4-2008

Alpha-Investing: A Procedure for Sequential Control of Expected False Discoveries

Dean Foster
University of Pennsylvania

Robert A. Stine
University of Pennsylvania

Recommended Citation

Foster, D., & Stine, R. A. (2008). Alpha-Investing: A Procedure for Sequential Control of Expected False Discoveries. *Journal of the Royal Statistical Society: Series B: Statistical Methodology, 70* (2), 429-444. http://dx.doi.org/10.1111/j.1467-9868.2007.00643.x

Alpha-Investing: A Procedure for Sequential Control of Expected False Discoveries

Abstract

Alpha-investing is an adaptive, sequential methodology that encompasses a large family of procedures for testing multiple hypotheses. 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 possesses 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.

Keywords

alpha spending, Bonferroni method, false discovery rate (FDR, mFDR), family-wise error rate (FWER), multiple comparisons

Disciplines

Business | Statistics and Probability

Alpha-investing: A procedure for sequential control of expected false discoveries

Dean P. Foster and Robert A. Stine*

Department of Statistics

The Wharton School of the University of Pennsylvania

Philadelphia, PA 19104-6340

July 24, 2007

Abstract

Alpha-investing is an adaptive, sequential methodology that encompasses a large family of procedures for testing multiple hypotheses. 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 possesses 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: alpha spending, Bonferroni method, false discovery rate (FDR, mFDR), family-wise error rate (FWER), multiple comparisons.

*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 hypotheses arrive sequentially in a stream. Streams of hypotheses arise naturally in 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 tests hypotheses 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). Genovese, Roeder and Wasserman (2006) describes other testing situations that offer domain knowledge.

Before we describe alpha-investing, we introduce a variation on the marginal false discovery rate (mFDR), an existing criterion for multiple testing. 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. A testing procedure controls mFDR at level α if

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

mFDR traditionally does not add 1 to the denominator; following our notation, we denote the traditional version mFDR_0. The 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, implying, $V \equiv R$ and $\text{mFDR}_0 = 1$.

An alpha-investing rule is an adaptive testing procedure that resembles an alpha-spending rule. An alpha-spending rule begins with an 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 by α_i. Once the alpha-wealth of the spending rule reaches zero, no further tests are allowed. Because the total chance for a Type I error is bounded by the initial alpha-wealth, alpha-spending naturally implements a Bonferroni rule. For multiple testing, however, Bonferroni rules are too conservative. An alpha-investing rule overcomes this conservatism by earning a contribution to its alpha-wealth for each rejected null hypothesis. Thus rejections beget more rejections. Alpha-investing rules further allow one to test an infinite stream of hypotheses, accommodate dependent tests, and incorporate domain knowledge, all the while controlling mFDR.

The sequential nature of alpha-investing allows us to enhance the type of control obtained through mFDR. By placing mFDR in a sequential setting, we can require that a testing procedure does 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 protection distinguishes what we call uniform control of mFDR. We show that alpha-investing uniformly controls mFDR.

In general, suppose we test hypotheses until the 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) implies that the expected value of V given that $R = r$ is less than or equal to $\alpha(r + 1)$. This conditional expectation requires that we introduce a stopping time. We defer these details to Section 5. As a preview, we offer

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 among tests of m hypotheses.*

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 false discovery rate (FDR). Next we discuss alpha-investing rules in Section 3. In Sections 4 and 5 we discuss uniform control of mFDR. We describe the design and performance of alpha-investing rules in Section 6. We close in Section 7 with a brief discussion.

2 Criteria and Procedures

Suppose that we have 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, $H_j : \theta_j = 0$.

We follow the standard notation for labeling correct and incorrect 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. A superscript θ distinguishes *unobservable* random variables from statistics such as $R(m)$. The random variable $V^\theta(m)$ denotes the number of false positives among the m tests, counting cases in which the testing procedure incorrectly rejects a true null hypothesis. $S^\theta(m) = R(m) - V^\theta(m)$ counts the number of correctly rejected null hypotheses. Under 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 $\mathcal{H}(m)$,

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

An important special case is control of FWER under the complete null hypothesis: $\text{P}_0(V^\theta(m) \geq 1) \leq \alpha$, where P_0 denotes the probability measure under the complete null hypothesis. We refer to controlling FWER under the complete null hypothesis as controlling FWER in the weak sense.

Bonferroni procedures control FWER. 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 simplest 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

$$\mathrm{P}_\theta(V^\theta(m) \geq 1) \leq \sum_{j=1}^{m} \mathrm{P}_\theta(V_j^\theta = 1) \leq \alpha \tag{3}$$

shows that this procedure controls $\mathrm{FWER}(m) \leq \alpha$. One need not distribute α equally over $\mathcal{H}(m)$; the inequality (3) requires only that the sum of the α-levels not exceed α. This observation suggests an 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). FDR is the expected proportion of false positives among rejected hypotheses,

$$\mathrm{FDR}(m) = E_\theta\left(\frac{V^\theta(m)}{R(m)} \mid R(m) > 0\right)\mathrm{P}(R(m) > 0)\ . \tag{4}$$

For the complete null hypothesis, $\mathrm{FDR}(m) = \mathrm{P}_0(R(m) > 0)$, which is $\mathrm{FWER}(m)$. Thus, test procedures that control $\mathrm{FDR}(m) \leq \alpha$ also control $\mathrm{FWER}(m)$ in the weak sense at level α. If the complete null hypothesis is rejected, FDR introduces a different type of control. Under the alternative, $\mathrm{FDR}(m)$ decreases as the number of false null hypotheses $m - m_0$ increases (Dudoit et al., 2003). As a result, $\mathrm{FDR}(m)$ 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 $\mathrm{P}(R > 0)$; see Storey, 2002, 2003) and the local false discovery rate $\mathrm{fdr}(z)$ (which estimates the false discovery rate as a function of the size of the test statistic; see Efron, 2005, 2007). Closer to our work, Meinshausen and Bühlmann (2004) and Meinshausen and Rice (2006) estimate

m_0, the total number of false hull hypotheses in $\mathcal{H}(m)$, and Genovese et al. (2006) weight p-values based on prior knowledge that identifies hypotheses that are likely to be false. Benjamini and Hochberg (1995) also considers mFDR_0 and mFDR_1, which they considered artificial. mFDR does not control a property of the realized sequence of tests; instead it controls a ratio of expectations.

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),

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)} \ . \tag{5}$$

The test of $H_{(1)}$ has p-value $p_{(1)}$, the test of $H_{(2)}$ has p-value $p_{(2)}$, and so forth. 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 $p_{(1)} \leq \alpha/m$, 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 $2\alpha/m$. In general, the BH step-down procedure rejects $H_{(1)}, \ldots, H_{(j_d-1)}$ for $j_d = \min\{j : p_{(j)} > j\alpha/m\}$. Clearly this sequence of increasing thresholds obtains more power than a Bonferroni procedure. If the p-values are independent, the inequality of Simes (1986) implies that this step-down procedure satisfies FDR. 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 the BH procedure controls FDR under certain types of dependence.

3 Alpha-Investing Rules

Alpha-investing resembles alpha-spending used in sequential clinical trials. In a sequential trial, investigators perform 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, 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."

While similar in that they allocate Type I error over multiple tests, 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. 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." 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.

More specifically, an alpha-investing rule $\mathcal{I}$ is a function 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 chooses which hypothesis to test next. (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. Conventionally, $W(0) = 0.05$ or 0.10. At step j, an alpha-investing rule sets the level α_j for testing H_j from 0 up to the maximum alpha-level it can afford. The rule must ensure that its wealth never goes negative. Let $R_j \in \{0, 1\}$ denote the outcome of testing H_j:

$$R_j = \begin{cases} 1, & \text{if } H_j \text{ is rejected } (p_j \leq \alpha_j), \text{ and} \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Using this notation, the investing rule $\mathcal{I}$ is a function of $W(0)$ and the prior outcomes,

$$\alpha_j = \mathcal{I}_{W(0)}(\{R_1, R_2, \ldots, R_{j-1}\}). \tag{7}$$

The outcome of testing $H_1, H_2, \ldots, H_j$ determines the alpha-wealth $W(j)$ available for testing H_{j+1}. If $p_j \leq \alpha_j$, the test rejects H_j and the investing rule earns a contribution to its alpha-wealth, called the *pay-out* and denoted by $\omega < 1$. We typically set $\alpha = \omega = W(0)$. If $p_j > \alpha_j$, its alpha-wealth decreases by $\alpha_j/(1 - \alpha_j)$, which is slightly more than the cost extracted in alpha-spending. The change in the alpha-wealth is thus

$$W(j) - W(j-1) = \begin{cases} \omega & \text{if } p_j \leq \alpha_j \ , \\ -\alpha_j/(1-\alpha_j) & \text{if } p_j > \alpha_j \ . \end{cases} \tag{8}$$

If the p-value is uniformly distributed on [0,1], then the expected change in the alpha-wealth is $-(1-\omega)\alpha_j < 0$. This suggests alpha-wealth decreases when testing a true null hypothesis. Other payment systems are possible; see the discussion in Section 7.

The notion of compensation for rejecting a hypothesis allows one to build context-dependent information into the testing procedure. Suppose that substantive insights suggest that the first few hypotheses are likely to be rejected and that subsequent false hypotheses come in clusters. In this instance, one might consider 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 heavily in the test of H_{k^*+1}. One rule that does this is, for $j > k^*$,

$$\mathcal{I}_{W(0)}(\{R_1, R_2, \ldots, R_{j-1}\}) = \frac{W(j-1)}{1+j-k^*} \; . \tag{9}$$

This rule invests 1/2 of its current wealth in testing H_1 or H_{k^*+1}. 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 needed p-values. To show that a testing procedure controls mFDR, we require that conditionally on the prior $j-1$ outcomes, the level of the test of H_j must not exceed α_j:

$$\forall \theta \in \Theta, \quad E_\theta(V_j^\theta \mid R_{j-1}, R_{j-2}, \ldots, R_1) \leq \alpha_j \; . \tag{10}$$

An equivalent statement is that for all $\theta \in H_j$, $\mathrm{P}_\theta(R_j = 1 \mid R_{j-1}, R_{j-2}, \ldots, R_1) \leq \alpha_j$. The tests need not be independent. Note that the test of H_j 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 than procedures that use only acceptance at the first look.

4 mFDR

The following definition generalizes the definition of mFDR given in (1).

Definition 1 *Consider a procedure that tests hypotheses H_1, $H_2, \ldots, H_m$. Then we define*

$$mFDR_\eta(m) = \sup_{\theta \in \Theta} \frac{E_\theta\left(V^\theta(m)\right)}{E_\theta\left(R(m)\right) + \eta} \,. \tag{11}$$

A multiple testing procedure *controls $mFDR_\eta(m)$ at level α* if $\mathrm{mFDR}_\eta(m) \le \alpha$. We typically set $\eta = 1 - \alpha$. Values of η near zero produce a less satisfactory criterion. Under the complete null hypothesis, no procedure can reduce mFDR_0 below 1 since $V^\theta(m) \equiv R(m)$. Control of mFDR_η provides control of FWER in the weak sense. Under the complete null hypothesis, $\mathrm{mFDR}_\eta(m) \le \alpha$ implies that

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

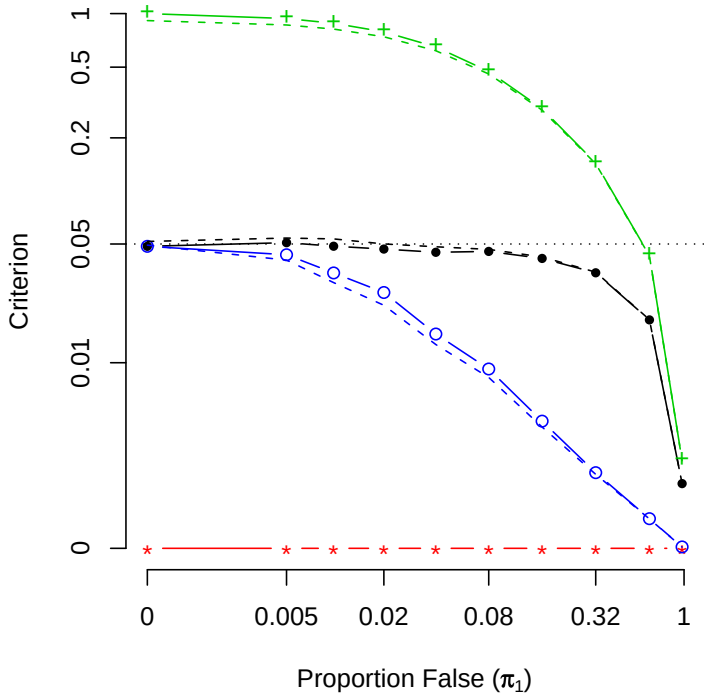
If $\eta = 1 - \alpha$, then $E_\theta(V^\theta(m)) \le \alpha$. Hence, control of $\mathrm{mFDR}_{1-\alpha}$ implies weak control of FWER at level α.

The following simulation illustrates the similarity of mFDR_η and FDR. The tested hypotheses $H_j : \mu_j = 0$ specify means of $m = 200$ populations. We set 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 is a normal distribution, so that

$$\mu_j \sim \begin{cases} 0 & w.p. \quad 1 - \pi_1 \\ N(0, \sigma^2) & w.p. \quad \pi_1 \end{cases} \,. \tag{12}$$

In the simulation, π_1 ranges from 0 (the complete null hypothesis) to 1 (in which case $V^\theta(m) \equiv 0$). We set $\sigma^2 = 2 \log m$ so that the standard deviation of the non-zero μ_j matches the bound commonly used in hard thresholding. The test statistics

Figure 1: *FDR and $mFDR_\eta$ provide similar control. The graph shows the simulated FDR (solid) and $mFDR_{0.95}$ (dashed) for the BH step-down procedure (•), oracle-based wBH (⋆), Bonferroni (∘), and a procedure that tests each hypothesis at level $\alpha = 0.05$ (+).*



are independent, normally distributed random variables $Z_j \overset{\text{iid}}{\sim} N(\mu_j, 1)$ for which the two-sided p-values are $p_j = 2\,\Phi(-|Z_j|)$.

Given these p-values, we computed FDR and $\text{mFDR}_{0.95}$ with 10,000 trials of four test procedures. Two procedures fix the level: one rejects H_j if $p_j \le \alpha = 0.05$ and the second rejects if $p_j \le \alpha/200 = 0.00025$ (Bonferroni). The other two are step-down tests: the BH and wBH procedures. For our implementation of the wBH procedure, we group the hypotheses into those that are false ($\mu_j \neq 0$) and those that are true. The hypotheses are weighted so that only false nulls are tested. Each false null receives weight $W_j = 1/S^\theta(m)$, and the true nulls get weight zero (see section 3 of Genovese et al., 2006). In effect, it is as though an oracle has revealed which hypotheses are false and the statistician applies the BH procedure to these rather than all m hypotheses.

Figure 1 confirms the similarity of control implied by FDR and mFDR. Both criteria identify the failure of naive testing; FDR and mFDR approach 1 unless π_1 approaches 1. The criteria also provide similar assessments of the other two procedures; Bonferroni and step-down testing control FDR and mFDR for all levels of signal. As π_1 increases and more hypotheses are rejected, both FDR and mFDR of all procedures fall toward zero. The FDR of the wBH procedure is identically zero because it only tests false null hypotheses and cannot incorrectly reject a hypothesis.

5 Uniform control of mFDR

Because alpha-investing proceeds sequentially, the testing can halt after some number of rejected hypotheses. To control such sequential testing, we extend the definition of $mFDR_\eta(m)$ to random stopping times. Recall the definition of a stopping time: T is a stopping time if the event $T \le j$ can be determined by information known when the jth test is completed.

Definition 2 *If T denotes a finite stopping time of a procedure for testing a stream of hypotheses H_1, H_2, ... then*

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

The supremum over stopping times of mFDR determines the uniform mFDR:

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

$$\forall (T \in \mathcal{T})\ mFDR_\eta(T) \le \alpha \tag{13}$$

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

Before proving that alpha-investing provides uniform control of mFDR, we prove half of theorem 1. We first define the relevant stopping time.

Definition 4 *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 testing procedure that uniformly controls $mFDR_\alpha$ at level α,*

$$E_\theta\, V^\theta(T_{R=r}) \le \alpha(r+1-\alpha)\ .$$

Proof. Let $\tau \ge 1$ 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 $E_\theta(R(T)) \le r$. We know by our hypothesis $E_\theta(V^\theta(T)) \le \alpha\,(E_\theta(R(T)) + 1 - \alpha) = \alpha(r+1-\alpha)$. Since this holds for all τ and $V^\theta(t)$ is bounded by r, we can take the limit as r increases.

□

mFDR and a variety of modifications of FDR become equivalent when stopped at a fixed number of rejections. A bit of algebra shows that

$$-\gamma_R^2 \le \mathrm{mFDR}_0 - \mathrm{FDR} - \gamma_R \frac{\rho\sigma_V}{\mu_R} \le 0\ , \tag{14}$$

where μ_V and σ_V are the mean and standard deviation, respectively, of $V^\theta(j)$, the coefficient of variation $\gamma_R = \sigma_R/\mu_R$, and ρ is the correlation between $V^\theta(j)$ and $R(j)$. When γ_R is small, 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, Hsueh and Chen (2003) who work in a Bayesian setting.

Theorem 2 *Suppose $T_{R=r} < \infty$ almost surely. Then a testing procedure that stops when 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. $FDR \le \alpha\frac{r+2}{r+1}$ *if the procedure has uniform control of* $mFDR_0$ *at level* α.

Further, for alpha-investing rules with $\alpha = \omega = W(0)$, then

4. $Var(V^\theta(r)) \le \alpha\, r$.
5. $P(V^\theta(r) \ge 1 + \alpha\, r + k\sqrt{r}) \le e^{-k^2/2}$.

The proofs of all five results are straightforward and hence omitted. Property 5 has a relationship to Genovese and Wasserman (2002, 2004). In Genovese and Wasserman

(2004) 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 hypotheses are considered. In both cases, an appropriate martingale converts expectations into tail probabilities.

The following theorem shows that an alpha-investing rule $\mathcal{I}_{W(0)}$ with wealth determined by (8) 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 (10) to hold.

Theorem 3 *An alpha-investing rule $\mathcal{I}_{W(0)}$ governed by (8) with initial alpha-wealth $W(0) \leq \alpha\,\eta$ and pay-out $\omega \leq \alpha$ controls $mFDR_\eta$ at level α.*

Theorem 3 also applies to the stopped version of mFDR and hence shows uniform control. A proof of this theorem is in the appendix.

Remark. It may not be obvious that the condition of a finite stopping time $T_{R=r} < \infty$ can in fact be met. Given an infinite sequence of hypotheses, define $S^\theta(\infty) = \lim_{m\to\infty} S^\theta(m)$. If $S^\theta(\infty) = \infty$ then $T_{R=r} < \infty$ for all r because $R(m) \geq S^\theta(m)$. A parameter θ that has the property (for all m and all $W(m) > 0$) that the chance that the alpha-investing procedure $\mathcal{I}$ will reject at least one more hypothesis after m is at least 0.5 is said to provide *continuous funding* for $\mathcal{I}$. Clearly for such a θ we have that $S^\theta(\infty) = \infty$.

We say that an alpha-investing procedure is *thrifty* if it never commits all of its current alpha-wealth to the current hypothesis. An alpha-investing procedure is *hopeful* if it always spends some wealth on the next hypothesis. A hopeful, thrifty procedure consumes some of its alpha-wealth to test every hypothesis in an infinite sequence. With these preliminaries, it can be shown that for any hopeful, thrifty alpha-investing procedure there exists a distribution P_θ that provides continuous funding so that $T_{R=r}$ is finite almost surely.

6 Examples

This section discusses practical issues of using alpha-investing. We start by offering general guidelines, or policies, on how to construct alpha-investing rules. An example constructs an alpha-investing rule that mimics step-down testing. We conclude with simulations that show the advantages of a good policy and compare alpha-investing to BH procedures.

Alpha-investing allows the statistician to incorporate prior beliefs into the design of the testing procedure while avoiding the quagmire of uncorrected multiplicities. This flexibility opens the question of how such choices should be made. We can recommend a few policies.

Best-foot-forward policy. Ideally, the initial hypotheses include those believed most likely to be rejected. For example, in a testing drugs, it is common to test the primary endpoint before others. Alpha-investing rewards this approach: the rejection of the leading hypotheses earns additional alpha-wealth toward tests of secondary endpoints.

Spending-rate policies. Compared to ordering the hypotheses, deciding how much alpha-wealth to spend on each is less important. That said, spending too slowly is inefficient. There is no reward for conserving alpha-wealth unused; the procedure could have used more powerful tests. Alternatively, spending too quickly may exhaust the alpha-wealth before testing every hypothesis. It seems reasonable to use a thrifty procedure that reserves some alpha-wealth for future tests.

Dynamic-ordering policies. Suppose you are lucky enough to have a drug that might cure cancer *and* heart disease. Clearly, these two hypotheses should be tested first. But what should come next if the procedure rejects one but not the other? The entire collection of subsequent tests depends on which of the initial hypotheses has been rejected. The nature of such dynamic-ordering policies is clearly domain-specific.

Revisiting policies. Our theorems make no assumptions on how the various hypothesis are related. This flexibility makes it possible to test hypotheses that

closely resemble others. In fact, our theorems hold if the *same* hypothesis is tested more than once, so long as subsequent tests condition on prior outcomes. For example, it might be sensible to test H_j initially 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 H_j the first time, it might make sense to test this hypothesis again at a higher level, say, 0.01. In this way, the procedure distributes its alpha-wealth among a variety of hypotheses – spending a little here and then a little there.

The following testing procedure illustrates how alpha-investing benefits when the investigator has accurate knowledge of the underlying science. If the investigator can order hypotheses *a priori* so that the procedure first tests those most likely to be rejected (best-foot-forward policy), then alpha-investing rejects more hypotheses than the step-down test of BH. The full benefit is only realized, however, when combined with a spending-rate policy. Suppose that the test procedure has rejected H_{k^*} and is about to test H_{k^*+1}. Rather than spread its current alpha-wealth $W(k^*)$ evenly over the remaining hypotheses, allocate $W(k^*)$ using a discrete probability mass function such as the following version of (9). This version consumes all remaining alpha-wealth by the last hypothesis by setting

$$\alpha_j = W(j-1)\left(\frac{1}{1+j-k^*} \vee \frac{1}{1+m-j}\right) \tag{15}$$

If a subsequent test 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*.

In the absence of domain knowledge, a revisiting policy produces an alpha-investing rule that imitates the step-down BH procedure. Begin by investing small amounts of alpha-wealth in the initial test of every 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 results in a procedure that resembles

step-down testing. In fact, as the size of these nibbles goes to zero, the order that hypotheses are rejected is precisely that from step-down testing. The point where each testing procedure stops is slightly different. Sometimes step-down testing stops first and sometimes alpha-spending stops first.

A few calculations clarify the connection to the BH procedure. To avoid taking many passes through the hypotheses without generating any rejects, increase the size of the nibbles so as to take as large a bite each time as possible. Set $\alpha = \omega = W(0)$. 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 level assures us that the procedure 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 and the alpha-wealth reaches zero.

On successive passes through the hypotheses, the fact that a hypothesis was previously tested must be used in computing the rejection region. Suppose that exactly one hypothesis is rejected at each pass. In the first pass, the p-value of one hypothesis is smaller than the initial level β_1. Following (8), 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

$$\begin{aligned} W(m) &= W(0) + \omega - (m-1)\beta_1/(1-\beta_1) \\ &= \alpha + \alpha/m \end{aligned} \tag{16}$$

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

For the second pass, the procedure again distributes its alpha-wealth equally over the remaining hypotheses. The level invested in each test at this second pass is

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

To determine which tests are rejected during the second pass, assume that the remaining p-values are uniformly distributed. Conditioning on $p_j > \beta_1$, this pass rejects any

hypothesis for which $p_j \le p^*$ with the threshold p^* determined by

$$\mathrm{P}_0\left(p_j \le p^* \mid p_j > \beta_1\right) = \frac{p^* - \beta_1}{1 - \beta_1} = \beta_2 \; .$$

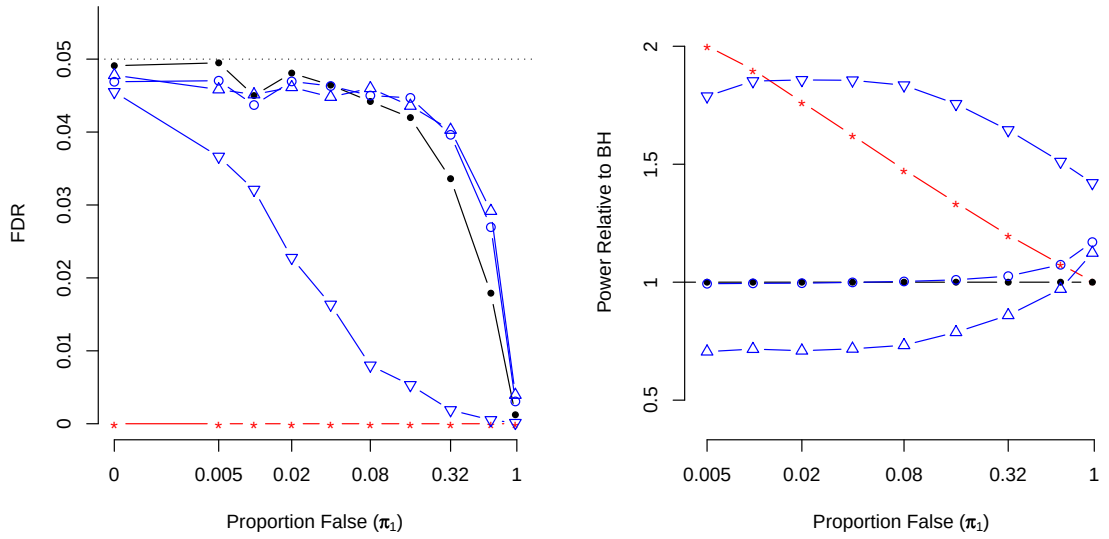
This implies that $p^* = \beta_1 + \beta_2 - \beta_1\beta_2 \approx 2\alpha/m$. Thus, the second pass approximately rejects any hypotheses with p-value is smaller than $2\,\alpha/m$, the second threshold of the step-down test. In this way, the investing rule gradually raises the threshold for rejecting hypotheses as in step-down testing. If any hypothesis is rejected during a pass over the remaining hypotheses, then alpha-wealth remains and the testing procedure continues recursively. Instead of spending equally on each hypothesis one could weight these hypotheses differently. This idea of using prior information is implicit in alpha-spending rules. The use of prior information also appears in the wBH procedure proposed in Genovese et al. (2006). Following the ideas of this section, we can show that the wBH procedure controls mFDR.

Simulations. Our first simulation compares three alpha-investing rules to step-down versions of the BH and wBH procedures. As in Section 3 for the wBH procedure, an oracle assigns weights

$$W_j = \begin{cases} 0 & \text{if } H_j \text{ is true} \\ 1/S^{\theta}(m) & \text{otherwise,} \end{cases}$$

to each hypothesis so that wBH only tests false hypotheses. For the sake of comparison, each replication of the simulation tests a fixed batch of $m = 200$ hypotheses. The 200 hypotheses are defined as in the simulation in Section 3 (see equation 12); this simulation also uses 10,000 samples. Of the alpha-investing rules, one implements the revisiting policy that mimics the BH procedure. The other two alpha-investing rules implement aggressive alpha-investing. To illustrate the impact of domain knowledge, we simulated the performance of alpha-investing using (15) in a best-case and a worst-case scenario. For the best case, the investigator tests the hypotheses in the order implied by $|\mu_j|$, testing the hypothesis with the largest $|\mu_j|$ first. The tests are ordered by the underlying means rather than the observed p-values. In the worst case, the hypotheses are tested in random order, indicating poor domain knowledge. We set the

Figure 2: *Comparison of the FDR and power of aggressive alpha-investing rules with accurate* ($\triangledown$) *and inaccurate* ($\triangle$) *domain knowledge to BH* ($\bullet$)*, oracle-based wBH* ($\star$)*, and a revisiting alpha-investing rule* ($\circ$)*. (a) All five procedures control FDR. (b) Better domain knowledge improves the power of alpha-investing relative to step-down testing. The vertical axis shows the ratio of the number of correctly rejected null hypotheses relative to BH.*



level for all procedures to $\alpha = 0.05$; for alpha-investing, the initial wealth $W(0) = \alpha = \omega$ and $\eta = 1 - \alpha$.

Figure 2(a) shows the FDR of each procedure. As in the prior simulation summarized in Figure 1, FDR and mFDR are quite similar in this simulation and so we have only shown FDR. All five procedures control FDR (and mFDR), as they should. As in Figure 1, the FDR of the wBH procedure is identically zero because it tests only false null hypotheses. For the other procedures that do not benefit from this oracle, it becomes easier to control FDR in the presence of more false null hypotheses (larger π_1). Accurate domain knowledge reduces the FDR of the aggressive procedure. When tested in random order (poor domain knowledge), the FDR of aggressive testing is similar to that of the BH step-down procedure.

Alpha-investing guarantees protection from too many false rejections, but how well

does it find signal? For each alpha-investing rule, Figure 2(b) shows the ratio of the number of correctly rejected hypotheses relative to the number rejected by BH, estimating

$$E_\theta\left(\frac{S^\theta(200, \text{test procedure})}{S^\theta(200, \text{BH})}\right)$$

from the simulation. With accurate domain knowledge, aggressive alpha-investing rejects in excess of 50% more hypotheses than the step-down BH procedure. Unless the problem offers few false null hypotheses ($\pi_1 < 0.02$, an average of 1 or 2 false null hypotheses), aggressive alpha-investing with good domain knowledge obtains the greatest power. Alpha-investing gains more from testing the hypotheses in the right order than BH gains from knowing which hypotheses to test. If the domain knowledge is poor, aggressive alpha-investing rejects no less than 70% of the number rejected by step-down testing. As expected from its design, alpha-investing using the revisiting policy performs similarly to step-down testing.

We also performed a simulation to investigate the performance of alpha-investing when applied to an infinite stream of hypotheses. Each null hypothesis specifies a mean ($H_t : \mu_t = 0$, $t = 1, 2, \ldots$), and the test statistic for each hypothesis is $Z_t \sim N(\mu_t, 1)$, independently. The simulation computes the FDR and power of aggressive alpha-investing using the rule (9) and a two-sided test of H_t. Two levels of signal are present in the simulation. Under one scenario, 10% of the null hypotheses are false; in the second, 20% are false. For each scenario, false null hypotheses cluster in bursts of varying size. We generated the sequence of means μ_t from a two-state Markov chain $\{Y_t\}$. In state 0, the null hypothesis holds, $\mu_t = 0$. In state 1, $\mu_t = 3$. The transition probabilities for the Markov chain are $p_{ij} = P(Y_{t+1} = j \mid Y_t = i)$. The probability of leaving state 0 varies over $p_{01} = (0.0025, 0.005, 0.010, 0.025, 0.05)$. To obtain a fixed percentage of false null hypotheses, we set $p_{10} = k\, p_{01}$ with $k = 4$ (20% false null hypotheses) and $k = 9$ (10%). Given k, increases in the transition probability p_{01} produce a more choppy sequence of hypotheses. We simulated 1,000 streams of hypotheses for each combination of k and p_{01}, beginning each with $Y_1 = 1$ (a false null). As in previous simulations, we set $W(0) = \alpha = \omega = 0.05$ and $\eta = 1 - \alpha$.

Figure 3 presents a snapshot of the results after testing 4,000 hypotheses. Although

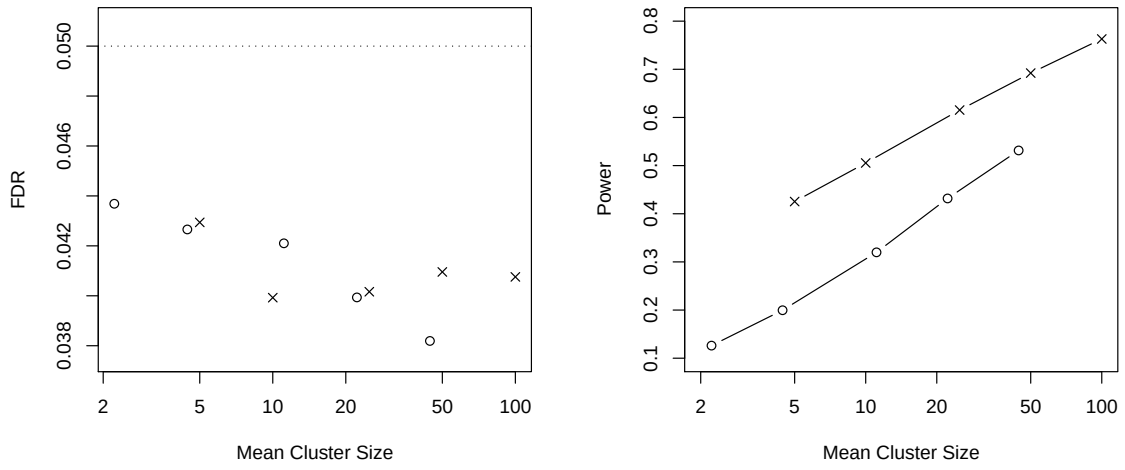
aggressive alpha-investing is thrifty and hopeful in the sense of Section 5, it can be shown that the sequence of means $\{\mu_t\}$ generated by the Markov chain do not provide continuous funding. Each sequence of tests eventually stops after a finite number of tests when the alpha-wealth runs out. Even so, at the point of the snapshot in Figure 3, each simulated realization of alpha-investing has enough alpha-wealth to reject further hypotheses. For example, the retained alpha-wealth $W(4000)$ averaged 0.003 with $k = 9$ and $p_{01} = 0.05$ (fewest, most choppy false hypotheses) up to 0.63 with $k = 4$ and $p_{01} = 0.0025$. In every case, Figure 3(a) shows that the FDR lies well below 0.05.

Figure 3(b) shows that aggressive alpha-investing has higher power when applied to sequences with a higher proportion of false null hypotheses that arrive in longer clusters. This situation affords the best opportunity to accumulate alpha-wealth that can be spent quickly to reject clustered false hypotheses. The power rises rapidly once a cluster of false null hypotheses is discovered. For example, the overall power is 0.51 with $k = 4$ and $p_{01} = 0.025$ (20% false nulls with mean cluster size 10). For finding the first null hypothesis of each cluster, however, the power falls to 0.20. This observation suggests that a revisiting policy that returns to hypotheses immediately preceding a rejected hypothesis would have higher power. In general, the power shown in Figure 3(b) rises roughly linearly in the log of the mean cluster size $1/p_{10}$. Aggressive alpha-investing also finds a higher percentage of the false null hypotheses when more are present. The power is consistently higher when 20% of the hypotheses are false ($k = 4$) than when 10% are false ($k = 9$). The gap between these scenarios diminishes slightly as the mean cluster size increases. As in the prior simulation, the higher power obtained for longer clusters is accompanied by smaller rates of false discoveries.

7 Discussion

The best-foot-forward policy raises a concern relevant to any multiple testing procedure that controls an FDR-like criterion. Suppose that $\mathcal{H}(m)$ is contaminated with trivially-false hypotheses that artificially produce alpha-wealth. Then all subsequent tests are tainted. As an extreme example, suppose the first hypothesis claims "gravity does not

Figure 3: *The power of aggressive alpha-investing increases with a higher proportion of false null hypotheses that arrive in longer clusters. The frames show (a) FDR after completing 4,000 tests and (b) power (∘ denotes 10% false nulls ($k = 9$) and ×, 20% false nulls ($k = 4$)).*



exist." After rejecting this hypothesis, the testing procedure has more alpha-wealth to use in subsequent tests than allocated by $W(0)$. Most readers would, however, be uncomfortable using this additional alpha-wealth to test the primary endpoint of a drug. Step-down tests share this problem. By rejecting the trivially false H_1, the level of the test of the first "real hypothesis" is $2\alpha/m$ rather than α/m. In this sense, it is important to allow an observer to ignore the list of tested hypotheses from some point onward. The design of the sequential test procedure should put the most interesting hypotheses first to insure that that when the reader stops, they have seen the most important results. By providing uniform control of mFDR, alpha-investing controls this criterion wherever the observer stops.

One can regulate alpha-investing using other methods of compensating, or charging, for each test. The increment in the alpha-wealth defined in (8) 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 $\alpha/(1-\alpha)$ reveals p_j (other than to

indicate if $p_j \leq \alpha_j$), subsequent tests need only condition on the sequence of rejections, $R_1, \ldots, R_j$. The following alternative method for regulating alpha-investing has the same expected pay-out, but varies the winnings when the test rejects H_j:

$$W(j) - W(j-1) = \begin{cases} \omega + \log(1 - p_j) & \text{if } p_j \leq \alpha_j \ , \\ \log(1 - \alpha_j) & \text{if } p_j > \alpha_j \ . \end{cases} \tag{17}$$

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. 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 (15) invests half of the current wealth in testing hypotheses following a rejection. One can devise other choices. Results in information theory (Rissanen, 1983; Foster, Stine and Wyner, 2002), however, suggest that one can find universal alpha-investing rules. A universal alpha-investing rule would reject on average as many hypotheses as the best rule within some class. We would expect such a rule to spend its alpha-wealth more slowly than the simple rule (15), but retain this general form.

Appendix

Proof of Theorem 3

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 $\omega \leq \alpha$. In other words we will show that $A(j)$ is "increasing" in the sense that

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

Theorem 3 uses the weaker fact that $E_\theta A(j) \geq A(0)$. By definition $V^\theta(0) = R(0) = 0$ so that $A(0) = \eta\,\alpha - W(0) \geq 0$ if $W(0) \leq \eta\,\alpha$. When $A(j)$ is a sub-martingale, the optional stopping theorem implies that for all finite stopping times M that $E_\theta A(M) \geq 0$. Thus,

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

The first inequality follows because the alpha-wealth $W(j) \geq 0$ $[a.s.]$, 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 $E_\theta\, V^\theta(M) \leq \alpha(E_\theta\, R(M) + \eta)$ and

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

Thus to show Theorem 3 we need to prove the following lemma:

Lemma 2 *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 $\omega \leq \alpha$ and alpha-wealth $W(m)$. Then the process*

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

is a sub-martingale,

$$E_\theta\left(A(m) \mid A(m-1), \ldots, A(1)\right) \geq A(m-1)\ . \tag{18}$$

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 (10). 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

$$A_j \geq (1 - R_j)\alpha_j/(1 - \alpha_j) - V_j^{\theta} \ . \tag{19}$$

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 (19) becomes

$$A_j \geq (1 - R_j)\alpha_j/(1 - \alpha_j) - R_j = (\alpha_j - R_j)/(1 - \alpha_j) \ . \tag{20}$$

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 (20) 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. (2005) Selection and estimation for large-scale simultaneous inference. *Tech. rep.*, Department of Statistics, Stanford University, http://www-stat.stanford.edu/brad/papers/.

— (2007) Size, power, and false discovery rates. *Annals of Statistics*, **35**, to appear.

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., Roeder, K. and Wasserman, L. (2006) False discovery control with p-value weighting. *Biometrika*, **93**, 509–524.

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.

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

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 Bühlmann, P. (2004) Lower bounds for the number of false null hypotheses for multiple testing of associations under general dependence. *Biometrika*, **92**, 893–907.

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.