Power of Ordered Hypothesis Testing

Lihua Lei, William Fithian

Department of Statistics, UC Berkeley

Ordered Hypothesis Testing Problem

Setup of ***Multiple Testing Problem***: a sequence of hypotheses $H_1, \ldots, H_n$.

- $\mathcal{H}_0 = \{i : H_i \text{ is true}\}$, $\mathcal{S} = \{i : H_i \text{ is rejected}\}$, $R = |\mathcal{S}|$, $V = |\mathcal{S} \cap \mathcal{H}_0|$;
- $\mathrm{FDP} = \frac{V}{\max\{R,1\}}$ be the *False Discovery Proportion*;
- $\mathrm{FDR} = \mathbb{E}\mathrm{FDP}$ be the *False Discovery Rate*.
- A procedure that control FDR at level 0.1 produces a rejection set $\mathcal{S}$ with roughly 90% being the true discoveries.

Setup of ***Ordered Testing Problem***: $H_1, \ldots, H_n$ sorted via prior knowledge.

- Domain knowledge might be used to indicate which hypothesis is more "promising", i.e. likely to be rejected;
- Heuristically, more focus should be put on "promising" hypotheses.

A Unified Framework of Existing Procedures

Most Existing Multiple Testing Procedures fall into the following framework:

- Input: a sequence of p-values $p_1, \ldots, p_n$ associated with the hypotheses $H_1, \ldots, H_n$, usually assuming $p_i \sim U([0,1])$ for null hypothesis;
- Rejection Rule: the rejection set $\mathcal{S}$ has the form
$$\mathcal{S}(s;k) = \{i : p_i \le s, i \le k\},$$
- Choice of s and k: maximize the number of rejection $R(s;k) = |\mathcal{S}(s;k)|$, subject to the constraint
$$\widehat{\mathrm{FDP}}(s;k) \le q,$$
with a target level q, where $\widehat{\mathrm{FDP}}$ is a procedure-specified estimator of FDP.

BH Procedure: (Benjamini & Hochberg, 1997) $k \equiv n$ and
$$\widehat{\mathrm{FDP}}_{BH}(s) = \frac{ns}{\sum_{i=1}^n I(p_i \le s) \vee 1};$$

Storey's BH Procedure: (Storey et al., 2004) $k \equiv n$ and
$$\widehat{\mathrm{FDP}}_{SBH}(s;\lambda) = \frac{s}{1-\lambda} \cdot \frac{\sum_{i=1}^n I(p_i > \lambda) + 1}{\sum_{i=1}^n I(p_i \le s) \vee 1};$$

Selective Seqstep (SS): (Barber & Candès, 2015) s is pre-fixed and
$$\widehat{\mathrm{FDP}}_{SS}(k;s) = \frac{s}{1-s} \cdot \frac{\sum_{i=1}^k I(p_i > s) + 1}{\sum_{i=1}^k I(p_i \le s) \vee 1};$$

Accumulation Test (AT): (Li & Barber, 2015) $s \equiv 1$ and for $h \ge 0$ with $\int_0^1 h(x)dx = 1$,
$$\widehat{\mathrm{FDP}}_{AT}(k) = \frac{1}{k} \sum_{i=1}^k h(p_i),$$

Seqstep: (Barber & Candès, 2015) AT with $h(x) = CI(x > 1 - 1/C)$;
ForwardStop: (G'Sell et al., 2015) AT with $h(x) = -\log(1-x)$.

Adaptive Seqstep and FDR Control

Adaptive Seqstep (AS): s is pre-fixed and
$$\widehat{\mathrm{FDP}}_{AS}(k;s,\lambda) = \frac{s}{1-\lambda} \cdot \frac{\sum_{i=1}^k I(p_i > \lambda) + 1}{\sum_{i=1}^k I(p_i \le s) \vee 1};$$

Motivation: Similar to Storey's correction of BH procedure. Notice that
$$|\mathcal{S}(s,k)| \approx |\mathcal{H}_0| \cdot s \triangleq ns \cdot \pi_0,$$
where $\pi_0 = |\mathcal{H}_0|/n$ is the fraction of null hypotheses. Thus,
$$\widehat{\mathrm{FDP}}_{BH}(s) \approx \frac{1}{\pi_0} \cdot \mathrm{FDP}(s),$$
is too conservative when π_0 is small. By contrast,
$$\widehat{\mathrm{FDP}}_{SBH}(s;\lambda) = \frac{s}{\sum_{i=1}^n I(p_i \le s) \vee 1} \cdot \frac{\sum_{i=1}^n I(p_i > \lambda) + 1}{1-\lambda} \approx \frac{|\mathcal{H}_0| \cdot s}{\sum_{i=1}^n I(p_i \le s) \vee 1}.$$
On the other hand, notice that
$$\widehat{\mathrm{FDP}}_{SBH}(k;s,\lambda) = \frac{s}{\sum_{i=1}^k I(p_i \le s) \vee 1} \cdot \frac{\sum_{i=1}^k I(p_i > s) + 1}{1-s},$$
the term in red is not an accurate estimate of π_0 when s is small. It can be improved by replacing s by a larger number λ, which gives $\widehat{\mathrm{FDP}}_{AS}(k;s,\lambda)$.

FDR Control in Finite Samples:

Theorem 1. *Assume that*

- $\{p_i : i \in \mathcal{H}_0\}$ *are independent of* $\{p_i : i \notin \mathcal{H}_0\}$;
- $\{p_i : i \in \mathcal{H}_0\}$ *are i.i.d. with distribution function* $F_0 \succeq U[0,1]$.

Then AS controls FDR at level q.

References

Barber, R. F., & Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, *43*(5), 2055–2085.

Benjamini, Y., & Hochberg, Y. (1997). Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics*, *24*(3), 407–418.

G'Sell, M. G., Wager, S., Chouldechova, A., & Tibshirani, R. (2015). Sequential selection procedures and false discovery rate control. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Li, A., & Barber, R. F. (2015). Accumulation tests for fdr control in ordered hypothesis testing. *arXiv preprint arXiv:1505.07352*.

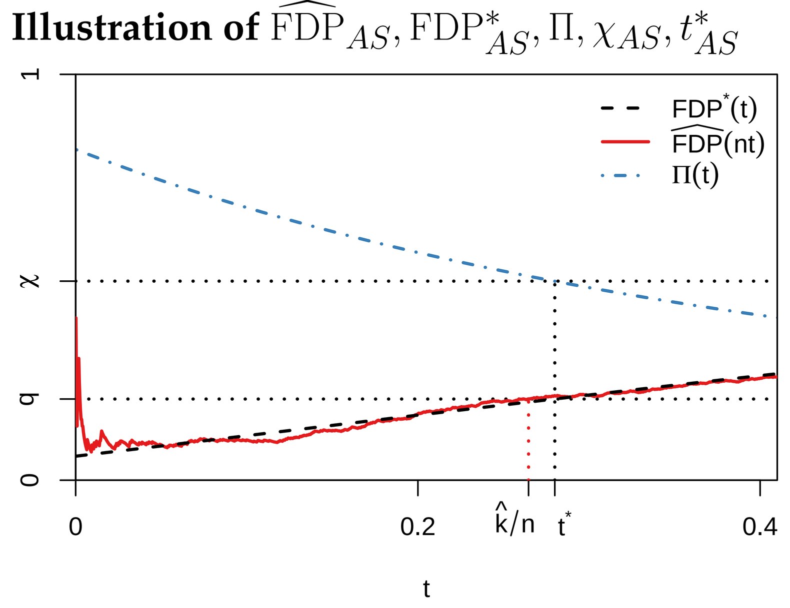
Storey, J. D., Taylor, J. E., & Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *66*(1), 187–205.

VCT Model and Asymptotic Power

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 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 a VCT model, the *Cumulative non-null fraction* is defined as
$$\Pi(t) = \frac{1}{t}\int_0^t \pi(s)ds \approx \frac{|\{i \le nt : H_i \text{ is non-null}\}|}{nt}.$$

Illustration of $\widehat{\mathrm{FDP}}_{AS}, \mathrm{FDP}^*_{AS}, \Pi, \chi_{AS}, t^*_{AS}$



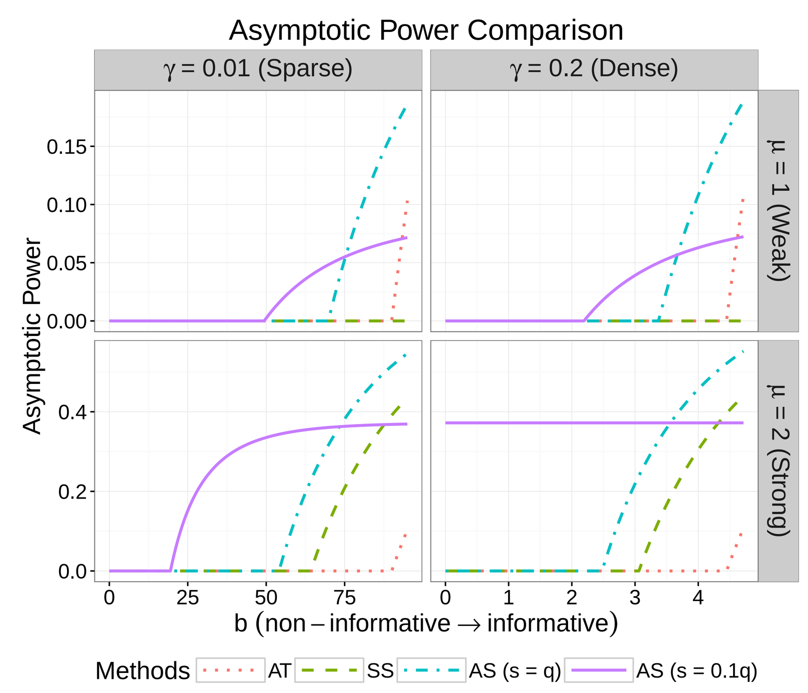
Heuristics: Under a VCT model,
$$\widehat{\mathrm{FDP}}_{AS}(\lfloor nt \rfloor; s, \lambda) \approx \frac{s}{1-\lambda} \cdot \frac{\sum_{i=1}^{\lfloor nt \rfloor} \mathbb{E}I(p_i > \lambda)}{\sum_{i=1}^{\lfloor nt \rfloor} \mathbb{E}I(p_i \le s)}$$
$$\approx \frac{s}{1-\lambda} \cdot \frac{(1-\Pi(t))(1-\lambda) + \Pi(t)(1 - F_1(\lambda))}{(1-\Pi(t))s + \Pi(t)F_1(s)}$$
$$\triangleq \mathrm{FDP}^*_{AS}(t).$$
Denote $\hat{k}_{AS}$ by $\max\{k : \widehat{\mathrm{FDP}}_{AS}(k;s,\lambda) \le q\}$, then
$$\hat{k}_{AS}/n \approx t^*_{AS} \triangleq \max\{t : \mathrm{FDP}^*_{AS}(t) \le q\}.$$
Note that $\mathrm{FDP}^*_{AS}(t)$ depends on t through $\Pi(t)$,
$$t^*_{AS} = \max\{t : \Pi(t) \ge \chi_{AS}\}$$
$$\chi_{AS} = \frac{1-q}{1 - \frac{1-F_1(\lambda)}{1-\lambda} + q\left(\frac{F_1(s)}{s} - 1\right)}$$

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 $\hat{k}_{AS}/n \xrightarrow{a.s.} t^*_{AS}$ *and*
$$\mathrm{Pow}_{AS} \xrightarrow{a.s.} F_1(s) \cdot \frac{t^*_{AS}\Pi(t^*_{AS})}{\Pi(1)} = F_1(s) \cdot \frac{\int_0^{t^*_{AS}} \pi(u)du}{\int_0^1 \pi(u)du},$$

Power Comparison: AS versus SS and AT



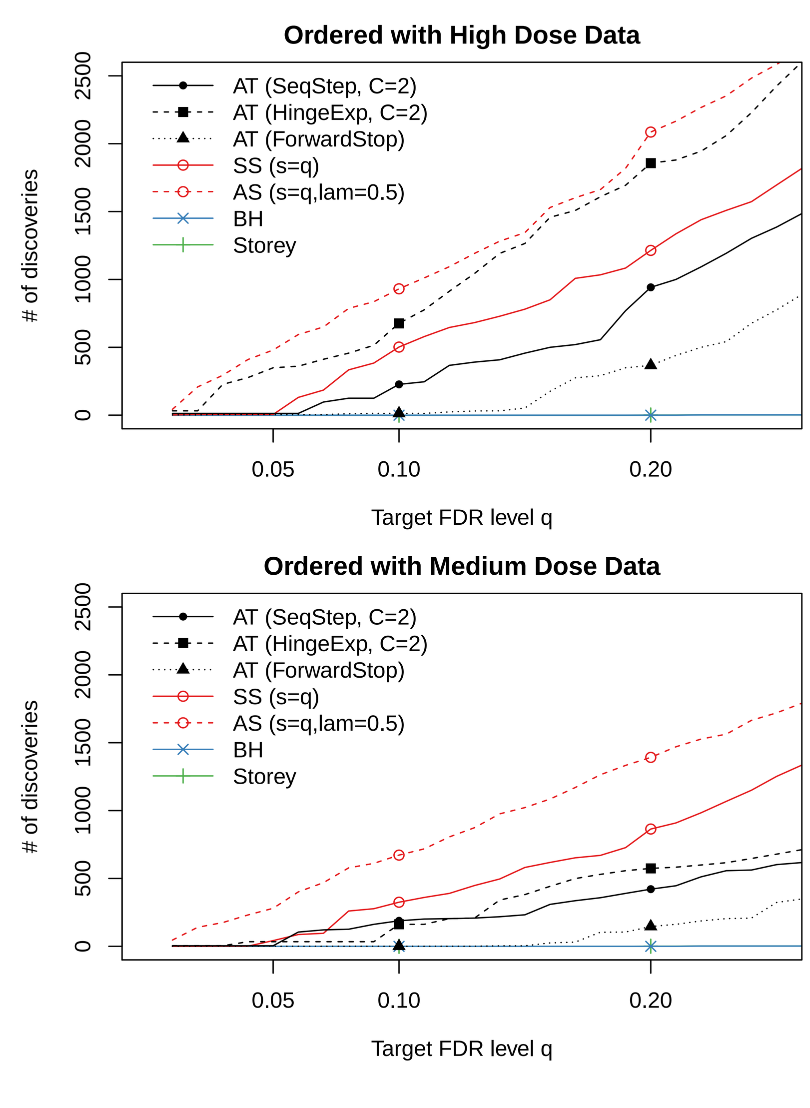
Settings:

- F_1 is the c.d.f. of $\Phi(z)$, a p-value derived from a one-sided z-test, where $z \sim N(\mu, 1)$;
- $\pi(t) \propto \gamma e^{-bt}$ with $\Pi(1) = \int_0^1 \pi(t)dt = \gamma$;
- μ: signal strength; γ: sparsity; b: quality of ordering.

Conclusions:

- AS is always more powerful than SS asymptotically;
- AT is asymptotically powerless unless $\Pi(0) = \pi(0) \ge \frac{1-q}{1-f_1(1)}$. Even when $f_1(1) = 0$, $\pi(0)$ is required to be at least $1-q$;
- AS is more robust to the ordering by setting a small s. If $f_1(0) = \infty$ as in many cases, AS can never be asymptotically powerless if s is chosen appropriately;
- AS is also more robust to weak and sparse signals than SS and AT.

Real Data Example: GEOquery Data



- GEOquery data(Li & Barber, 2015) consists of gene expression measurements in response to estrogen in breast cancer cells;
- Consists of $n = 22283$ genes and five groups (four treatment group with different dosage levels and one control group) with 5 trials in each group;
- Test $H_i : F_{0i} = F_{1i}$, where F_{0i} and F_{1i} are the distributions of gene expression of gene i in the control group and the low-dosage group, respectively.

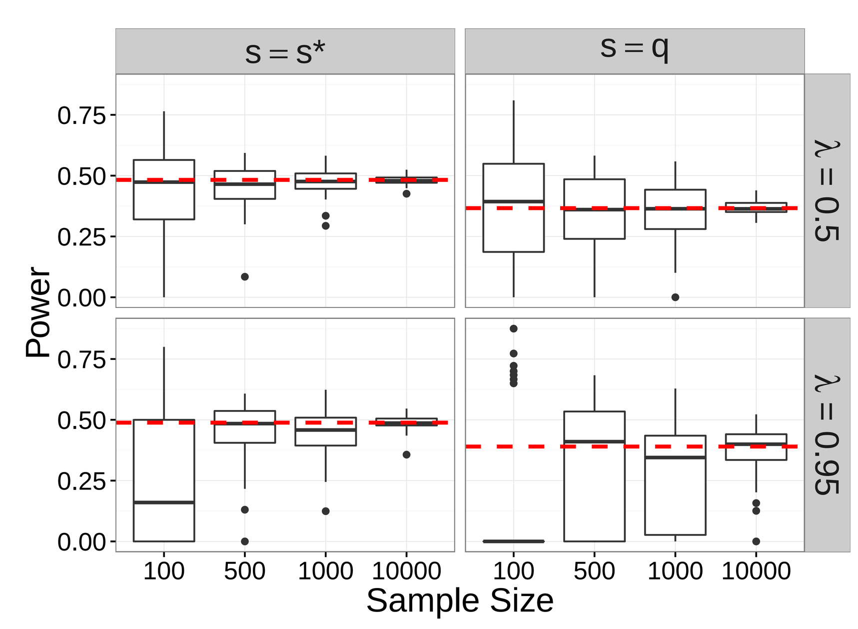
Step 1 Carry out a permutation test by comparing *Highest* with *Low* + *Control* (using t-statistics), and obtain the p-values $\tilde{p}_1, \ldots, \tilde{p}_n$;

Step 2 Sort $H_1, \ldots, H_n$ by $\tilde{p}_1, \ldots, \tilde{p}_n$ and denote the sorted hypotheses by $H_{(1)}, \ldots, H_{(n)}$;

Step 3 Carry out another permutation test by comparing *Low* with *Control* (using t-statistics), and obtain the p-values $p_{(1)}, \ldots, p_{(n)}$;

Step 4 Apply ordered testing procedures on $p_{(1)}, \ldots, p_{(n)}$.

Parameter Selection: s and λ



- We take $s = q$ and $\lambda = 0.5$ as default and the left figure shows the simulated power in finite samples (with $q = 0.1, \mu = 2, \gamma = \Pi(1) = 0.2, \Pi(0) = 0.75$);
- $\lambda = 0.5$ is a rule of thumb and it is much more stable than a large λ, as suggested by theory;
- The choice of s depends on the quality of ordering. Unless the ordering is very bad (either $\Pi(0) \approx 0$ or $\Pi(0) \approx \Pi(1)$), $s = q$ gives a reasonable performance.
- We could try a grid of values for s, e.g. $\{q, 0.5q, 0.25q, \ldots\}$ to maximize the number of rejections. We will explore the validity of processes of this type in future researches.