-investing: A new multiple hypothesis testing procedure that controls mFDR

Dean P. Foster and Robert A. Stine[*]

Department of Statistics

The Wharton School of the University of Pennsylvania

Philadelphia, PA 19104-6340

October 5, 2006

Abstract

We propose alpha-investing for testing multiple hypotheses. Alpha-investing is an adaptive, sequential methodology that encompasses a large family of procedures. All control mFDR, which is the ratio of the expected number of false rejections to the expected number of rejections. mFDR is a weaker criterion than FDR, which is the expected value of the ratio. We compensate for this weakness by showing that alpha-investing controls mFDR at every rejected hypothesis. Alpha-investing resembles alpha-spending used in sequential trials, but possess a key difference. When a test rejects a null hypothesis, alpha-investing earns additional probability toward subsequent tests. Alpha-investing hence allows one to incorporate domain knowledge into the testing procedure and improve the power of the tests. In this way, alpha-investing enables the statistician to design a testing procedure for a specific problem while guaranteeing control of mFDR.

Key words and phrases: Bonferroni method, false discovery rate (FDR, mFDR), family-wise error rate (FWER), multiple comparison procedure.

[*] All correspondence regarding this manuscript should be directed to Prof. Stine at the address shown with the title. He can be reached via e-mail at stine@wharton.upenn.edu.

1 Introduction

We propose an adaptive, sequential methodology for testing multiple hypotheses. Our approach, called alpha-investing, works in the usual setting in which one has a batch of several hypotheses as well as in cases in which the hypotheses arrive sequentially in a stream. Streams of hypotheses arise naturally in variety of contemporary modeling applications, such as genomics and variable selection for large models. In contrast to the comparatively small problems that spawned multiple comparison procedures, modern applications can involve thousands of tests. For example, micro-arrays lead one to compare a control group to a treatment group using measured differences on over 6,000 genes (Dudoit, Shaffer and Boldrick, 2003). If one considers the possibility for interactions, then the number of tests is virtually infinite. In contrast, the example used by Tukey to motivate multiple comparisons compares the means of only 6 groups (Tukey, 1953, available in Braun (1994)). Because alpha-investing allows the testing to proceed sequentially, the choice of future hypotheses can depend upon the results of previous tests. Thus, having discovered differences in certain genes, an investigator could, for example, direct attention toward genes that share common transcription factor binding sites (Gupta and Ibrahim, 2005).

Before we describe alpha-investing, we introduce a slight variation on the marginal false discovery rate (mFDR), a existing criterion for multiple testing. mFDR is defined as follows. Let the observable random variable R denote the total number of hypotheses rejected by a testing procedure, and let V denote the unobserved number of falsely rejected hypotheses. Ideally V is small. Then controlling a version of mFDR, which we denote mFDR_1, at level α means

$$\mathrm{mFDR}_1 \equiv \frac{E(V)}{E(R)+1} \leq \alpha. \tag{1}$$

mFDR traditionally does not add 1 to the denominator; following our notation, we denote the traditional version mFDR_0. The notation mFDR_1 will remind purists that this isn't exactly the usual definition. The addition of a positive constant to the denominator avoids statistical problems under the complete null hypothesis. Under the complete null hypothesis, all hypotheses are false. In this case, $V \equiv R$ and mFDR_0

$= 1$ for all procedures. We contrast mFDR_0 with $\text{FDR} \equiv E(V/R)$ after we introduce alpha-investing.

An alpha-investing rule is an adaptive testing procedure that resembles an alpha-spending rule. An alpha-spending rule begins with an initial allowance for Type I error, what we call the initial alpha-wealth of the procedure. Each test at level α_i reduces the alpha-wealth of the spending rule by α_i. Once the alpha-wealth of the spending rule reaches zero, no further tests are allowed. Alpha-spending naturally implements a Bonferroni rule. By the union rule of probability, the total chance of making a Type I error is bounded by the initial alpha-wealth. In multiple testing, however, Bonferroni rules are too conservative. We invented alpha-investing rules to overcome this conservatism. In what follows, we show that all alpha-investing rules control mFDR_1.

We escape the conservatism of Bonferroni rules in the following way. Each time an alpha-investing rule rejects a null hypothesis, it earns a contribution to its alpha-wealth. Thus rejections beget more rejections. This idea of alpha-wealth begetting more alpha-wealth leads us to call this class of procedures alpha-investing rules. Alpha-investing rules allow one to test a possibly infinite stream of hypotheses, accommodate dependent tests, and incorporate domain knowledge. Like alpha-spending rules alpha-investing rules free the statistician to allocate the available Type I error among the various hypotheses, all the while protecting from excessive false rejections.

The sequential nature of alpha-investing allows us to enhance the type of control obtained through mFDR. Though mFDR is successful in identifying good multiple testing procedures, it is often compared unfavorably to the more famous FDR. Both FDR and mFDR_0 were suggested in Benjamini and Hochberg (1995) along with mFDR_1, which they considered somewhat artificial. After all, mFDR does not control a property of the realized sequence of tests, instead controlling a ratio of expectations over the ensemble of test sequences. In contrast, FDR controls the expectation of V/R for a given sequence. Nonetheless, FDR does not produce the type of control we require. Ideally we would like to be guaranteed that at the point we reject k hypotheses, at most some fraction of these, say $0.05\,k$ have been incorrectly rejected on average. Neither

FDR, nor mFDR, permit such claims.

By placing mFDR in a sequential setting, we can require that a procedure do well if stopped early. Suppose rather than testing all m hypotheses, the statistician stops after rejecting 10. We would like to be able to assure her that no more than, say, 2 of these were false rejections, on average. This further control distinguishes what we describe as "uniform control" of mFDR.

Suppose we test hypotheses until the total number of rejections R reaches some target r. Let T_r identify the index of this test. Define $V(T_r)$ to be the number of nulls that have been incorrectly rejected at this point. A test procedure uniformly controls mFDR_1 if this stopped process controls mFDR_1 in the sense of (1). In words, equation (1) shows that the expected value of V_R given that $R = r$ is less than or equal to $\alpha(r + 1)$. This conditional expectation requires that we introduce a stopping time which stops the testing procedure when $R = r$. We leave these details for Section 5. As a preview of our results, we give the following theorem:

Theorem 1 *An alpha-investing rule with control parameters set to α has the property that $E\ V(T_r) \leq \alpha(r + 1)$ where T_r is the stopping time defined by occurrence of the r^{th} rejection, and $V(m)$ is the number of false rejections when m hypothesis have been tested.*

In fact, when a sequence of tests is stopped at a fixed number of rejections, many of the variations on FDR are equivalent (see Theorem 2.)

The rest of this paper develops as follows. We first review several ideas from the literature on multiple comparisons, particularly those related to the family-wise error rate and FDR. Next we discuss alpha-investing rules in Section 3. In Sections 4 and 5 we discuss uniform control of mFDR. In Section 6, we show that alpha-investing rules uniformly control mFDR. We describe the design of alpha-investing rules and give several examples in Section 7. We close in Section 8 with a brief summary. We generally defer proofs to the appendix.

2 Criteria and Procedures

To set the stage for describing alpha-investing and our modification of mFDR, we review the two criteria most commonly applied in testing multiple hypotheses: the family-wise error rate and the false discovery rate. These criteria generalize the notion of the Type I error rate (α-level) to tests of several hypotheses. Just as there are many α-level tests of a simple hypothesis, so too are there various procedures for testing multiple hypotheses. We confine our review to two: the Bonferroni rule (as implemented using alpha-spending) and step-down tests. These two are closely related to the alpha-investing rules described in Section 3.

Suppose that we have a set of m null hypotheses $\mathcal{H}(m) = \{H_1, H_2, \ldots, H_m\}$ that specify values for parameters $\theta = \{\theta_1, \theta_2, \ldots, \theta_m\}$. Each parameter θ_j can be scalar or vector-valued, and Θ denotes the space of parameter values. In the most familiar case, each null hypothesis specifies that a scalar parameter is zero, $H_j : \theta_j = 0$. The complete null hypothesis implies that all m null hypotheses hold.

We follow the standard notation for labeling the true and false rejections (Benjamini and Hochberg, 1995). Assume that m_0 of the null hypotheses in $\mathcal{H}(m)$ are true. The *observable* statistic $R(m)$ counts how many of these m hypotheses are rejected. The *unobservable* random variable $V^\theta(m)$ denotes the number of false positives among the m tests, those cases in which the testing procedure incorrectly rejects a true null hypothesis. Similarly, $S^\theta(m) = R(m) - V^\theta(m)$ counts the number of correctly rejected null hypotheses. We index these random variables with a superscript θ to distinguish them from a statistic such as $R(m)$; $V^\theta(m)$ and $S^\theta(m)$ are not observable without θ. For the complete null hypothesis, $m_0 = m$, $V^\theta(m) \equiv R(m)$ and $S^\theta(m) \equiv 0$.

The original intent of multiple testing was to control the chance for *any* false rejection. The *family-wise error rate* (FWER) is the probability of falsely rejecting *any* null hypothesis from $\mathcal{H}(m)$, regardless of the underlying parameters,

$$\mathrm{FWER}(m) \equiv \sup_{\theta \in \Theta} \mathrm{P}_\theta(V^\theta(m) \geq 1) \ . \tag{2}$$

An important special case is control of FWER under the complete null hypothesis:

$$\mathrm{P}_0(V^\theta(m) \geq 1) \leq \alpha \ , \tag{3}$$

where P_0 denotes the probability measure under the complete null hypothesis. We refer to this goal as controlling FWER in the weak sense. All of the procedures that we describe control FWER in the weak sense, but not all control FWER more generally.

Bonferroni procedures are familiar and represent an important benchmark for comparison. Let $p_1, \ldots, p_m$ denote the p-values of tests of $H_1, \ldots, H_m$. Given a chosen level $0 < \alpha < 1$, the usual Bonferroni procedure rejects those H_j for which $p_j \leq \alpha/m$. Let the indicators $V_j^\theta \in \{0,1\}$ track incorrect rejections; $V_j^\theta = 1$ if H_j is incorrectly rejected and is zero otherwise. Then $V^\theta(m) = \sum V_j^\theta$ and the inequality

$$P_\theta(V^\theta(m) \geq 1) \leq \sum_{j=1}^{m} P_\theta(V_j^\theta = 1) \leq \alpha \tag{4}$$

shows that this Bonferroni procedure controls $\text{FWER}(m) \leq \alpha$. One need not distribute α equally over $\mathcal{H}(m)$; the inequality (4) requires only that the sum of the α-levels not exceed α. This observation suggests the alpha-spending characterization of the Bonferroni procedure. As an alpha-spending rule, the Bonferroni procedure allocates α over a collection of hypotheses, devoting a larger share to hypotheses of greater interest. In effect, the procedure has a budget of α to spend. It can spend $\alpha_j \geq 0$ on testing each hypothesis H_j so long as $\sum_j \alpha_j \leq \alpha$. Although such alpha-spending rules control FWER, they are often criticized for having little power. Clearly, the power of the traditional Bonferroni procedure decreases as m increases because the threshold α/m for detecting a significant effect decreases. The testing procedure introduced in Holm (1979) offers more power while controlling FWER, but the improvements are small.

To obtain substantially more power, Benjamini and Hochberg (1995) (BH) introduces a different criterion, the false discovery rate (FDR). Benjamini and Hochberg define FDR as the expected proportion of false positives among rejected hypotheses,

$$\text{FDR}(m) = E_\theta\left(\frac{V^\theta(m)}{R(m)} \mid R(m) > 0\right) P(R(m) > 0)\ . \tag{5}$$

For the complete null hypothesis, $R(m) \equiv V^\theta(m)$ and $\text{FDR}(m) = P_0(R(m) > 0)$, which is just $\text{FWER}(m)$. Thus, test procedures that control $\text{FDR}(m) \leq \alpha$ control the $\text{FWER}(m)$ in the weak sense at level α. If the complete null hypothesis is rejected, FDR introduces a different type of control. Under the alternative, FDR(m) decreases as the number of false null hypotheses $m - m_0$ increases (Dudoit et al., 2003). As a result, FDR(m) becomes more easy to control in the presence of non-zero effects, allowing more powerful procedures. Variations on FDR include pFDR (which drops the term $\mathrm{P}(R > 0)$ Storey, 2002, 2003) and the local false discovery rate fdr(z) (which estimates the false discovery rate as a function of the size of the test statistic Efron, 2005a,b). Closer to our work, Meinshausen and Rice (2006) and Meinshausen and Buehlmann (2004) consider estimates of m_0, the total number of false hull hypotheses in $\mathcal{H}(m)$.

Benjamini and Hochberg (1995) also introduces a step-down testing procedure that controls FDR. Order the collection of m hypotheses so that the p-values of the associated tests are sorted from smallest to largest (putting the most significant first),

$$p_{(1)} \le p_{(2)} \le \cdots \le p_{(m)} \ . \qquad (6)$$

The test of $H_{(1)}$ has p-value $p_{(1)}$, the test of $H_{(2)}$ has p-value $p_{(2)}$, and so forth. This step-down procedure of first compares the smallest p-value to the Bonferroni threshold. If $p_{(1)} > \alpha/m$, the BH procedure stops and does not reject any hypothesis. This step controls *FWER* in the weak sense at level α. If $p_{(1)} \le \alpha/m$, the procedure rejects $H_{(1)}$ and moves on to $H_{(2)}$. Rather than compare $p_{(2)}$ to α/m, however, the BH procedure compares $p_{(2)}$ to a larger threshold, $2\alpha/m$. In general, if we define $j_d = \min\{j : p_{(j)} > j\alpha/m\}$, then the BH step-down procedure rejects $H_{(1)}, \ldots, H_{(j_d-1)}$. Clearly this sequence of increasing thresholds produces more power than obtained by a Bonferroni procedure. If the p-values are independent, the inequality of Simes (1986) implies that this step-down procedure satisfies FDR. (This theorem was first proved by Benjamini and Hochberg (1995) who used a different proof.) This sequence of thresholds, however, does not control FWER(m) in general. This is the price we pay for the improvement in power. Subsequent papers (such as Benjamini and Yekutieli, 2001; Sarkar, 1998; Troendle, 1996) consider situations in which step-down testing controls FDR under certain types of dependence.

3 Alpha-Investing Rules

Alpha-investing introduces a framework for devising multiple testing procedures that control mFDR in a dynamic setting that allows streams of hypotheses. Alpha-investing resembles alpha-spending as used in sequential clinical trials. In a sequential trial, investigators routinely monitor the accumulating results for safety and efficacy. This monitoring leads to a sequence of tests of one (or perhaps a few) null hypotheses as the data accumulate. An alpha-spending (or error-spending) rule controls the level of such tests. Given an overall Type I error rate for the trial, say $\alpha = 0.05$, an alpha-spending rule allocates, or spends, α over a sequence of tests. As Tukey (1991) writes, "Once we have spent this error rate, it is gone." When repeatedly testing one null hypothesis H_0 as data accumulate, alpha-spending guarantees that $P(\text{reject } H_0) \leq \alpha$ when H_0 is true.

While similar in that they allocate Type I error over multiple tests, alpha-investing differs from alpha-spending in the following way. An alpha-investing rule earns additional probability toward subsequent Type I errors with each rejected hypothesis. Rather than treat each test as an expense that consumes its Type I error rate, an alpha-investing rule treats tests as investments, motivating our choice of name. In keeping with this analogy, we call the Type I error rate available to the rule its alpha-wealth. As with an alpha-spending rule, an alpha-investing rule can never spend more than its current alpha-wealth. Unlike an alpha-spending rule, however, an alpha-investing rule earns an increment in its alpha-wealth each time that it rejects a null hypothesis. For alpha-investing, Tukey's remark becomes "Rules that invest the error rate wisely earn more for further tests." A procedure that invests its alpha-wealth in testing hypotheses that are rejected accumulates additional wealth toward subsequent tests. The more hypotheses that are rejected, the more alpha-wealth it earns. If the test of H_j is not significant, however, an alpha-investing rule loses the invested α-level and its alpha-wealth decreases. The more wealth a rule invests in testing hypotheses that are not rejected, the less alpha-wealth remains for subsequent tests.

More specifically, an alpha-investing rule is a function $\mathcal{I}$ that determines the α-level for testing the next hypothesis in a sequence of tests. We assume an exogenous system external to the investing rule determines the next hypothesis to test. (Though not part of the investing rule itself, this exogenous system can use the sequence of rejections to pick the next hypothesis.) Let $W(k) \geq 0$ denote the alpha-wealth accumulated by an investing rule after k tests; $W(0)$ is the initial alpha-wealth. For example, one might conventionally set $W(0) = 0.05$ or 0.10. At step j, an alpha-investing rule sets the level α_j for testing H_j up to the maximum alpha it can afford. In other words, they rule should ensure that the wealth never goes negative. The level α_j for testing H_j typically depends upon the sequence of prior outcomes. Let $R_j \in \{0, 1\}$ be a sequence of binary random variables denoting the outcome of testing H_j:

$$R_j = \begin{cases} 1, & \text{if } H_j \text{ is rejected } (\alpha_j > p_j), \text{ and} \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

In general, an alpha-investing rule is a function of the initial wealth $W(0)$ and the string of prior test outcomes,

$$\alpha_j = \mathcal{I}_{W(0)}(\{R_1, R_2, \ldots, R_{j-1}\}). \tag{8}$$

The outcome of testing $H_1, H_2, \ldots, H_j$ determines the alpha-wealth $W(j)$ available for testing H_{j+1}. Let p_j denote the p-value of the test of H_j. If $p_j \leq \alpha_j$, the test rejects H_j. In this case, the investing rule earns an increment toward its alpha-wealth, called the *pay-out* and denoted by ω. If $p_j > \alpha_j$, the procedure does not reject H_j and its alpha-wealth decreases by $\alpha/(1-\alpha)$, slightly more than the cost extracted in alpha-spending. The change in the alpha-wealth is thus

$$W(j) - W(j-1) = \begin{cases} \omega & \text{if } p_j \leq \alpha_j \ , \\ -\alpha_j/(1-\alpha_j) & \text{if } p_j > \alpha_j \ . \end{cases} \tag{9}$$

If the p-value is uniformly distributed on [0,1], then the expected change in the alpha-wealth is $-(1-\omega)\alpha_j$. This suggests alpha-wealth decreases when testing a true null hypothesis. Other payment systems are possible; see the discussion in Section 8.

In the next section, we show that alpha-investing rules control mFDR. The initial alpha-wealth $W(0)$ controls the chance for rejecting the complete null hypothesis. Under the complete null hypothesis, an alpha-investing rule resembles an alpha-spending rule and controls FWER $\leq W(0)$. Results in the next section describe how the initial wealth $W(0)$ and pay-out ω lead to control of mFDR. Whereas $W(0)$ controls the probability of rejecting the complete null hypothesis, the pay-out ω controls how the alpha-investing rule performs once it rejects the complete null hypothesis.

The notion of compensation for rejecting a hypothesis captured in (9) allows one to build context-dependent information into the testing procedure. Suppose that the substantive context suggests that the first few hypotheses are likely to be rejected and that false hypotheses come in clusters. In this instance, one might consider using an alpha-investing rule that invests heavily at the start and after each rejection, as illustrated by the following rule. Assume that the most recently rejected hypothesis is H_{k^*}. (Set $k^* = 0$ when testing H_1.) If false hypotheses are clustered, an alpha-investing rule should invest most of the available wealth $W(k^*)$ to test H_{k^*+1}. One rule that does this is, for $j > k^*$,

$$\mathcal{I}_{W(0)}(R_1,\, R_2, \ldots, R_{j-1}) = \frac{W(j-1)}{1+j-k^*} \ . \tag{10}$$

This rule invests 1/2 of its current wealth in testing H_1 or the null hypothesis H_{k^*+1} that follows a rejected hypothesis. The α-level falls off quadratically if subsequent hypotheses are tested and not rejected. If the substantive insight is correct and the false hypotheses are clustered, then tests of H_1 or H_{k^*+1} represent "good investments." An example in Section 6 illustrates these ideas.

While it is relatively straightforward to devise investing rules, it may be difficult *a priori* to order the hypotheses in such a way that those most likely to be rejected come first. Such an ordering relies on the specific situation. Another complication is the construction of tests for which one can obtain the p-values that determine the alpha-wealth of an investing rule. In order to show that a procedure controls mFDR, we require the test of H_j to have the property that

$$\forall \theta \in \Theta, \quad E_\theta(V_j^\theta \mid R_{j-1},\, R_{j-2}, \ldots, R_1) \leq \alpha_j \ . \tag{11}$$

An alternative statement is that for all $\theta \in H_j$, $P_\theta(R_j = 1 \mid R_{j-1},\, R_{j-2}, \ldots, R_1) \leq \alpha_j$. Either statement amounts to requiring that, conditionally on having either accepted or rejected the prior $j-1$ hypotheses, the level of the test of H_j does not exceed α_j. The tests need not be independent.

Remark. Alpha-investing requires that the test of H_j maintain the stated α-level conditionally on the binary random variables $R_1, R_2, \ldots, R_{j-1}$. In particular, we note that the test is not conditioned on the test statistic (such as a z-score) or parameter estimate. Adaptive testing in a group sequential trial (e.g. Lehmacher and Wassmer, 1999) uses the information on the observed z-statistic at the first look. Tsiatis and Mehta (2003) show that using this information leads to a less powerful test compared to traditional group sequential tests that only use acceptance at the first look.

4 mFDR

Since we are dealing with a sequence of hypothesis, we define a sequence of evaluations of mFDR. The following definition also generalizes our subscript 1 on mFDR to an arbitrary value.

Definition 1 *Consider a procedure that sequentially tests hypotheses $H_1, H_2, \ldots$, where the hypothesis tested at step j can depend on the prior test results. Then we define*

$$mFDR_\eta(m) = \sup_{\theta\in\Theta} \frac{E_\theta\left(V^\theta(m)\right)}{E_\theta\left(R(m)\right)+\eta} \; . \tag{12}$$

A multiple testing procedure controls $\mathrm{mFDR}_\eta(m)$ at level α *if $mFDR_\eta(m) \leq \alpha$.*

We typically set $\eta = 1-\alpha$. Values of η near zero produce a less satisfactory criterion since under the complete null hypothesis, no system can generate an mFDR_0 better than 1 since $V^\theta(m) \equiv R(m)$.

Control of mFDR_η can be used to insure control of FWER in the weak sense. Under the complete null hypothesis $V^\theta(m) \equiv R(m)$; hence, control of $\mathrm{mFDR}_\eta(m) \leq \alpha$ implies that

$$E_\theta(V^\theta(m)) \leq \frac{\alpha\,\eta}{1-\alpha} \; .$$

With our typical choice $\eta = 1 - \alpha$, this inequality shows that $E_\theta(V^\theta(m)) \le \alpha$. As a result, control of $\mathrm{mFDR}_{1-\alpha}$ implies control of FWER in the weak sense at level α.

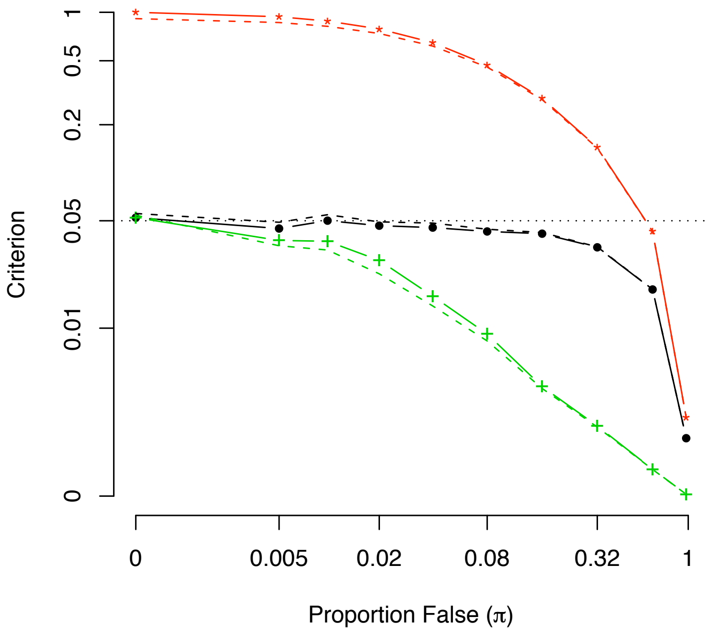
The following simulation shows that mFDR_η and FDR provide similar control of test procedures. Figure 1 shows simulated values of FDR and mFDR for testing $m = 200$ hypotheses using three procedures. The test procedures illustrate a range of methods. One is a naive, fixed-level test that rejects H_j if $p_j \le \alpha = 0.05$. The second is the standard Bonferroni procedure with $\alpha = 0.05/m$, and the third is the step-down BH procedure. The tested hypotheses $H_j : \mu_j = 0$ specify means of 200 normal populations. We set the values of the μ_j by sampling a spike-and-slab mixture. The mixture puts $100(1 - \pi_1)\%$ of its probability in a spike at zero; $\pi_1 = 0$ identifies the complete null hypothesis. The slab of this mixture – the signal – is a normal distribution, so that

$$\mu_j \sim \begin{cases} 0 & w.p. \quad 1-\pi_1 \\ N(0,\sigma^2) & w.p. \quad \pi_1 \end{cases} . \tag{13}$$

We set the variance of the signal component of the mixture to $\sigma^2 = 2\log m$ so that the standard deviation of the non-zero μ_j matches the bound commonly used in hard thresholding. The test statistics are independent, normally distributed random variables $Z_j \overset{\text{iid}}{\sim} N(\mu_j, 1)$ for which the two-sided p-values are $p_j = 2(1 - \Phi(|Z_j|))$. Given these p-values, we computed FDR and $\mathrm{mFDR}_{0.95}$ in a simulation with 10,000 trials. In the simulation, the amount of signal varies from $\pi_1 = 0$ 0 (the complete null hypothesis) to 1 (in which case $V^\theta(m) \equiv 0$).

Figure 1 shows the similarity of controls implied by FDR and mFDR. Both criteria identify the failure of naive testing; FDR and mFDR for naive testing approach the maximum value 1 for all but the largest amounts of signal (π_1 approaching 1). The criteria also provide similar assessments of the other two procedures; Bonferroni and step-down testing control both FDR and mFDR for all levels of signal. As π_1 increases and more hypotheses are rejected, FDR and mFDR both fall toward zero. The Bonferroni procedure is clearly more conservative, with smaller values of both criteria.

Figure 1: *Simulations show that FDR and mFDR$_\eta$ provide similar evaluations of testing procedures. The graphs show FDR (solid) and mFDR (dashed, $\eta = 0.95$) for the BH step-down procedure (●), the standard Bonferroni procedure (+), and a naive procedure that rejects each hypothesis at level $\alpha = 0.05$ (∗). The proportion of false null hypotheses varies from $\pi = 0$ (complete null) to 1.*



5 Uniform control of mFDR

Alpha-investing allows one to test hypotheses sequentially, with the choice of the next hypothesis possibly dependent on prior outcomes. Rather than begin with the full collection of ordered p-values, the testing proceeds one-at-a-time. This flexibility allows the statistician to stop testing, for example, after the first rejected hypothesis. To control a sequential testing procedure, we extend our definition of $mFDR_\eta(m)$ to random stopping times. Recall the definition of a stopping time: T is a stopping time, if the event $T \leq j$ can be determined by information known when the jth test is completed.

Definition 2 *The $mFDR_\eta(T)$ of a procedure for testing a stream of hypotheses H_1, H_2, … when the procedure is stopped at a finite stopping time T is*

$$mFDR_\eta(T) = \sup_{\theta\in\Theta} \frac{E_\theta\left(V^\theta(T)\right)}{E_\theta\left(R(T)\right)+\eta} \; .$$

We call the supremum over all stopping times of the sequential mFDR the universal mFDR:

Definition 3 *A procedure provides* universal control *of $mFDR_\eta$ at level α if*

$$\forall (T \in \mathcal{T}) \;\; mFDR_\eta(T) \leq \alpha \tag{14}$$

where $\mathcal{T}$ is the set of finite stopping times.

Before we turn to proving that alpha-investing provides universal control of mFDR, let us first prove half of theorem 1, namely that any procedure that universally controls mFDR$_\eta$ at level α has the property that "$E(V^\theta | R = r) \leq \alpha(r+\eta)$."

Definition 4 *Define the stopping time $T_{R=r} \equiv \inf_t\{t | R(t) = r\}$, where we take $T_{R=r}$ to be ∞ if the set is empty.*

Lemma 1 *For any scheme which uniformly controls $mFDR_\alpha$ at level α,*

$$E\; V^\theta(T^\theta_{R=r}) \leq \alpha(r+1-\alpha)\; .$$

Proof. Let $\tau \geq 1$ denote an arbitrary, but finite, number of tests. We can stop the process at $T \equiv T_{R=r} \wedge \tau$ to make it a finite stopping time. Thus $E_\theta(R(T)) \leq r$. We know by our hypothesis $\frac{E_\theta(V^\theta(T))}{E_\theta(R(T))+1-\alpha} \leq \alpha$ and so $E_\theta(V^\theta(T)) \leq \alpha(r+1-\alpha)$. Since this holds for all τ and $V^\theta(t)$ is bounded by r, we can take the limit to compute $E_\theta(V(T_{R=r})) \leq \alpha(r+1-\alpha)$.

□

5.1 Relating mFDR to FDR

mFDR and a variety of other modifications of FDR have been shown to be equivalent under a Bayesian setting (Tsai, Hsueh and Chen (2003)). We will show that when stopped at a fixed number of rejections, all of these definitions are equivalent.

To start, a bit of algebra shown in the appendix gives

$$-\gamma_R^2 \leq \mathrm{mFDR}_0 - \mathrm{FDR} - \gamma_R \frac{\rho \sigma_V}{\mu_R} \leq 0 \tag{15}$$

where μ_V and σ_V are the mean and standard deviation, respectively, of $V^\theta(j)$, $\gamma_R = \sigma_R/\mu_R$, and ρ is the correlation between $V^\theta(j)$ and $R(j)$. When the coefficient of variation γ_R is small, we see that FDR and mFDR_0 are close.

If $T_{R=r}$ is finite almost surely, then the standard deviation of R is identically zero, and hence $\gamma_R = 0$. So, mFDR_0 and FDR are identical. The following theorem is similar in spirit to Tsai et al. (2003).

Theorem 2 *Suppose $T_{R=r} < \infty$ almost surely. Then the procedure that stops when exactly r rejections have occurred has the following properties:*

1. $\gamma_R = 0$.
2. *$FDR = mFDR_0 = cFDR = eFDR = pFDR = E\,(V^\theta(r))/r$.*
3. *$FDR \leq \alpha\frac{r+2}{r+1}$ if the procedure has universal control of $mFDR_0$ at level α.*

Proof. Since $T_{R=r}$ is finite, we know that $R = r$ when we stop the process. Hence the standard deviation of R is zero and $\gamma_R = 0$. The second property follows from the various definitions:

$$\begin{aligned}
\text{FDR} &\equiv E(V^\theta(r)/R(r)) = E(V^\theta)/r \\
\text{mFDR}_0 &\equiv E(V^\theta(r))/E(R(r)) = E(V^\theta(r))/r \\
\text{cFDR} &\equiv E(V^\theta(r)/R(r)|R=r) = E(V^\theta(r)/R(r)) = E(V^\theta(r))/r \quad \text{[almost surely]} \\
\text{eFDR} &\equiv E(V^\theta(r))/R(r) = E(V^\theta(r))/r \quad \text{[almost surely]} \\
\text{pFDR} &\equiv E(V^\theta(r)/R(r)|R(r) \geq 1) = E(V^\theta(r)/R(r)) = E(V^\theta(r))/r \quad \text{[almost surely]}
\end{aligned}$$

and the fact that $R = r$ almost surely. The third property follows by the definitions.

□

If we restrict ourselves to alpha-investing rules with both ω and $W(0)$ set to α then it directly follows that

4. $\mathit{Var}(V^\theta(r)) \leq \alpha\, r$.

5. $P(V^\theta(r) \geq 1 + \alpha\, r + k\sqrt{r}) \leq e^{-k^2/2}$.

Property 5 has a relationship to Genovese and Wasserman (2002) and Genovese and Wasserman (2004a). In Genovese and Wasserman (2004a) a stochastic process of hypothesis tests is considered. Their approach contrasts with ours in that they use a stochastic process indexed by p-values, whereas we use a process indexed by the a priori order in which they are considered. In both cases, an appropriate martingale is constructed that allows converting expectation results into tail probabilities.

It is not obvious that the condition of $T_{R=r} < \infty$ can in fact be met. In the appendix, we explicitly construct a model for which this holds for many alpha-investing rules. In fact, we will construct a class of such models with this property.

Theorem 3 *There exists a model and an alpha-investing scheme such that $T_{R=r}$ is almost surely finite for all r.*

6 Alpha-Investing Rules Control mFDR

We start by observing that it is always possible to construct a procedure for which mFDR $\leq \alpha$. Concrete examples are the Bonferroni procedure or alpha-spending. The following theorem states that an alpha-investing rule $\mathcal{I}_{W(0)}$ with wealth determined by (9) controls $\text{mFDR}_{1-W(0)}$ so long as the pay-out ω is not too large. The theorem follows by showing that a stochastic process related to the alpha-wealth sequence $W(0), W(1), \ldots$ is a sub-martingale. Because the proof of this result relies only on the optional stopping theorem for martingales, we do not require independent tests, though this is the the easiest context in which to show that the p-values are honest in the sense required for (11) to hold.

Theorem 4 *An alpha-investing rule $\mathcal{I}_{W(0)}$ governed by (9) with initial alpha-wealth $W(0) \le \alpha\,\eta$ and pay-out $\omega \le \alpha$ controls $mFDR_\eta$ at level α.*

The above theorem also applies for the stopped version of mFDR and hence shows universal control. A proof of the theorem is in the appendix.

7 Examples

The examples in this section illustrate alpha-investing. We start with some general guidelines on how to construct alpha-investing rules. Then we build a generic example using these rules. Finally we will discuss how to construct an alpha-investing rule that is closely related to step-down procedures.

7.1 Designing alpha-investing rules

Alpha-investing frees the statistician to adapt the test procedure to the problem at hand. Prior hopes and beliefs can be incorporated into the design of the procedure. Patterns or structure among the hypotheses can also be incorporated. It is even possible to change dynamically the order of testing the hypothesis. Regardless of how these decisions are made, alpha-investing controls the mFDR. The statistician is thereby given maximal freedom for making choices without landing in the quagmire of uncorrected multiplicities.

This leaves the question of how these choices should be made. We can recommend a few policies. Whether none, several, or all of these suggestions are followed does not affect the applicability of our theorem that the mFDR will be controlled. Instead of following these suggestions, the hypothesis could be tested in a random order with a fixed alpha-spending rule.

Best-foot-forward policy: When using alpha-investing, it makes sense to test important hypotheses first. There are two reasons for doing so. Ideally, the initial hypotheses include those we believe most likely to be rejected. For example, in a testing drugs, it is common to test the primary endpoint before testing others. The drug being tested has been designed to have its largest impact on primary endpoint. The rejection of the leading hypotheses earns additional alpha-wealth toward tests of secondary endpoints.

There is, however, a devious side to this policy that should be avoided. If $\mathcal{H}(m)$ is contaminated with trivially-false hypotheses to produce alpha-wealth, then all subsequent tests are tainted. As an extreme example, suppose the first hypothesis is "gravity does not exist." After rejecting this hypothesis, a procedure has more alpha-wealth to use in later tests than present in $W(0)$. Most readers would, however, be uncomfortable using this additional alpha-wealth to test the primary endpoint of a drug. In a sense, it is important to allow an observer to ignore the list of tested hypotheses from some point onward. The design of the test procedure should then put the most interesting hypothesis first to insure that that when the reader stops, they have seen the most important results.

Spending-rate policies: Compared to ordering the hypotheses, deciding how much alpha-wealth to spend on each is relatively less important. If a procedure spends its alpha-wealth too slowly, it will have alpha-wealth left at the end of the sequence of tests. Since there is no reward for leaving this alpha-wealth unused, the procedure could have used more powerful tests. In this, spending at too slow a rate is inefficient.

Alternatively, spending too quickly means that the procedure may not have a chance to reject later hypotheses due to running out of alpha-wealth before reaching these. It seems reasonable to save at least a small amount of alpha-wealth for the future.

Spending-rate policies come in various forms. The previous policies assume that the order of the hypotheses completely capture our beliefs: best first, worst last. In other instances, it might be that hypotheses are clumped. For example, hypotheses 10-20 might test the activity of one family of chemicals, and hypotheses 21-30 might test a different family. Once a rejection is made, it carries implications for the entire family. In this case, it makes sense to use an alpha-investing rule that provides a burst of spending after a rejection is made in hopes that the next hypothesis is similar to the current hypothesis.

Dynamic-ordering policies: Suppose you are lucky enough to have a drug that might cure cancer *and* heart disease. Clearly, these two hypotheses should come first on the list. But what should come next? If the procedure rejects one of them but not the other, then the entire collection of subsequent tests depend on which one has been rejected. The appropriate test to try next is very different in these cases. We call a policy that adapts the order of the tests a dynamic-ordering policy. The nature of these policies is clearly domain-specific. Besides encouraging such policies as good designs, we cannot give explicit suggestions as to how to implement these.

Revisiting policies: Our theorems make no assumptions on how the various hypothesis are related. This flexibility makes it possible to test hypotheses that are very "close" to others. In fact, our theorems hold even if the same hypothesis is tested more than once. Of course, the second test must be conditional on the random variable R_j which represents the outcome of the first test. We call this "revisiting a hypothesis." For example, it might be sensible initially to test H_j at a small level $\alpha_j = 0.001$ (so as not to risk much alpha-wealth) and then test other hypotheses. If the test does not reject the first time, it might make sense to test it again at a higher level, say, 0.01. In this way, the procedure can manage its alpha-wealth among a variety of hypotheses – spending a little here and then a little there. This policy is very useful in mimicking step-down testing.

7.2 Example: Leveraging Domain Knowledge

The performance of alpha-investing improves if the investigator “knows the science.” If the investigator is able to order hypotheses *a priori* so that the procedure first tests those most likely to be rejected, then alpha-investing can reject more hypotheses than the step-down test of Benjamini and Hochberg. The full benefit is only realized, however, when one exploits this knowledge in the spending-rate policy.

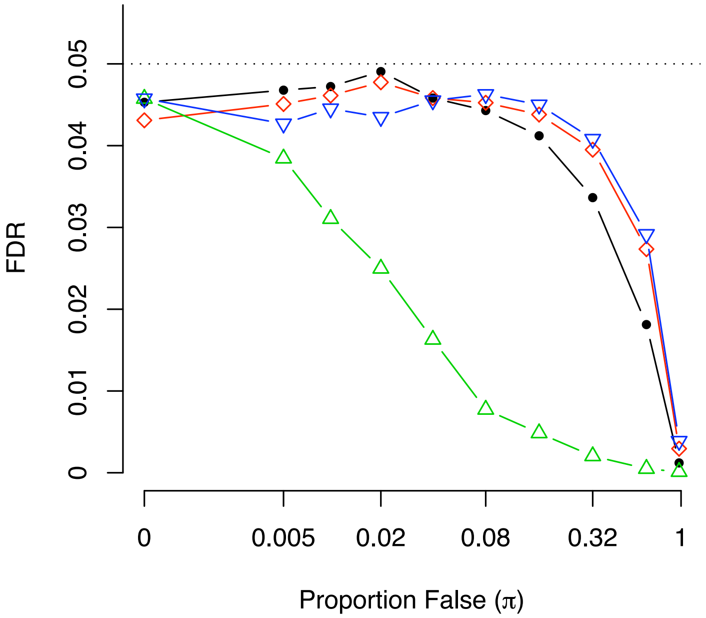
Suppose that the test procedure rejects H_{k^*} and is about to test H_{k^*+1}. Rather than spread its current alpha-wealth $W(k^*)$ evenly over the remaining hypotheses, a rule can invest more in testing the next hypothesis. For example, one can allocate $W(k^*)$ using a discrete probability mass function such as this version of the investing rule (10). A minor improvement insures that we run through all remaining alpha-wealth by the last hypothesis. Such a modification sets

$$\alpha_j = W(j-1)\left(\frac{1}{1+j-k^*} \vee \frac{1}{1+m-j}\right) \tag{16}$$

If one of these tests rejects a hypothesis, the procedure reallocates its wealth so that all is spent by the time the procedure tests H_m. Mimicking the language of financial investing, we describe this type of alpha-investing rule as “aggressive.”

The simulation summarized in Figures 2 and 3 compares step-down testing using the method of Benjamini and Hochberg to several alpha-investing rules. Two of these rules implement to aggressive alpha-investing; a third procedure implements a revisiting policy. To illustrate the behavior of aggressive alpha-investing, we show the performance of the alpha-investing using the rule (16) in a best case and a worst case scenario. For the best case, we assume that the investigator tests the hypotheses in the “correct” order implied by $|\mu_j|$. (Note that this does not imply that the tests are in order of increasing p-values.) In the worst case, the hypotheses are tested in random order, as if the domain knowledge is not accurate. The $m = 200$ hypotheses are defined as in the simulation in Section 3. (See equation 13.) This simulation also uses 10,000 samples. We set the level for all procedures to $\alpha = 0.05$; for alpha-investing, the initial wealth $W(0) = 0.05$, the pay-out $\omega = 0.05$, and $\eta = 0.95$. Figure 2 shows only FDR; as in Figure 1, FDR and mFDR are similar in this simulation. All four procedures

Figure 2: *Aggressive (△ accurate, ▽ inaccurate) alpha-investing rules control FDR, as does step-down testing (●) and a revisiting alpha-investing rule (◊).*



control FDR (and mFDR), as they should.

Aggressive alpha-investing easily controls FDR when the side-information accurately describes the hypotheses. Accurate side-information not only improves the power of the method (as shown below), but it also reduces the FDR of the procedure. When the tests are performed in random order, the FDR of aggressive testing is similar to that of the BH step-down procedure. The FDR of aggressive alpha-investing remains slightly below α when the side-information (the order of testing) is inaccurate until the level of signal π_1 approaches 1. When most hypotheses are false, all of these procedures easily control FDR.

Alpha-investing guarantees protection from too many false rejections, but how well does it find signal? Figure 3 compares the power of alpha-investing rules to that of step-down testing in the simulation. For each alpha-investing rule, Figure 3 shows the ratio of the number of rejected hypotheses, estimating

$$\frac{E_\theta\, R(200, \text{alpha-investing})}{E_\theta\, R(200, \text{BH})}$$

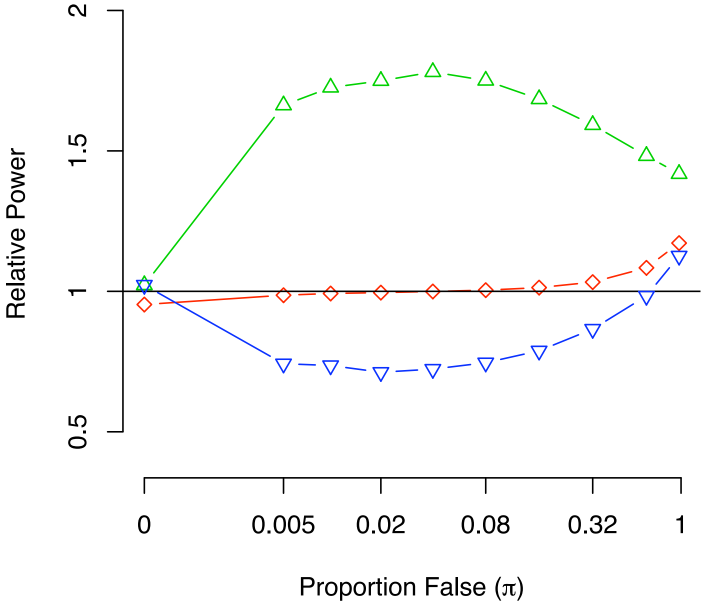
from the simulation for various choices of π_1. With accurate side information and some signal, aggressive alpha-investing rejects about 50% more hypotheses than step-down testing. If the side information is inaccurate, aggressive alpha-investing rejects no less than 75% of the number rejected by step-down testing. Alpha-investing using the revisiting policy described in the next section performs similarly to step-down testing. We designed this rule for alpha-investing to approximate step-down testing. The results in Figure 3 show that this rule has slightly less power under the complete null and that the power rises with the level of signal.

7.3 Comparison to Step-Down Testing

This section discusses testing a fixed set hypotheses as a stream of hypotheses. We assume that the set of hypotheses have no intrinsic order to them; they should all be treated symmetrically. A testing procedure that orders them breaks this symmetry and favors the test of some over others. To remove this favoritism, we randomly order the hypotheses and treat them as a sequence. This randomization treats them symmetrically, but leads to an inefficient test. Some sort of Rao-Blackwellization could help here, but would require introducing randomized tests. This section describes a simpler solution.

To achieve the desired symmetry, we use an idea mentioned in the previous section, a revisiting policy. The alpha-investing rule begins by investing small amounts of alpha-wealth in the initial test of each hypothesis. This conservative policy means that the procedure runs out of hypotheses well before it runs out of alpha-wealth. To improve its power, the rule uses its remaining alpha-wealth to take a second pass through the hypotheses that were not rejected in the first pass. Although we do not advocate this as a general procedure, it is allowed by our theorems. Gradually "nibbling" away at the hypotheses in this fashion produces a procedure that is similar to step-down testing. In fact, as the size of these nibbles goes to zero, the order that hypotheses are rejected

Figure 3: *Ratio of the number of rejected null hypotheses for alpha-investing to step-down testing. Aggressive alpha-investing using (16) exploits domain knowledge. When the domain knowledge accurately orders the H_j ($\triangle$), the procedure achieves higher power. The procedure has less power if the hypotheses are tested randomly ($\triangledown$). Alpha-investing using a revisiting policy ($\Diamond$) approximates the power of step-down testing.*



is precisely that from step-down testing. The point that each testing procedure stops is slightly different. Sometimes step down stops first and sometimes alpha-spending stops first.

To avoid taking many passes through the hypotheses without generating any rejects, we modify nibbling so as to take as bite a bite each time as is possible. Figures 2 and 3 show the FDR and relative power of this approach. Set the initial alpha-wealth $W(0) = \alpha$ and the pay-out $\omega = \alpha$. The procedure begins by testing each hypothesis in $\mathcal{H}(m)$ at level $\beta_1 = \alpha/(\alpha + m) \approx \alpha/m$, approximating the Bonferroni level. This choice for the level assures us that the procedure exactly exhausts its alpha-wealth if no hypothesis is rejected. If, however, some hypothesis is rejected, the procedure uses the earned alpha-wealth to revisit the hypotheses that were not rejected in the first pass. At the start of each pass, the algorithm divides its alpha-wealth equally among all remaining unrejected hypotheses and tests each at this common level. These steps continue until a pass does not reject a hypothesis. At this time the alpha-wealth is exactly zero and the procedure stops. Keep in mind that on the successive passes through the hypotheses, the fact that a hypothesis was previously tested must be used in computing the rejection region. The following worked example should clarify the details.

Example 1 *This example displays the exact computations necessary for one possible pattern of rejections. Namely, suppose that exactly one hypothesis is rejected at each pass. So to start, only one hypothesis has p-value less than the initial level β_1. Following (9), the rule pays $\beta_1/(1-\beta_1)$ for each test that it does not reject and earns α for rejecting $H_{(1)}$. Hence, after the initial test of each hypothesis at level β_1, the alpha-wealth grows slightly to*

$$\begin{aligned} W(m) &= W(0) + \omega - (m-1)\beta_1/(1-\beta_1) \\ &= \alpha + \alpha/m \end{aligned} \tag{17}$$

For large m, its alpha-wealth is virtually unchanged from $W(0)$.

For the second pass through the remaining null hypotheses, this testing procedure behaves as in the first pass. It again distributes its available alpha-wealth equally over *the remaining hypotheses so as to exhaust its alpha-wealth if none are rejected in this second look. The level invested in each test at this second pass, denoted β_2, is*

$$\beta_2 = \frac{W(m)}{W(m)+m-1} > \frac{\alpha}{m} \ .$$

To determine which tests are rejected in this second pass, assume that the p-values are uniformly distributed. Conditioning on $p_j > \beta_1$, this pass rejects any hypothesis for which $p_j \leq p^*$ with the threshold p^* determined by

$$P_0\left(\beta_1 < p_j \leq p^* \mid p_j > \beta_1\right) = \frac{p^* - \beta_1}{1-\beta_1} = \beta_2 \ ,$$

which implies that $p^* = \beta_1 + \beta_2 - \beta_1\beta_2 \approx 2\alpha/m$. Thus, the second pass rejects hypothesis with p-value smaller than $2\,\alpha/m$, approximately, mimicking the second threshold of the step-down test. As in the first pass through the hypotheses, the second pass approximately conserves the alpha-wealth of the procedure if one hypothesis is rejected. In this way, the investing rule gradually raises the threshold for rejecting hypotheses to the level used in step-down testing. If any hypothesis is rejected so that alpha-wealth remains after this second look, the testing procedure continues recursively.

Instead of spending equally on each hypothesis one could weight each hypothesis differently. This idea of using prior information is implicit in alpha-spending rules. Recently, Genovese and Wasserman (2004b) uses prior information on the hypotheses to devise a weighted Benjamini-Hochberg (called wBH) procedure. Following the ideas of this section, we can show that the wBH procedure satisfies the mFDR.

8 Discussion

One can regulate alpha-investing using other methods of compensating, or charging, for each test. The increment in the alpha-wealth defined in (9) is natural, with a fixed reward and penalty determined by whether the test rejects a hypothesis, say H_j. Because neither the payout ω nor the cost $\alpha/(1-\alpha)$ reveal p_j (other than to indicate if $p_j \leq \alpha_j$), subsequent tests need only condition on the sequence of rejections, $R_1, \ldots, R_j$. Under the complete null hypothesis, the expected pay-out is $-(1-\omega)\alpha_j$.

The following alternative method for regulating alpha-investing has the same expected pay-out, but varies the winnings when the test rejects H_j:

$$W(j) - W(j-1) = \begin{cases} \omega + \log(1-p_j) & \text{if } p_j \le \alpha_j \ , \\ \log(1-\alpha_j) & \text{if } p_j > \alpha_j \ . \end{cases} \tag{18}$$

We can show that alpha-investing governed by this "regulator" also satisfies the theorems shown previously. Because the reward reveals p_j when H_j is rejected, however, the investing rule must condition on p_j for any rejected prior hypotheses. This would seem to complicate the design of tests in applications in which the p-values are not independent. With the rewards defined as in (9), the tests need only condition on the binary outcomes R_j. (See equation 11.) Other methods for regulating the alpha-wealth could be desirable in other situations. We hope to pursue these ideas in future work.

We speculate that the greatest reward from developing a specialized testing strategy will come from developing methods that select the next hypothesis rather than specific functions to determine how α is spent. The rule (16) invests half of the current wealth in testing hypotheses following a rejection. One can devise other choices. Our work and those of others in information theory (Rissanen, 1983; Foster, Stine and Wyner, 2002), however, suggest that one can find universal alpha-investing rules. Given a procedure for ordering the hypotheses, a universal alpha-investing rule would lead to rejecting as many hypotheses as the best rule within some class. We would expect such a rule to spend its alpha-wealth a bit more slowly than the simple rule (16), but retain this general form.

Part of our motivation for alpha-investing arose in our work using stepwise regression for data mining (Foster and Stine, 2004). In this application, we compared forward stepwise regression to tree-based classifiers for predicting the onset of personal bankruptcy. To make regression competitive, we expanded the search for explanatory variables to include *all* possible interactions among more than 350 variables. This expansion of the scope produced more than 67,000 possible variables. Because so many of these variables are interactions (more than 98%), it is not surprising that most of the predictors identified by the search were interactions. Furthermore, because of the wide scope of this search, the procedure lacked power to find subtle effects that, while small, improve the predictive accuracy. It became apparent that a hybrid search that considers interactions $X_j * X_k$, say, *only after* including either X_j or X_k as main effects might be very effective. At the time, however, we lacked a method for controlling the variable selection when the scope of the search dynamically expands. We expect to exploit alpha-investing in this work in the future.

Another application for alpha-investing is in group-sequential clinical trials. In other work (Foster and Stine, 2006) we address the concept of adaptive design with a modification for alpha-investing. We show that the complaints raised in Tsiatis and Mehta (2003) about the efficiency of such tests can be mitigated by proper alpha-investing. At the same time, we allow the researcher freedom to design rules that guide how to spend or invest their alpha-wealth.

Appendix

Bounds defined in equation 15

From the inequality $1/x \geq 2 - x$ for all $x \geq 0$, we see see that for two random variables $X \geq 0$ and $Y \geq 0$ that,

$$\begin{aligned} E\left(\frac{X}{Y}\right) &= \frac{E(X/(Y/EY))}{EY} \\ &\geq \frac{E(X(2-(Y/EY))}{EY} \\ &= \frac{EX}{EY} + \frac{E(X(EY-Y))}{(EY)^2} \\ &= \frac{EX}{EY} - \frac{\mathrm{Cov}(X,Y)}{(EY)^2} \end{aligned}$$

Hence it follows that

$$\frac{\mathrm{Cov}(X,Y)}{(E\,Y)^2} \geq E\left(\frac{X}{Y}\right) - \frac{E\,X}{E\,Y}\ . \tag{19}$$

To simplify the notation in this section we omit the argument m and superscript/subscript θ, abbreviating, for example, $S = S^\theta(m)$. The inequality (19) implies that

$$\delta \ \equiv \ mFDR_0 - FDR = E(V)/E(R) - E(V/R)$$

$$\leq \quad \frac{\text{Cov}(V,R)}{E(R)^2} = \frac{\rho\sigma_V\sigma_R}{E(R)^2}$$

Hence, $\delta \leq \gamma_R \rho \sigma_V / \mu_R$. We can convert to using S (recall $S^\theta(m) = R(m) - V^\theta(m)$):

$$\begin{aligned} \delta &= E(R-S)/E(R) - E((R-S)/R) \\ &= 1 - E(S)/E(R) - 1 + E(S/R) \\ &= E(S/R) - E(S)/E(R) \end{aligned}$$

It then follows that $\delta \geq -\text{Cov}(S,R)/E(R)^2$. Since $\text{Cov}(S,R) = \sigma_R^2 - \text{Cov}(V,R)$ we get

$$\delta \geq (-\sigma_R^2 + \text{Cov}(V,R))/(ER)^2$$

As a result, $\delta \geq \text{Cov}(V,R)/(ER)^2 - \gamma_R^2$. Putting these together we see that

$$0 \geq \delta - \text{Cov}(V,R)/(ER)^2 \geq -\gamma_R^2.$$

8.1 Proof of Theorem 3

We first define the assumptions that define the class of models. We need an infinite sequence of hypothesis that we are testing. Under this set up we can define: $S^\theta(\infty) = \lim_{m\to\infty} S^\theta(m)$, namely the number of correct rejections over the entire sequence. We will be interested in the case where this is unbounded, namely, $S^\theta(\infty) = \infty$. If this occurs, then $T_{R=r}$ will be almost surely finite for all r. We state this precisely as:

Lemma 2 *If $S^\theta(\infty) = \infty$ then $T_{R=r} < \infty$ for all r.*

Proof. $R(m) \geq S^\theta(m) \to \infty$ as $m \to \infty$. Hence, for all r there exists an T such that $R(T) > r$. □

Notice that this lemma does not require a large fraction of the hypotheses to be false. It simply means that regardless of how many hypotheses that have been tested so far, there is always at least one more to be found that is significant enough that is likely that it will be found.

We say that an alpha-investing scheme is *thrifty* if it never commits all of its current alpha-wealth to the current hypothesis. A thrifty scheme never gives up the search for more signal; it always saves a bit more for the future. We say that an alpha-investing scheme is *hopeful* if it always spends at least some of its wealth on the next hypothesis. A hopeful, thrifty scheme consumes some of its alpha-wealth to test every hypothesis in an infinite sequence.

A parameter θ which has the property that for all m and all $W_m > 0$, the chance that the alpha-investing scheme $\mathcal{I}$ will reject at least one more hypothesis after time m is at least 0.5 is said to provide *continuous funding* for $\mathcal{I}$. Clearly for such a θ we have that $S^{\theta}(\infty) = \infty$.

Lemma 3 *Assume we are testing a sequence of hypothesis where each is a test of whether a normal distribution has mean zero. Let $\mathcal{I}$ be a thrifty and hopeful alpha-investing scheme. Consider an infinite set of integers: $J \subset \mathcal{N}$, $|J| = \infty$. Then there exists a θ, such that $\theta_j \neq 0$ only if $j \in J$ and that it will provide continuous funding for $\mathcal{I}$.*

Proof. Define $\underline{\alpha}_j$ to be the least amount that could ever be spent to test hypothesis j by the alpha-investing rule $\mathcal{I}$. Since there are only a finite number of possible sequences of accepts and rejects before time j, this is a minimum over a finite set. Since our rule is thrifty and hopeful, we know that the level of each test in this finite set is non-zero: $\underline{\alpha}_j > 0$.

Now for each $j \in J$, pick θ_j large enough so that the power of a $\underline{\alpha}_j$ level test is at least 0.5. Since J is infinite, we know that regardless of the sequence we have observed, there is a 0.5 chance of rejecting at least one more hypothesis.

□

We have now established the following version of Theorem 3.

Theorem 5 *For any thrifty and hopeful alpha-investing scheme there exists model such that it will be provided with continuous funding. Under such a condition, $T_{R=r}$ is finite almost surely.*

Proof of Theorem 4

We begin by defining a stochastic process indexed by j, the number of hypotheses that have been tested:

$$A(j) \equiv \alpha R(j) - V^\theta(j) + \eta\,\alpha - W(j)\ .$$

Our main lemma shows that $A(j)$ is a sub-martingale for alpha-investing rules with pay-out $\omega \le \alpha$. In other words we will show that $A(j)$ is "increasing" in the sense that

$$E_\theta\left(A(j) \mid A(j-1),\, A(j-2), \ldots, A(1)\right) \ge A(j-1)\ .$$

Theorem 4 uses the weaker fact that $E_\theta A(j) \ge A(0)$. By definition $V^\theta(0) = R(0) = 0$ so that $A(0) = \eta\,\alpha - W(0) \ge 0$ if we start with $W(0) \le \eta\,\alpha$. When $A(j)$ is a sub-martingale, the optional stopping theorem implies that for all finite stopping times M that $E_\theta A(M) \ge 0$. Thus,

$$\begin{aligned} E_\theta\Big(\alpha(R(M)+\eta) - V^\theta(M)\Big) &= E_\theta\left(W(M) + A(M)\right) \\ &\ge E_\theta\, A(M) \\ &\ge A(0) \ge 0\ . \end{aligned}$$

The first inequality follows because the alpha-wealth $W(j) \ge 0$ [*a.s.*], and the second inequality follows from the sub-martingale property. Thus, once we have shown that $A(j)$ is a sub-martingale, it follows that

$$E_\theta\, V^\theta(M) \le \alpha(E_\theta\, R(M) + \eta)\ ,$$

and

$$\mathrm{mFDR}_\eta(M) = \frac{E_\theta\, V^\theta(M)}{E_\theta\, R(M) + \eta} \le \alpha.$$

Thus to show Theorem 4 we need to prove the following lemma:

Lemma 4 *Let $V^\theta(m)$ and $R(m)$ denote the cumulative number of false rejections and the cumulative number of all rejections, respectively, when testing a sequence of null hypotheses $\{H_1,\, H_2,\, \ldots\}$ using an alpha-investing rule $\mathcal{I}_{W(0)}$ with pay-out $\omega \le \alpha$ and alpha-wealth $W(m)$. Then the process*

$$A(j) \quad \equiv \quad \alpha R(j) - V^\theta(j) + \eta\,\alpha - W(j)$$

is a sub-martingale,

$$E_\theta\left(A(m) \mid A(m-1), \ldots, A(1)\right) \geq A(m-1) \ . \tag{20}$$

Proof. Write the cumulative counts $V^\theta(m)$ and $R(m)$ as sums of indicators V_j^θ, $R_j \in \{0, 1\}$,

$$V^\theta(m) = \sum_{j=1}^m V_j^\theta \ , \qquad R(m) = \sum_{j=1}^m R_j \ .$$

Similarly write the accumulated alpha-wealth $W(m)$ and $A(m)$ as sums of increments, $W(m) = \sum_{j=0}^m W_j$ and $A(m) = \sum_{j=0}^m A_j$. Let α_j denote the alpha level of the test of H_j that satisfies the condition (11). The change in the alpha-wealth from testing H_j can be written as:

$$W_j = R_j\omega - (1-R_j)\alpha_j/(1-\alpha_j) \ ,$$

Substituting this expression for W_j into the definition of A_j we get

$$A_j = (\alpha - \omega)R_j - V_j^\theta + (1-R_j)\alpha_j/(1-\alpha_j) \ .$$

Since $R_j \geq 0$ and $\alpha - \omega \geq 0$ by the conditions of the lemma, it follows that

$$A_j \geq (1-R_j)\alpha_j/(1-\alpha_j) - V_j^\theta \ . \tag{21}$$

If $\theta_j \notin H_j$, then $V_j^\theta = 0$ and $A_j \geq 0$ almost surely. So we only need to consider the case in which the null hypothesis H_j is true. When H_j is true, $R_j \equiv V_j^\theta$ and (21) becomes

$$A_j \geq (1-R_j)\alpha_j/(1-\alpha_j) - R_j = (\alpha_j - R_j)/(1-\alpha_j) \ . \tag{22}$$

Abbreviate the conditional expectation

$$E_\theta^{j-1}(X) = E_\theta\left(X \mid A(1),\, A(2),\, \ldots, A(j-1)\right) \ .$$

Under the null, $E_\theta^{j-1} R_j \leq \alpha_j$ by the definition of this being an α_j level test. Taking conditional expectations in (22) gives $E_\theta^{j-1} A_j \geq 0$.

□

References

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statist. Soc., Ser. B*, **57**, 289–300.

Benjamini, Y. and Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, **29**, 1165–1188.

Braun, H. I. (ed.) (1994) *The Collected Works of John W. Tukey: Multiple Comparisons*, vol. VIII. New York: Chapman & Hall.

Dudoit, S., Shaffer, J. P. and Boldrick, J. C. (2003) Multiple hypothesis testing in microarray experiments. *Statistical Science*, **18**, 71–103.

Efron, B. (2005a) Large scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the Amer. Statist. Assoc.*, **100**, 96–104.

— (2005b) Selection and estimation for large-scale simultaneous inference. *Tech. rep.*, Department of Statistics, Stanford University, http://www-stat.stanford.edu/brad/papers/hivdata.

Foster, D. P. and Stine, R. A. (2004) Variable selection in data mining: Building a predictive model for bankruptcy. *Journal of the Amer. Statist. Assoc.*, **99**, 303–313.

— (2006) Theoretical foundations for adaptive testing using alpha-investing rules. *Tech. rep.*, Statistics Department, University of Pennsylvania.

Foster, D. P., Stine, R. A. and Wyner, A. J. (2002) Universal codes for finite sequences of integers drawn from a monotone distribution. *IEEE Trans. on Info. Theory*, **48**, 1713–1720.

Genovese, C. and Wasserman, L. (2002) Operating charateristics and extensions of the false discovery rate procedure. *Journal of the Royal Statist. Soc., Ser. B*, **64**, 499–517.

— (2004a) A stochastic process approach to false discovery control. *Annals of Statistics*, **32**, 1035–1061.

Genovese, Christopher, K. R. and Wasserman, L. (2004b) False discovery control with p-value weighting. *in progress*.

Gupta, M. and Ibrahim, J. G. (2005) Towards a complete picture of gene regulation: using Bayesian approaches to integrate genomic sequence and expression data. *Tech. rep.*, University of North Carolina, Chapel Hill, NC.

Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65–70.

Lehmacher, W. and Wassmer, G. (1999) Adaptive sample size calculations in group sequential trials. *Biometrics*, **55**, 1286–90.

Meinshausen, N. and Buehlmann, P. (2004) Lower bounds for the number of false null hypotheses for multiple testing of associations under general dependence. *Tech. Rep. 121*, ETH Zurich, http://stat.ethz.ch/ nicolai/.

Meinshausen, N. and Rice, J. (2006) Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Annals of Statistics*, **34**, 373–393.

Rissanen, J. (1983) A universal prior for integers and estimation by minimum description length. *Annals of Statistics*, **11**, 416–431.

Sarkar, S. K. (1998) Some probability inequalities for ordered Mtp$_2$ random variables: A proof of the Simes conjecture. *Annals of Statistics*, **26**, 494–504.

Simes, R. J. (1986) An improved bonferroni procedure for multiple tests of significance. *Biometrika*, **73**, 751–754.

Storey, J. D. (2002) A direct approach to false discovery rates. *Journal of the Royal Statist. Soc., Ser. B*, **64**, 479–498.

— (2003) The positive false discovery rate: a Bayesian interpretation and the q-value. *Annals of Statistics*, **31**, 2013–2035.

Troendle, J. F. (1996) A permutation step-up method of testing multiple outcomes. *Biometrics*, **52**, 846–859.

Tsai, C.-A., Hsueh, H.-m. and Chen, J. J. (2003) Estimation of false discovery rates in multiple testing: Application to gene microarray data. *Biometrics*, **59**, 1071 – 1081.

Tsiatis, A. A. and Mehta, C. (2003) On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika*, **90**, 367–378.

Tukey, J. W. (1953) The problem of multiple comparisons. Unpublished lecture notes.

— (1991) The philosophy of multiple comparisons. *Statistical Science*, **6**, 100–116.