p-filter : Multilayer FDR control for grouped hypotheses

Aaditya Ramdas* (UC Berkeley), with Rina Foygel Barber* (U Chicago)

Introduction

Central Question:
When testing n different null hypotheses simultaneously, how do we determine which effects are significant? **and** take prior structural knowledge into account while doing this?

When a null hypothesis is rejected, we say *a discovery has been made.*

False Discovery Rate (FDR)

Unknown set of true nulls : $\mathcal{H}^0 \subseteq [n]$.
Declared set of rejected nulls (discoveries) : $\widehat{S} \subseteq [n]$.

- False discovery proportion:

$$\text{FDP} = \frac{\#\text{ false discoveries}}{\text{total }\#\text{ discoveries}} = \frac{|\mathcal{H}^0 \cap \widehat{S}|}{|\widehat{S}|}$$

- False discovery rate $\text{FDR} = \mathbb{E}\,[\text{FDP}]$.

  Aim: Make (many) discoveries with the guarantee that the FDR is smaller than pre-specified level α.

Benjamini-Hochberg (BH)

Let $P := \{P_1, ..., P_n\}$ denote our list of p-values.

Benjamini-Hochberg'95 (BH) procedure: Reject all P_i smaller than a data-dependent threshold $t_{BH} = t(P) \in [0,1]$.

- Suppose we declare as a discovery all p-values below threshold t,

$$\text{FDP}(t) = \frac{|\mathcal{H}^0 \cap \widehat{S}|}{|\widehat{S}|} \approx \frac{t \cdot |\mathcal{H}^0|}{\#\{i : P_i \le t\}} \le \frac{t \cdot n}{\#\{i : P_i \le t\}} = \widehat{\text{FDP}}(t)$$

- $t_{BH} :=$ max t with $\widehat{\text{FDP}}(t) \le \alpha$
  Rephrase: find largest j such that $P_{(j)} \le \alpha j/n$, reject $P_{(1)}, ..., P_{(j)}$.
- Guaranteed to control FDR at level α
  if p-values are independent or positively dependent (PRDS)

Simes test for the global null

Global Null GH_0: test if P is entirely null.

Simes'86 (Improved Bonferroni): we reject GH_0 if

$$\exists j \;:\; P_{(j)} \le \frac{\alpha j}{n} \quad \text{iff} \quad \min_{1 \le k \le n} \frac{P_{(k)} \cdot n}{k} \le \alpha$$

Closely related to BH: Simes rejects GH_0 iff P passes BH(α).

Incorporating Structure

- If the hypotheses have a natural clustered / hierarchical structure, how can we take this into account?
- You may want to **group** together hypotheses that are likely to be null together or be non-null together.
- In spatio-temporal applications, it might be natural to group hypotheses by space or time or space-time blocks. "Discovery at time/location x makes discoveries around x more likely".
- In genetics, certain genes/proteins might be known to act together, or have similar molecular structure.

Goal (in English)

- Given n hypotheses with p-values $P := \{P_1, \ldots, P_n\}$
  Eg: Imagine they are placed in a $r \times c$ grid, $n = rc$.
- Given M partitions (disjoint subsets of P, whose union is P)
  Partition 1 could be the set of all singletons,
  Partition 2 could be the set of all rows, and
  Partition 3 could be the set of all columns.
- Goal: select set $\widehat{S} \subseteq [n]$ such that FDR is bounded simultaneously for partition $1, 2, \ldots, M$.
  Few falsely discovered singletons,
  Few falsely discovered rows,
  Few falsely discovered columns.

Goal (in Math)

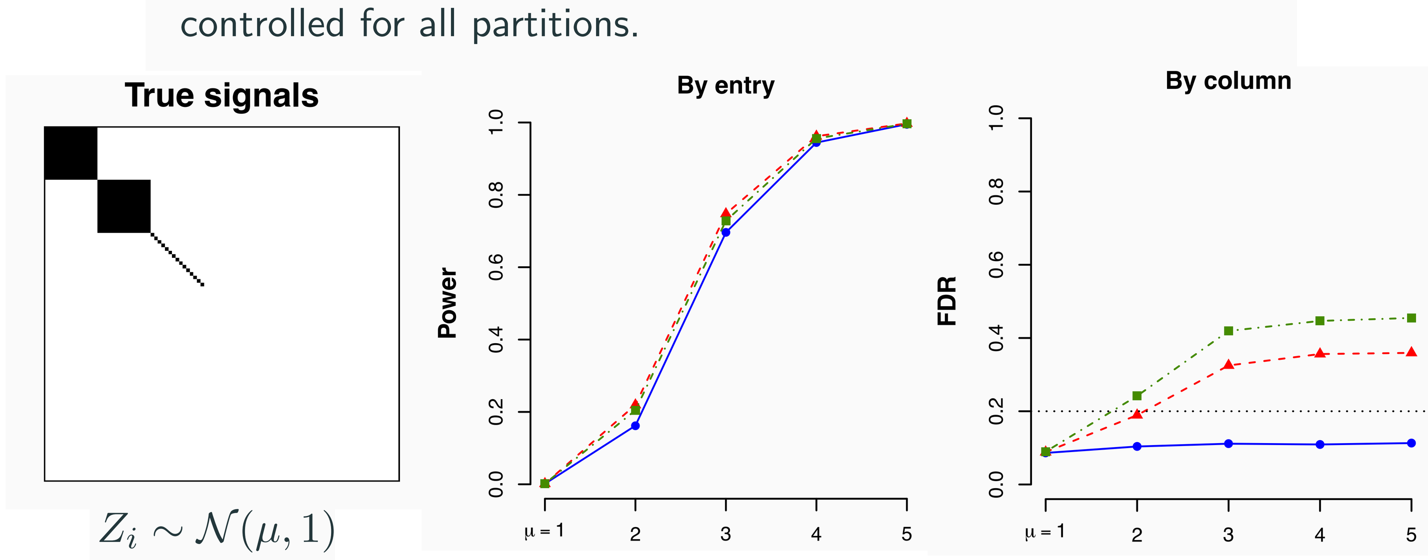
Input partitions:

- mth Partition of $[n] = A_1^m \cup \cdots \cup A_{G_m}^m$
- Null groups $\mathcal{H}_m^0 = \{g : A_g^m \subseteq \mathcal{H}^0\}$

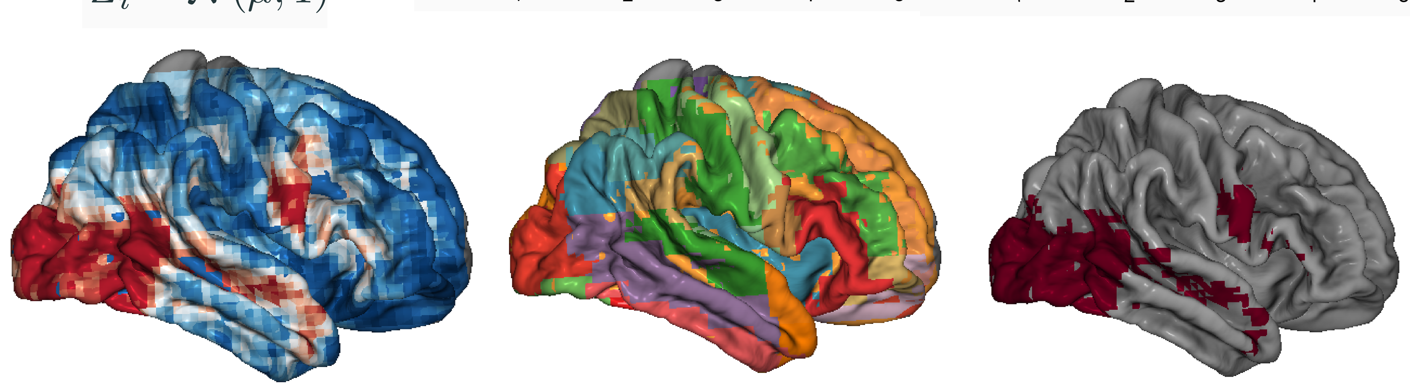
Output discoveries: $\widehat{S} \subseteq [n]$

- Selected groups $\widehat{S}_m = \{g : A_g^m \cap \widehat{S} \ne \varnothing\}$
- FDR control: $\mathbb{E}\left[\frac{|\mathcal{H}_m^0 \cap \widehat{S}_m|}{|\widehat{S}_m|}\right] \le \alpha_m$

p-Filter: will discover $\widehat{S} \subseteq [n]$ such that FDR is simultaneously controlled for all partitions.



$Z_i \sim N(\mu, 1)$


Input Groups Output

The p-filter algorithm

Intuition from the one-partition case

Single partition of G groups: Simes + threshold

- Summarize each group by its Simes p-value. Let

$$P^* = \{\text{Simes}(P^1), \text{Simes}(P^2), ..., \text{Simes}(P^G)\}$$

- Reject all groups with Simes p-value smaller than $t_{BH}(P^*, \alpha)$.

Claim: This procedure controls group-FDR. Why?

Fact: Simes(P^g) is a p-value! (if $P^g \subseteq \mathcal{H}^0$, Simes($P^g$) $\sim U[0,1]$)

Conservative under PRDS.

p-filter : Generalization to multiple partitions

Input: n p-values, M partitions, M FDR levels

Let $t_1 = \alpha_1, \ldots, t_M = \alpha_M$. Repeat $m = 1, \ldots, M$, until no change:

- For the mth partition, Simes+thresholding
  — Calculate Simes p-values $P^m := \{P_1^m, \ldots, P_G^m\}$
  — Reject all groups whose $P_g^m \le t_m$.
- $\widehat{S} := \{P_i :$ in every partition, P_i's group was selected$\}$, intersect
  Let $\widehat{S}_m$ be the discovered groups in partition m, induced by $\widehat{S}$.
- Estimate FDP's for each partition: correction

$$\widehat{\text{FDP}}_m = \frac{t_m \cdot G_m}{|\widehat{S}_m|} \begin{array}{l} \leftarrow \text{approx. }\#\text{ false discoveries} \\ \leftarrow \#\text{ discoveries} \end{array}$$

  If $\widehat{\text{FDP}}_m > \alpha_m$, reduce t_m until $\widehat{\text{FDP}}_m$ is $\le \alpha_m$ (discrete search)

Note: Simes and BH are special cases when $M = 1$.

Assumptions and Guarantees

Let $\widehat{\mathcal{T}}$ be the set of legal thresholds $(t_1, ..., t_M)$, i.e. s.t. $\widehat{\text{FDP}}_m \le \alpha_m$

Conservative null p-value assumption: for each $i \in \mathcal{H}^0$,

$$\frac{\mathbb{P}\{P_i \le t\}}{t} \text{ is an increasing function of } t$$

PRDS assumption: for each $i \in \mathcal{H}^0$,

$$\mathbb{P}\{P \in \text{increasing set} \mid P_i = t\} \text{ is an increasing function of } t$$

Theorem 2

p-Filter finds $\max(\widehat{\mathcal{T}})$, and it controls FDR simultaneously $\forall m$:

$$\text{FDR for partition } m = \mathbb{E}\left[\frac{|\mathcal{H}_m^0 \cap \widehat{S}_m|}{|\widehat{S}_m|}\right] \le \alpha_m \cdot \frac{|\mathcal{H}_m^0|}{G_m} \quad \forall\, m.$$

Furthermore, it halts in $G_1 + G_2 + ... + G_M + 1$ outer loops.