Power of Ordered Hypothesis Testing

Lihua Lei LIHUA.LEI@BERKELEY.EDU
Department of Statistics, University of California, Berkeley

William Fithian WFITHIAN@BERKELEY.EDU
Department of Statistics, University of California, Berkeley

Abstract

Ordered testing procedures are multiple testing procedures that exploit a pre-specified ordering of the null hypotheses, from most to least promising. We analyze and compare the power of several recent proposals using the asymptotic framework of Li & Barber (2015). While accumulation tests including ForwardStop can be quite powerful when the ordering is very informative, they are asymptotically powerless when the ordering is weaker. By contrast, Selective SeqStep, proposed by Barber & Candès (2015), is much less sensitive to the quality of the ordering. We compare the power of these procedures in different régimes, concluding that Selective SeqStep dominates accumulation tests if either the ordering is weak or non-null hypotheses are sparse or weak. Motivated by our asymptotic analysis, we derive an improved version of Selective SeqStep which we call Adaptive SeqStep, analogous to Storey's improvement on the Benjamini-Hochberg procedure. We compare these methods using the GEOQuery data set analyzed by (Li & Barber, 2015) and find Adaptive SeqStep has favorable performance for both good and bad prior orderings.

1. Introduction

Since the invention of the Benjamini–Hochberg (BH) procedure (Benjamini & Hochberg, 1995), control of the false discovery rate (FDR) has gained widespread adoption as a reasonable measure of error in multiple hypothesis testing problems. In a typical setup, we observe a sequence of p-values $p_1, \dots, p_n$ corresponding to null hypotheses $H_1, \dots, H_n$, then apply some procedure to reject a subset of them. If we make R total rejections (also called "discoveries") of which V are true nulls (false discoveries), then the false discovery proportion (FDP) and false discovery rate (FDR) are defined respectively as

$$\mathrm{FDP} = \frac{V}{R \vee 1}, \quad \mathrm{FDR} = \mathbb{E}\,\mathrm{FDP}.$$

Let $\mathcal{S} = \{i : \ H_i \text{ is rejected}\}$ and $\mathcal{H}_0 = \{i : \ H_i \text{ is true}\}$, so that $R = |\mathcal{S}|$ and $V = |\mathcal{S} \cap \mathcal{H}_0|$.

We can classify testing problems into three types: batch testing, ordered testing and online testing. In batch testing, the ordering of hypotheses is irrelevant. The BH procedure and its many variants (e.g., Benjamini & Hochberg, 1997; Benjamini et al., 2006; Storey, 2002; Genovese et al., 2006) have been shown effective in this setting both in finite samples and asymptotically (Genovese & Wasserman, 2002; Storey, 2002; Storey et al., 2004; Ferreira & Zwinderman, 2006).

By contrast, in ordered testing, the ordering of hypotheses encodes prior information, typically telling us which hypotheses are most "promising" (i.e., most likely to be discoveries). For example, in genomic association studies, biologists could have prior knowledge about which genes are more likely to be associated with a disease of interest, and use this knowledge to concentrate statistical power on the more promising genes. Because prior information of this type is quite prevalent in scientific research, procedures that exploit it are attractive. Alternatively, the ordering may arise from the mathematical structure of the problem. For example, the co-integration test (Engle & Granger, 1987), which is widely used in macro-economics, involves testing $H_j : \mathrm{rank}(A) \le j$ where A is a coefficient matrix. Because the hypotheses are nested, it makes no sense to accept H_j and reject H_{j+1}. Other examples include sequential goodness-of-fit testing for the LASSO and other forward selection procedures such as Lockhart et al. (2014); Kozbur (2015); Fithian et al. (2015), which test $H_k : \mathcal{M}^* \subset \mathcal{M}_{k-1}$ where $\mathcal{M}^*$ is the true model and $\mathcal{M}_{k-1}$ is the model selected in $(k-1)$-th step. Section 2 reviews methods for ordered testing include ForwardStop (G'Sell et al., 2015), Accumulation Tests (Li & Barber, 2015), SeqStep and Se-

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

lective SeqStep (Barber & Candès, 2015).

Finally, in online testing, the ordering of hypotheses does not necessarily encode prior knowledge; rather, it imposes a constraint on the selection procedure, requiring that we decide to accept or reject H_i before seeing data for later hypotheses. Online procedures include α-investing (Foster & Stine, 2008), generalized α-investing (Aharoni & Rosset, 2014), LOND and LORD (Javanmard & Montanari, 2015). We will not address the online setting here.

In Section 2 we summarize existing ordered testing procedures and propose a new procedure, Adaptive SeqStep (AS), generalizing Selective SeqStep (SS). Our motivation is analogous to Storey (2002)'s improvement on the BH procedure. In Section 3, we introduce the varying coefficient two-groups (VCT) model and derive an explicit formula for asymptotic power of AS and SS under this model, comparing it to analogous results obtained by Li & Barber (2015) under similar asymptotic assumptions. Section 4 presents a detailed comparison of the asymptotic power of AS, SS, and accumulation tests (AT) under various regimes. In Section 5, we discuss selection of parameters and evaluate the finite-sample performance by simulation. In Section 6, we re-analyze the dosage response data from Li & Barber (2015), illustrating the predictions of our theory in real data. Section 7 concludes.

2. Ordered Testing and Adaptive SeqStep

Let π_0 denote the fraction of null p-values. Unless otherwise stated, we assume that null p-values are independent of the non-null p-values, and are drawn i.i.d. from the uniform distribution $U[0,1]$.

We now summarize several batch testing and ordered testing procedures and relate them to each other. For all of the procedures discussed below, the set of discoveries is of the form $\mathcal{S}(s,k) = \{i \le k : \ p_i \le s\}$: all p-values below some threshold s, which arrive before some stopping index k. Similarly $R(s,k), V(s,k)$, and $\mathrm{FDP}(s,k)$ denote the resulting values of V, R, and FDP if we select $\mathcal{S}(s,k)$.

Moreover, each method operates by defining some estimator of $\mathrm{FDP}(s,k)$, then maximizing the number of rejections $R(s,k) = |\mathcal{S}(s,k)|$ subject to a constraint that $\widehat{\mathrm{FDP}}(s,k) \le q$, the target FDR control level. For example, the BH procedure rejects all H_i with $p_i \le \hat{s}_{BH} = \max\{s : \ s \le qR(s,n)/n\}$, and may be formulated as

$$\max_{s\in[0,1]} R(s,n) \quad \text{s.t. } \widehat{\mathrm{FDP}}_{BH}(s) \le q; \tag{1}$$

$$\widehat{\mathrm{FDP}}_{BH}(s) = \frac{ns}{\sum_{i=1}^n I(p_i \le s) \vee 1} = \frac{\frac{1}{\pi_0}\mathbb{E}V(s,n)}{R(s,n)\vee 1}.$$

Benjamini & Hochberg (1995) show that $\mathrm{FDR}_{BH} \le \pi_0 q$. The procedure is very conservative when π_0 is small because $\widehat{\mathrm{FDP}}_{BH}(s)$ overestimates the true FDP. If π_0 were known, we could reduce $\widehat{\mathrm{FDP}}_{BH}(s)$ by a factor π_0, obtaining a more liberal threshold s (and therefore more rejections) while still controlling the FDR at level q.

In most problems, π_0 is unknown. Storey et al. (2004) propose an estimator based on counting the number of p-values *above* some fixed threshold $\lambda \in (0,1)$:

$$\hat{\pi}_0(\lambda) = \frac{1+\sum_{i=1}^n I(p_i > \lambda)}{n(1-\lambda)} = \frac{1+A(\lambda,n)}{n(1-\lambda)},$$

where $A(\lambda,k) = k - R(\lambda,k) = \sum_{i=1}^k I(p_i > \lambda)$ counts p-values exceeding the threshold. The logic is that, for high enough λ, the count $A(\lambda,n)$ will exclude most non-null p-values (commonly $\lambda = 0.5$). The Storey-BH (SBH) procedure then modifies (1), solving instead

$$\begin{aligned}\max_{s\in[0,\lambda]} R(s,n) \quad &\text{s.t. } \widehat{\mathrm{FDP}}_{SBH}(s;\lambda) \le q;\\ \widehat{\mathrm{FDP}}_{SBH}(s;\lambda) &= \hat{\pi}_0(\lambda)\,\widehat{\mathrm{FDP}}_{BH}(s)\\ &= \frac{s}{1-\lambda}\cdot\frac{1+A(\lambda,n)}{R(s,n)\vee 1},\end{aligned}$$

Storey et al. (2004) show that

$$\mathrm{FDR}_{SBH} \le (1-\lambda^{|\mathcal{H}_0|})q,$$

which can be much closer to q than $\pi_0 q$.

In ordered testing procedures, the choice variable is not the threshold s but rather the stopping index k. For example, Selective SeqStep (SS) (Barber & Candès, 2015) rejects all hypotheses H_i with $p_i \le s$ and $i \le \hat{k}_{SS}$ where

$$\hat{k}_{SS} = \max_{k\le n}\left\{k : \frac{1+\sum_{i=1}^k I(p_i > s)}{\sum_{i=1}^k I(p_i \le s)\vee 1} \le \frac{1-s}{s}q\right\},$$

for a given $s \in (0,1)$. This can be reformulated as

$$\begin{aligned}\max_{k\in\{0,\dots,n\}} R(s,k) \quad &\text{s.t. } \widehat{\mathrm{FDP}}_{SS}(k;s) \le q;\\ \widehat{\mathrm{FDP}}_{SS}(k;s) &= \frac{s}{1-s}\cdot\frac{1+A(s,k)}{R(s,k)\vee 1}.\end{aligned}$$

The close resemblance between $\widehat{\mathrm{FDP}}_{SS}(k;s)$ and $\widehat{\mathrm{FDP}}_{SBH}(s;\lambda)$ suggests writing $\widehat{\mathrm{FDP}}_{SS}$ as

$$\widehat{\mathrm{FDP}}_{SS}(k;s) = \hat{\pi}_0(s,k)\,\widehat{\mathrm{FDP}}_{BH}(s,k),$$

where the second argument k indicates evaluation on only the first k p-values.

If the threshold s is low, then $A(s,k)$ may include many non-null p-values, leading to an upwardly-biased estimate of ϕ_0. This observation motivates introducing an additional

parameter to improve the procedure, analogous to the improvement of SBH over BH. Defining

$$\widehat{\mathrm{FDP}}_{AS}(k; s, \lambda) = \frac{s}{1-\lambda} \cdot \frac{1 + A(\lambda, k)}{R(s,k) \vee 1},$$

we arrive at our proposal, which we call *Adaptive SeqStep* (AS): for some $0 \le s \le \lambda \le 1$, reject all hypotheses with $p_i \le s$ and $i \le \hat{k}_{AS}$, where

$$\hat{k}_{AS} = \max\{k : \widehat{\mathrm{FDP}}_{AS}(k; s, \lambda) \le q\}. \tag{2}$$

If $\lambda > s$ (say, $s = 0.1$ and $\lambda = 0.5$), then $\hat{\pi}_0(\lambda; k)$ may be much less upwardly biased than $\hat{\pi}_0(s; k)$, leading to a more powerful procedure. We investigate this power comparison in Sections 3–4.

The following theorem shows that AS achieves exact FDR control in finite samples.

Theorem 1. *Let $\mathcal{H}_0 \subset \{1, \dots, n\}$ denote the set of nulls, and assume that $\{p_i : i \in \mathcal{H}_0\}$ are independent of $\{p_i : i \notin \mathcal{H}_0\}$, and i.i.d. with distribution function F_0 that stochatically dominates $U[0,1]$. For $\hat{k}_{AS}$ defined as in (2),*

$$\mathrm{FDR}(\hat{k}_{AS}; s, \lambda) = \mathbb{E}\left(\frac{\sum_{i \in \mathcal{H}_0, i \le \hat{k}_{AS}} I(p_i \le s)}{\sum_{i \le \hat{k}_{AS}} I(p_i \le s) \vee 1} \right) \le q.$$

The proof of Theorem 1 is given in Appendix A.

Another class of ordered testing procedures are *accumulation tests* (AT) (Li & Barber, 2015), which include ForwardStop (G'Sell et al., 2015) and SeqStep (Barber & Candès, 2015) as special cases. Accumulation tests estimate FDP via

$$\widehat{\mathrm{FDP}}_{AT}(k) = \frac{1}{k} \sum_{i=1}^{k} h(p_i),$$

for some function $h \ge 0$ with $\int_0^1 h(x)dx = 1$, and rejects all hypotheses H_i with $i \le \hat{k}$ where

$$\hat{k} = \max\left\{ k : \widehat{\mathrm{FDP}}_{AT}(k) \le q \right\}.$$

ForwardStop corresponds to the case where $h(x) = -\log(1-x)$ and SeqStep corresponds to the case where $h(x) = CI(x > 1 - 1/C)$ for some $C > 0$.

In terms of our framework, accumulation tests solve

$$\max_{k \in \{0,\dots,n\}} R(1,k) \quad \text{s.t. } \widehat{\mathrm{FDP}}_{AT}(k) \le q.$$

The main difference between AT and AS is that the former rejects all hypotheses before $\hat{k}$, while the latter rejects only those smaller than threshold s. This means that AT will have full power if $\hat{k} \to n$, while the power of AS or SS is at most the average probability that a non-null p-value is less than s. On the other hand, unless nearly all of the early hypotheses are non-null, AT is likely to stop very early, as we will explore in Section 4.

3. Asymptotic Power Calculation

3.1. Varying Coefficient Two-group (VCT) Model

We now derive the asymptotic power of AS and SS under the following simple model:

Definition 1 (Varying Coefficient Two-groups (VCT) Model)**.** An VCT$(F_0, F_1; \pi(\cdot))$ model is a sequence of independent p-values $p_i \in [0,1]$ such that

$$p_i \sim (1 - \pi(i/n)) F_0 + \pi(i/n) F_1$$

for some distinct distributions F_0 and F_1 and a non-negative function $\pi(t) : [0,1] \to [0,1]$. F_0 and F_1 are the null and non-null distributions and $\pi(t)$ is the local non-null probability for $k = nt$.

For simplicity, we will take F_0 to be uniform. Following Genovese et al. (2006), we also assume that F_1 is strictly concave, so the density f_1 of non-null p-values is strictly decreasing; in other words, smaller p-values imply stronger evidence against the null.

The cumulative non-null probability $\Pi(t)$ is

$$\Pi(t) = \frac{1}{t} \int_0^t \pi(s) ds.$$

The quantity $\Pi(t)$ is essential to our results. It can be regarded as the average proportion of non-null hypotheses in the first nt-hypotheses since

$$\frac{\#\{i \le nt : i \notin \mathcal{H}_0\}}{nt} \approx \frac{1}{t} \int_0^t \pi(s) ds = \Pi(t).$$

Our setting is very similar to that of Li & Barber (2015) except that they impose conditions on $\Pi(t)$ directly. Proposition 1 in Appendix B reveals the relation between the VCT model and the assumptions of Li & Barber (2015).

3.2. Asymptotic Power for AS and SS

Because the SS method is a special case of the AS method with $\lambda = s$, it is sufficient to analyze the general case of

AS. Assuming a VCT model, for large n, we have

$$\begin{aligned}\widehat{\mathrm{FDP}}_{AS}(\lfloor nt \rfloor) &= \frac{s}{1-\lambda} \cdot \frac{1 + A(\lambda, \lfloor nt \rfloor)}{R(s, \lfloor nt \rfloor)} \\ \approx &\frac{s}{1-\lambda} \cdot \frac{(1-\Pi(t))(1-\lambda) + \Pi(t)(1-F_1(\lambda))}{(1-\Pi(t))s + \Pi(t)F_1(s)} \\ = &\frac{1 + \Pi(t)\left(\frac{1-F_1(\lambda)}{1-\lambda} - 1\right)}{1 + \Pi(t)\left(\frac{F_1(s)}{s} - 1\right)} \triangleq \mathrm{FDP}^*_{AS}(t).\end{aligned}$$

Because F_1 is strictly concave, we have $\frac{1-F_1(\lambda)}{1-\lambda} < 1 < \frac{F_1(s)}{s}$, and $\mathrm{FDP}^*_{AS}(t)$ is a strictly decreasing function of $\Pi(t)$. Thus, in the limit, $\widehat{\mathrm{FDP}}_{AS}(k)$ is determined by the fraction of non-nulls $\Pi(t)$, with more non-nulls leading to a lower estimate of FDP. Setting $\mathrm{FDP}^*_{AS}(t) = q$ and solving for $\Pi(t)$, we obtain the critical non-null fraction

$$\chi_{AS}(s, \lambda, q, F_1) = \frac{1-q}{1 - \frac{1-F_1(\lambda)}{1-\lambda} + q\left(\frac{F_1(s)}{s} - 1\right)}. \quad (3)$$

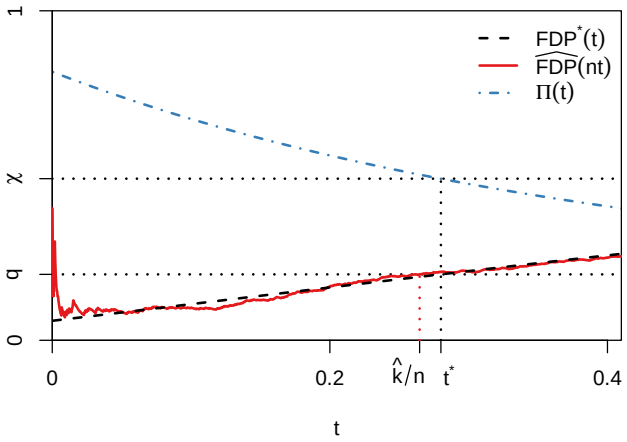

Figure 1. Illustration of the asymptotic behavior of AS. The broken curves show population limits for the simulation of Section 4, with parameters set to $\lambda = 0.5, s = q = \gamma = 0.2, \mu = 2, b = 5$. At t^*_{AS}, $\Pi(t) = \chi_{AS}$, leading to $\mathrm{FDP}^*_{AS}(t) = q$. The red curve shows a realization of $\widehat{\mathrm{FDP}}_{AS}(nt)$ with $n = 3000$.

If $\Pi(k/n) > \chi_{AS}$ then, with high probability, we will have $\widehat{\mathrm{FDP}}_{AS}(k) \le q$, implying $\hat{k}_{AS} \ge k$. The proportion $\hat{k}_{AS}/n$ of scanned hypotheses is approximately

$$t^*_{AS} = \max\{t : \Pi(t) \ge \chi_{AS}\}, \quad (4)$$

and the realized power is approximately

$$\begin{aligned}\mathrm{Pow}_{AS} &= \frac{\#\{i \le \hat{k} : i \notin \mathcal{H}_0, p_i \le s\}}{\#\{i \le n : i \notin \mathcal{H}_0\}} \\ &= \frac{\hat{k}}{n} \cdot \frac{\#\{i \le \hat{k} : i \notin \mathcal{H}_0, p_i \le s\}/\hat{k}}{\#\{i \le n : i \notin \mathcal{H}_0\}/n} \\ &\approx F_1(s) \cdot \frac{t^*_{AS}\Pi(t^*_{AS})}{\Pi(1)}. \quad (5)\end{aligned}$$

Theorem 2 confirms our heuristic approximations.

Theorem 2. *Consider a VCT model with*

- $\Pi(t)$ *is strictly decreasing and Lipschitz on* $[0,1]$ *with* $\Pi(1) > 0$*;*
- F_0 *is the uniform distribution on* $[0,1]$ *and* $f_1 = F_1'$ *is strictly decreasing on* $[0,1]$.

Then $\hat{k}_{AS}/n \overset{a.s.}{\to} t^*_{AS}$ *and*

$$\mathrm{Pow}_{AS} \overset{a.s.}{\to} F_1(s) \cdot \frac{t^*_{AS}\Pi(t^*_{AS})}{\Pi(1)} = F_1(s) \cdot \frac{\int_0^{t^*_{AS}} \pi(u)du}{\int_0^1 \pi(u)du},$$

with t^*_{AS} *defined as in* (4).

Interpreting (4–5), we see that if $\Pi(1) > \chi_{AS}$ then $t^*_{AS} = 1$ and $\hat{k}_{AS} = n$ with high probability: all $p_i < s$ are rejected and power is roughly $F_1(s)$. Conversely, if $\chi_{AS} > \sup_{t\in[0,1]} \Pi(t)$ then $\hat{k}_{AS} = o_p(n)$ and the method is asymptotically powerless. Figure 1 illustrates an intermediate case with $0 < t^*_{AS} < 1$.

From Theorem 2 we see there are two ways to increase the asymptotic power: either increase s (which we can do directly), or increase t^*_{AS}. To increase t^*_{AS} we must *decrease* $\chi_{AS}(s, \lambda, q, F_1)$, which itself is increasing in s and decreasing in λ.

Increasing λ always increases the asymptotic power. Because SS is a special case of AS, this implies that SS can always be improved by increasing λ above s, yielding a less biased estimator of the null fraction. Note, however, that taking $\lambda \to 1$ is not practical in finite samples: we still need large enough $A(\lambda, k)$ for the estimator to be stable.

Increasing s has an ambiguous effect on the asymptotic power. A smaller s leads to a smaller χ_{AS}, and therefore a larger stopping index $\hat{k}_{AS}$; however, it also applies a more stringent rejection threshold for hypotheses with $i \le \hat{k}_{AS}$. By contrast, larger s is more liberal for $i \le \hat{k}_{AS}$ but tends to give smaller $\hat{k}_{AS}$. If s is too large, χ could even exceed $\Pi(0)$, leading to a total loss of power. Small s avoids this catastrophe: if $\lim_{s\to 0} F_1(s)/s = \infty$ then $\lim_{s\to 0} \chi_{AS} = 0$. This implies that we can always have nonzero power if we take s small enough, but the power

can never exceed $F_1(s)$ even if $\hat{k}_{AS} \approx n$. Intuitively, then, using a large s is more aggressive, gambling that Π is large enough to overcome the larger value of χ_{AS}.

3.3. Asymptotic Power for AT

For AT, Li & Barber (2015) prove that

$$\mathrm{Pow}_{AT} \overset{a.s.}{\to} \frac{t^*_{AT}\Pi(t^*_{AT})}{\Pi(1)} \tag{6}$$

where $t^*_{AT} = \max\{t : \Pi(t) \geq \chi_{AT}\}$, where

$$\chi_{AT}(h, q, F_1) = \frac{1-q}{1-\nu}, \tag{7}$$

and $\nu = \mathbb{E}_{p\sim F_1} h(p)$. They also show that SeqStep, which uses $h(x) = CI(x > 1 - 1/C) : C \in (0, 1)$, is most powerful among all accumulation tests with h bounded by C. Reparameterizing with $\lambda = 1 - 1/C$, we can write $h(x) = \frac{1}{1-\lambda} I(x > \lambda)$. Then, $\nu = \frac{1-F_1(\lambda)}{1-\lambda}$ and

$$\chi_{AT} = \frac{1-q}{1-\frac{1-F_1(\lambda)}{1-\lambda}}. \tag{8}$$

Comparing (8) with (3) and recalling that $F_1(s) > s$ by concavity, we see that $\chi_{AS} < \chi_{AT}$. Therefore, $t^*_{AS} \geq t^*_{AT}$, implying that AT will tend to stop earlier than AS. Even so, AT could be more powerful due to the extra factor $F_1(s)$ in (5) which is absent from (6). If $f_1(x) = F_1'(x)$, we have

$$\nu = \frac{\int_0^1 h(p) f_1(p) dp}{\int_0^1 h(p) dp} \geq \inf_{x\in[0,1]} f_1(x),$$

where the last term equals $f_1(1)$ if F_1 is strictly concave. Thus, for any choice of h (bounded or otherwise), we have $\chi_{AT} \geq \frac{1-q}{1-f_1(1)}$.

4. Power Comparisons

In this section we analyze the results of Section 3 to extract further information about when each of AS, SS, and AT performs better or worse, and how and when the choice of s affects the performance of AS. There are three salient features of the VCT model to consider:

Signal density $\Pi(1) = \int_0^1 \pi(t)dt$ gives the expected total number of nulls. Note $\Pi(1) = 1 - \pi_0$.

Signal strength If the non-null p-values tend to be very small, we say the signals are strong.

Quality of the ordering If the prior information is very good then $\Pi(t)$ is steep, with $\Pi(0) = 1$ in the limit of very good information; if the prior ordering is completely useless then $\Pi(t) = \Pi(1)$ for all t.

First, note that if signals are very strong, then most of the non-null p-values are close to 1. In that case,

$$\frac{1-F_1(\lambda)}{1-\lambda} \approx 0 \quad \Rightarrow \quad \chi_{AS} \approx \frac{1-q}{1+q\left(\frac{F_1(s)}{s} - 1\right)},$$

even for relatively small values of λ, possibly including $\lambda = s$. As a result, λ plays a very small role in determining χ_{AS} and AS will behave similarly to SS. By contrast, if the signals are weaker, the difference is greater.

Second, if the ordering is very good, with $\Pi(0) \approx 1$ and $\Pi(t)$ correspondingly very steep, then we can afford to use a larger s for the AS procedure without worrying that $\chi_{AS} > \Pi(0)$ (though we still cannot allow χ_{AS} to exceed 1). By contrast, if the ordering is poor and $\Pi(t)$ is very flat, then a small change in s could move χ_{AS} from below $\Pi(1)$ (for which $\hat{k}_{AS} \approx n$) to above $\Pi(0)$ (for which $\hat{k}_{AS} = o_p(n)$), and so we are forced to be very cautious.

Finally, examining (7), we see that AT is highly aggressive compared to AS. Suppose $q = 0.1$. Then, regardless of the choice of h, AT is powerless unless at least 90% of the early hypotheses are non-null, requiring that either the signals are very dense or the ordering is very informative. In addition, the signals must be quite strong: even if $\Pi(0) = 1$, AT is asymptotically powerless unless

$$\nu = \mathbb{E}_{p\sim F_1} h(p) < q \ll 1 = \mathbb{E}_{p\sim F_0} h(p).$$

4.1. Numerical Results

We now illustrate the above comparisons with a numerical example. We consider the VCT model where F_0 is uniform and F_1 is the distribution of one-tailed p-values from a normal test. That is, $p = \bar{\Phi}(z) = 1 - \Phi(z)$ where $z \sim N(\mu, 1)$ and Φ is the standard normal CDF. Thus,

$$F_1(x) = \bar{\Phi}(\bar{\Phi}^{-1}(x) - \mu),$$

with μ determining the signal strength.

For the local non-null density, we take

$$\pi(t) = \gamma e^{-bt} \cdot \frac{b}{1-e^{-b}}, \quad \gamma \in (0, 1), \;\; b > 0.$$

The factor $b/(1 - e^{-b})$ is a normalization constant guaranteeing $\Pi(1) = \int_0^1 \pi(t)dt = \gamma$. Thus, γ determines the signal density, while b determines the quality of the prior ordering, with a larger b corresponding to a better ordering and $b \to 0$ corresponding to a useless ordering. b is implicitly upper-bounded by the constraint $\pi(0) = \gamma \cdot \frac{b}{1-e^{-b}} \leq 1$; let $b_{\max}$ denote the maximal value.

Figure 2 shows the asymptotic power for four methods, all using $q = 0.1$: AS with $s = q$ and $\lambda = 0.5$, AS with

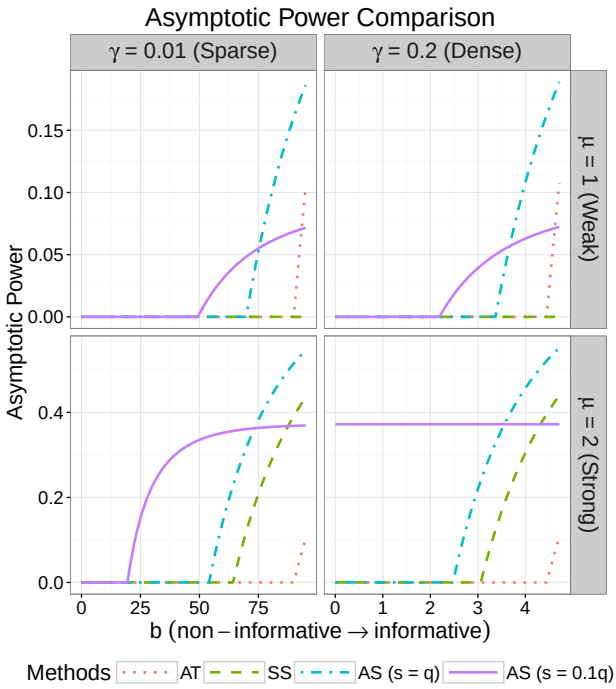

Figure 2. Asymptotic Power of AS (with $s = q$ and $s = 0.1q$), SS (with $s = q$) and AT (with $\nu = 0$) under four regimes: (sparse/weak) $\gamma = 0.01, \mu = 1$; (sparse/strong) $\gamma = 0.01, \mu = 2$; (dense/weak) $\gamma = 0.2, \mu = 1$; (dense/strong) $\gamma = 0.2, \mu = 2$. The x-axis measures b and a larger b corresponds to a more informative ordering.

$s = 0.1q$ and $\lambda = 0.5$, SS with $s = q$, and AT. AT is not implemented with a specific h, but rather with $\nu = 0$, giving the best possible power that any h could achieve. Four regimes are shown corresponding to two levels each of μ and γ: $\mu = 1$ (weak signals) vs. $\mu = 2$ (strong signals), and $\gamma = 0.01$ (sparse signals) vs. $\gamma = 0.2$ (dense signals). In each regime, we plot the asymptotic power of each method for $b \in (0, b_{\max}]$.

Unsurprisingly, all of the methods perform better with stronger, denser signals and better prior orderings, but their sensitivities to these parameters are quite different. Comparing the two AS methods, we see that smaller (less aggressive) s makes the method less sensitive to the ordering quality: its power is usually positive, but it is outperformed by AS($s = q$) when the ordering is excellent. AT is even more aggressive than the other two, and is asymptotically powerless unless the ordering is excellent.

SS is dominated by AS($s = q$) in all cases, as predicted, but the improvement is less dramatic when the signals are strong; in that case $\frac{1-F_1(\lambda)}{1-\lambda} \approx 0.05$ and $\frac{1-F_1(s)}{1-s} \approx 0.26$ are both small compared to $1 + q\left(\frac{F_1(s)}{s} - 1\right) \approx 1.66$.

5. Selection of Parameters

5.1. Selecting λ

As explained in Section 3.2, a large λ reduces χ_{AS} and improves asymptotic power. However, in finite samples, the procedure will be unstable if λ is too close to 1. One natural suggestion is to set $\lambda = 0.5$, analogous to Storey's suggestion for the Storey-BH procedure (Storey et al., 2004).

5.2. Selecting s

As discussed in Section 3.2, s has an ambiguous effect on the asymptotic power. The oracle choice s^*, which maximizes asymptotic power, is unknown in practice and depends on knowing parameters like $\Pi(t)$ and $F_1(x)$. Although we could plug in estimators of the parameters b and μ, or simply choose the value of s giving us the largest power on our data, the validity of such procedures would not be guaranteed by our results here.

In our view $s > q$ is intuitively unappealing because it would mean using a more liberal rejection cutoff than unadjusted marginal testing. We suggest $s = q$ as a heuristic, moderately aggressive default. This will give non-zero power as long as

$$\frac{F_1(q)}{1-q} > \frac{1-\Pi(0)}{\Pi(0)}. \tag{9}$$

(9) can be easily derived from (3) and (4), provided λ is close to 1 such that $\frac{1-F_1(\lambda)}{1-\lambda} \approx 0$, and is not too stringent. For example, if $q = 0.1$, $F_1(0.1) \geq 0.5$, then (9) holds provided $\Pi(0) > 0.64$. That is, if the non-nulls have reasonably strong signal and most of the early p-values are non-null, then $s = q$ is small enough. If we do not find these values of $F_1(0.1)$ and $\Pi(0)$ plausible, we can repeat this reasoning for smaller values of s until we arrive at assumptions we do find plausible.

5.3. Finite Sample Performance

Now we evaluate the finite sample performances of the above two heuristics for λ and s. Figure 3 displays the distribution of realized power for AS using $\lambda = 0.5$ vs. $\lambda = 0.95$, and $s = q$ vs. $s = s^*$. We set $q = 0.1, \gamma = 0.2, \mu = 2, b = 3.65$, in which case $\Pi(0) = 0.75, \Pi(1) = 0.2$. Each panel shows power for $n = 100, 500, 1000$, and $10,000$. For each setting we simulate 500 realizations of the fraction of all non-nulls that the method discovers. It is clear that large λ is less stable especially when n is small.

We see from Figure 3 that the performance of $\lambda = 0.5$ is more stable than that of $\lambda = 0.95$. On the other hand, the choice $s = q$ has a comparable power to the oracle approach. This justifies the simple choice $s = q$ as a moderately aggressive default choice for fairly strong signals and

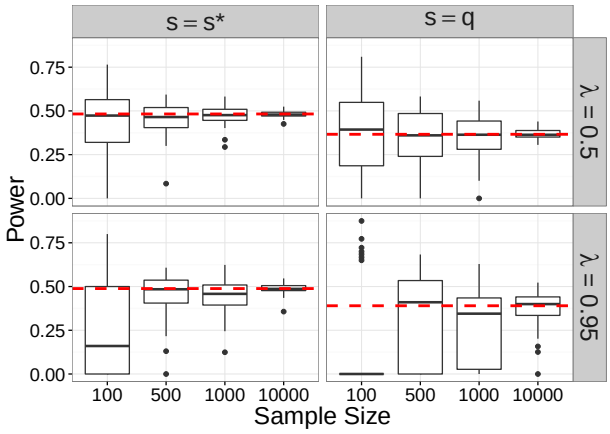

Figure 3. Finite-sample power using $s \in \{s^*, q\}$ and $\lambda \in \{0.5, 0.95\}$. Red dashed line corresponds to the asymptotic power.

a good prior ordering; note however that if the signals were much weaker or the ordering much worse, $s = q$ could be powerless.

6. Data Example: Dosage Response Data

Li & Barber (2015) analyzed the performance of several ordered testing methods using the GEOquery data of Davis & Meltzer (2007). In this section we reproduce and extend their analysis, adding the SS and AS methods as competitors. Where possible we have re-used the R code provided by Li & Barber (2015) at their website.

The GEOquery data consist of gene expression measurements at 4 different dosage levels of an estrogen treatment for breast cancer, plus control (no dose). At each of the 5 dosage levels, the gene expression of $n = 22,283$ genes is measured in 5 trials. The problem is to test whether each gene is differentially expressed in the lowest dosage versus the control condition, while using data from other dosage levels to obtain a prior ordering on the genes.

For each gene, Li & Barber (2015) carry out a t-test comparing expression under the highest dose versus the expression under the lowest dose and control, pooled together. Let $\widetilde{T}_i$ denote the t-statistic and $\tilde{p}_i$ the p-value for gene i using the high-dose data. Next, they compute one-sided permutation p-values p_i comparing lowest dose to control, using the sign of $\widetilde{T}_i$ to determine which side. Finally, they order the p-values $p_1, \ldots, p_n$ according to the ordering of $\tilde{p}_1, \ldots, \tilde{p}_n$ and apply an ordered testing procedure. For a more detailed explanation of the experiment, see Li & Barber (2015).

The top panel of Figure 4 reproduces Figure 6 of Li & Barber (2015) (with different axis limits), but including the SS and AS procedures analyzed here as competing methods. Both the HingeExp and AS methods perform quite well compared to the other methods, with SS coming in third place. In light of the foregoing theory, we can conclude that the high-dose data are doing an excellent job discriminating between null and non-null hypotheses — for example, the HingeExp method rejects the first 600 hypotheses at the $q = 0.1$ level, essentially implying that at least 540 of the first 600 genes in the ordering are truly non-null. The BH and Storey-BH methods, which are performed without any regard to the (highly informative) ordering, are unable to make any rejections.

For the lower panel of Figure 4, we repeat the same analysis, but with one change: instead of comparing with the highest dose to obtain $\widetilde{T}_i$, we instead compare with the second-lowest dose. This has the affect of attenuating the signal strength of $\widetilde{T}_i$, and thereby deteriorating the quality of the prior ordering. With a weaker ordering, all of the AT methods suffer major losses in power, so that the AS method is the clear winner, with SS in second place. As before, the BH and Storey methods have no power. This panel confirms the message of our theoretical analysis that AS and SS are more robust to weaker orderings.

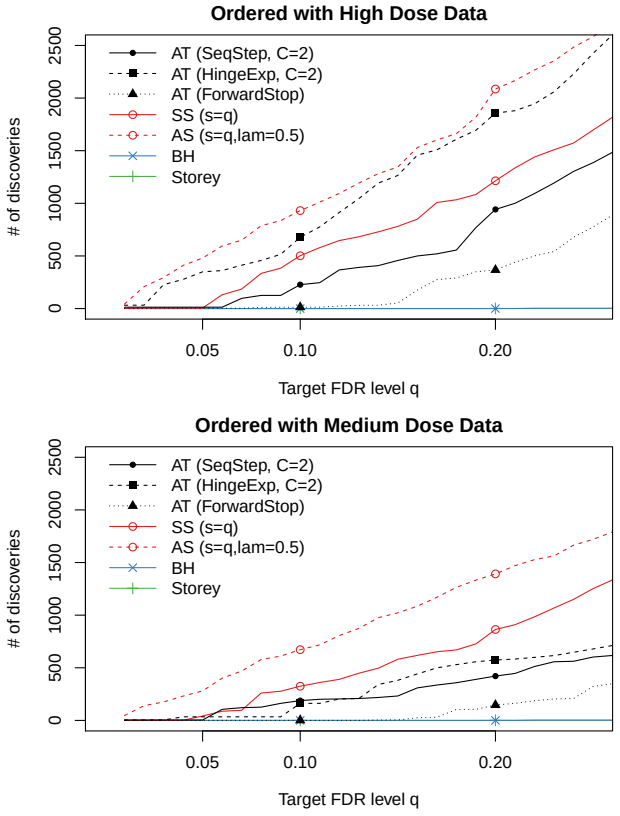

Figure 4. Power of AS, SS, and several AT methods on the dosage response data analyzed by Li & Barber (2015).

7. Conclusions and Future Directions

We have proposed Adaptive SeqStep (AS), which extends Selective SeqStep (SS) and improves on it in a manner analogous to Storey's improvement on the BH procedure. We have shown it controls FDR exactly in finite samples and analyzed its asymptotic power in detail, using the varying coefficient two-groups (VCT) model as a benchmark for comparing ordered testing procedures. For VCT models, we show that AS dominates SS asymptotically and outperforms AT except possibly in regimes with very good hypothesis ordering and very strong signals. Note that perfect ordering of hypotheses is implicit in the mathematical structure of many problems such as sequential goodness-of-fit testing; as a result AT could still be a suitable choice for these. Although we have proposed the heuristic $s = q$ for selecting s, it would be interesting to investigate whether there is a good way to estimate a good s from the data.

A natural extension of AS is to allow s and λ to be different across the hypotheses. Intuitively, for those which have a higher chance to be non-null, we could use a more liberal threshold. Once the conditions for exact FDR control are established, we can derive the "optimal" s-sequence and λ-sequence under the asymptotic framework. We leave this as future work.

Another interesting direction is to compare AS and AT with BH-type methods. Genovese & Wasserman (2002) has obtained the explicit formula for the power of BH procedure and it is not hard to obtain it in our more framework. The comparison should reveal more similarities and differences between these two genres.

Finally, AS is a natural fit for the "multiple knockoffs" extension of the knockoffs procedure suggested at the end of Barber & Candès (2015). Because the original knockoff procedure only produces 1-bit p-values, AS and SS are essentially equivalent, with $s = \lambda = 0.5$ the most natural settings of those parameters. However, a multiple-knockoff procedure could yield p-values lying in $\{\frac{i}{k+1} : i = 1, 2, \ldots, k+1\}$ by using k knockoffs for each predictor variable. It would be interesting to see whether using the AS instead of SS procedure would give a meaningful improvement.

Acknowledgments

We thank the anonymous reviewers for their helpful comments, which greatly improved this work.

References

Aharoni, Ehud and Rosset, Saharon. Generalized α-investing: definitions, optimality results and application to public databases. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):771–794, 2014.

Barber, Rina Foygel and Candès, Emmanuel J. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

Benjamini, Yoav and Hochberg, Yosef. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 289–300, 1995.

Benjamini, Yoav and Hochberg, Yosef. Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics*, 24(3):407–418, 1997.

Benjamini, Yoav, Krieger, Abba M, and Yekutieli, Daniel. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507, 2006.

Davis, Sean and Meltzer, Paul S. Geoquery: a bridge between the gene expression omnibus (geo) and bioconductor. *Bioinformatics*, 23(14):1846–1847, 2007.

Engle, Robert F and Granger, Clive WJ. Co-integration and error correction: representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, pp. 251–276, 1987.

Ferreira, JA and Zwinderman, AH. On the benjamini–hochberg method. *The Annals of Statistics*, 34(4):1827–1849, 2006.

Fithian, William, Taylor, Jonathan, Tibshirani, Robert, and Tibshirani, Ryan. Selective sequential model selection. *arXiv preprint arXiv:1512.02565*, 2015.

Foster, Dean P and Stine, Robert A. α-investing: a procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(2):429–444, 2008.

Genovese, Christopher R. and Wasserman, Larry. Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):499–517, 2002.

Genovese, Christopher R., Roeder, Kathryn, and Wasserman, Larry. False discovery control with p-value weighting. *Biometrika*, 93(3):509–524, 2006.

G'Sell, Max Grazier, Wager, Stefan, Chouldechova, Alexandra, and Tibshirani, Robert. Sequential selection procedures and false discovery rate control. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2015.

Javanmard, Adel and Montanari, Andrea. On online control of false discovery rate. *arXiv preprint arXiv:1502.06197*, 2015.

Kozbur, Damian. Testing-based forward model selection. *arXiv preprint arXiv:1512.02666*, 2015.

Li, Ang and Barber, Rina Foygel. Accumulation tests for fdr control in ordered hypothesis testing. *arXiv preprint arXiv:1505.07352*, 2015.

Lockhart, Richard, Taylor, Jonathan, Tibshirani, Ryan J, and Tibshirani, Robert. A significance test for the lasso. *Annals of statistics*, 42(2):413, 2014.

Storey, John D. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

Storey, John D, Taylor, Jonathan E, and Siegmund, David. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1): 187–205, 2004.