Leveraging prior information and group structure for false discovery rate control

Rina Foygel Barber

Dept. of Statistics, University of Chicago

http://www.stat.uchicago.edu/~rina/

Multiple comparisons & FDR control

When testing n different questions simultaneously,
how to determine which effects are significant?

- 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:

$$\text{FDR} = \mathbb{E}\left[\text{FDP}\right]$$

Multiple comparisons & FDR control

Benjamini-Hochberg (BH) procedure (1995):
set a data-dependent threshold for rejecting p-values,
to adapt to the amount of signal present in the data

- If we reject all p-values below a fixed threshold t,

$$\mathrm{FDP}(t) \approx \frac{t \cdot |\mathcal{H}^0|}{\#\{i : P_i \leq t\}} = \widehat{\mathrm{FDP}}(t)$$

- Choose adaptive threshold: max t with $\widehat{\mathrm{FDP}}(t) \leq \alpha$
- Guaranteed to control FDR at level α
  if p-values are independent or positively dependent (PRDS)

Benjamini & Hochberg 1995; Benjamini & Yekutieli 2001

Multiple comparisons & FDR control

How can we incorporate additional information into the FDR control problem?

- If some of the hypotheses are more likely to contain true signals, should we give them priority?
- If the hypotheses have a grouped / clustered / hierarchical structure, how can we take this into account?

Outline

1. Accumulation tests: testing a ranked list of hypotheses
    - Joint work with Ang Li

2. The p-filter: FDR control across groups
    - Joint work with Aaditya Ramdas

Ordered hypothesis testing

Setting:

a multiple comparisons problem with a pre-defined ordering.

p-values: $P_1,\ P_2,\ P_3,\ \ldots\ ,\ P_N$

$\longleftrightarrow$

select first / most likely to be a true signal

select last / least likely to be a true signal

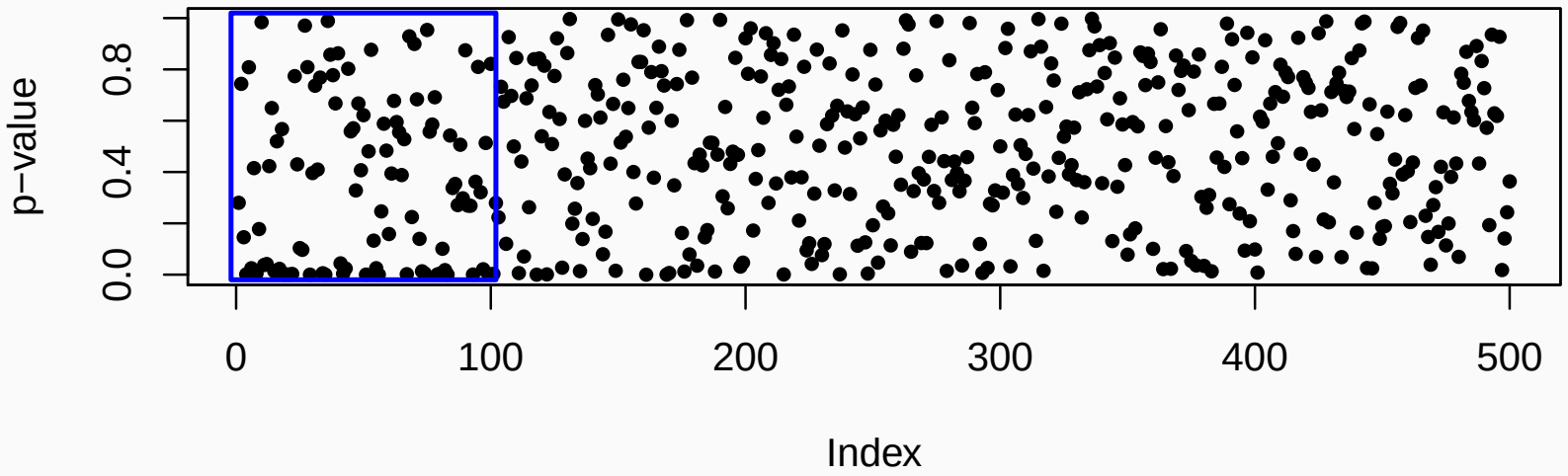
Ordered hypothesis testing

Where does the ordering come from?

- Data from related experiments: e.g. gene expression levels in a different tissue, with a related drug compound, etc
- Regression setting:
  For sequential procedures (forward selection, LASSO, etc), recent work produces valid p-values for variables in the order that they are selected:
  - Post-selection inference
    (Fithian, Taylor, Tibshirani, Tibshirani, Lockart, ...)
  - Knockoff method (Barber & Candès): one-bit p-values

Ordered hypothesis testing

SeqStep method (Barber & Candès):



Want to estimate # nulls among first k p-values

$\rightsquigarrow$ count how many p-values are > 0.5

Ordered hypothesis testing

Null p-values are equally likely to be above 0.5 or below 0.5

$$\Downarrow$$

$\approx$ half the null p-values, among the first k p-values, will be > 0.5

$$\Downarrow$$

$$\text{FDP}(k) \approx \frac{2 \cdot (\#\text{ p-values} > 0.5\text{, among first } k)}{k} = \widehat{\text{FDP}}_{\text{SeqStep}}(k)$$

Then stop at $\widehat{k}_{\text{SeqStep}} =$ last time that $\widehat{\text{FDP}}_{\text{SeqStep}}(k) \leq \alpha$

Ordered hypothesis testing

A related method — ForwardStop (G'Sell et al 2013):

To estimate FDP among the first k p-values,

$$\widehat{\mathsf{FDP}}_{\mathsf{ForwardStop}}(k) = \frac{\sum_{i=1}^{k} \log\left(\frac{1}{1-P_i}\right)}{k}$$

Then stop at $\widehat{k}_{\mathsf{ForwardStop}}$ = last time that $\widehat{\mathsf{FDP}}_{\mathsf{ForwardStop}}(k) \leq \alpha$

Accumulation tests

Accumulation test: reject the first $\widehat{k}_{\mathsf{h}}$ p-values, where

$$\widehat{k}_{\mathsf{h}} = \max\left\{k : \widehat{\mathsf{FDP}}_{\mathsf{h}}(k) \leq \alpha\right\},$$

for

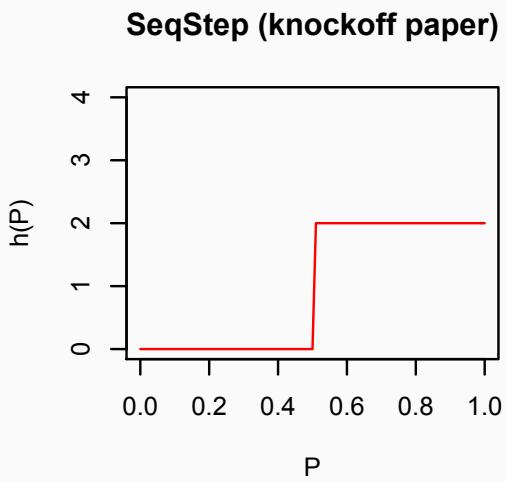
$$\mathsf{FDP}(k) = \frac{\#\text{ nulls among } \{1,\ldots,k\}}{k} \approx \underbrace{\frac{\mathsf{h}(P_1) + \cdots + \mathsf{h}(P_k)}{k}}_{\text{Estimated FDP}=\widehat{\mathsf{FDP}}_{\mathsf{h}}(k)}$$

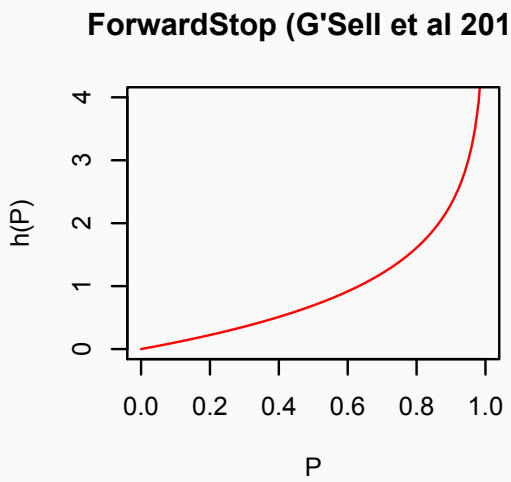
h is a function $[0,1] \to [0,\infty]$ with

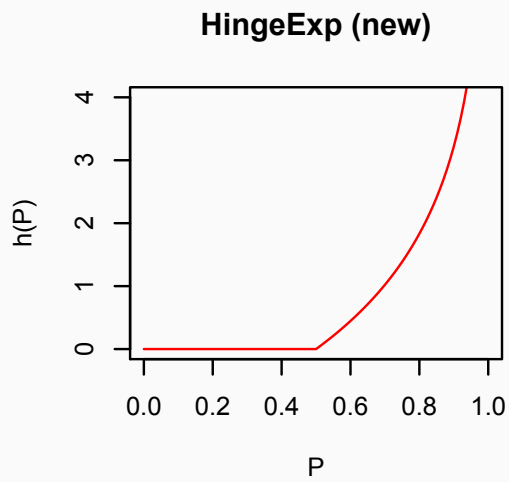
- $\int_{t=0}^{1} \mathsf{h}(t)\, \mathrm{d}t = 1 \Rightarrow \mathbb{E}\left[\mathsf{h}(P_i)\right] = 1$ for the nulls
- $\mathsf{h} \approx 0$ near $0 \Rightarrow \mathbb{E}\left[\mathsf{h}(P_i)\right] \approx 0$ for strong signals

Accumulation tests

Existing & new choices for the function h:







Accumulation tests

Theorem

If h is an accumulation function bounded by C, then

$$\mathbb{E}\left[\frac{\#\text{ nulls among } \{1,\ldots,k\}}{k + C/\alpha}\right] \leq \alpha.$$

(See paper for a guarantee when h is unbounded.)

Advantage over BH & other multiple testing corrections:

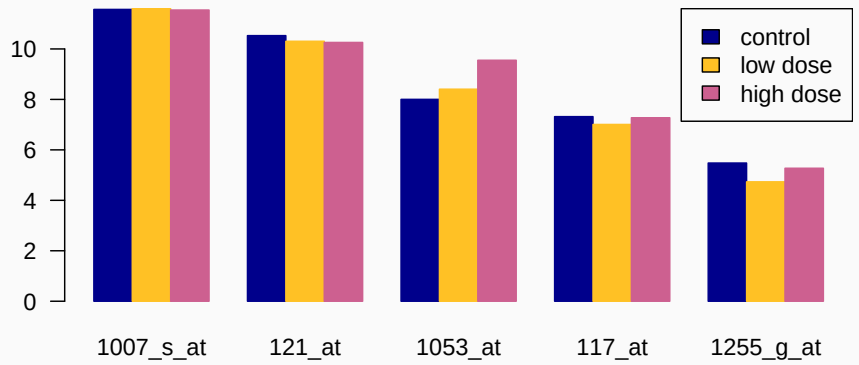
No dependence on $n = \#$ of hypotheses tested

Gene dosage data

- Expression levels for $n = 22283$ genes measured at different dosage levels:

  Sample size: 5 control (zero dose), 5 low dose, 5 high dose

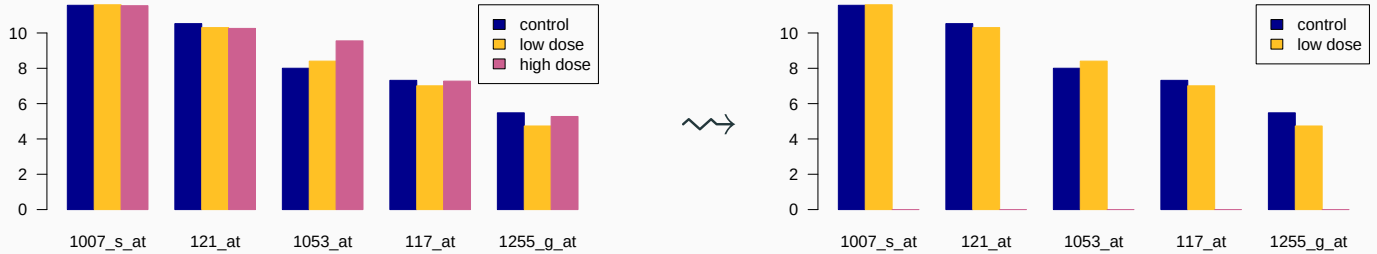
- Can we identify genes with *differential expression* at the lowest dosage level?



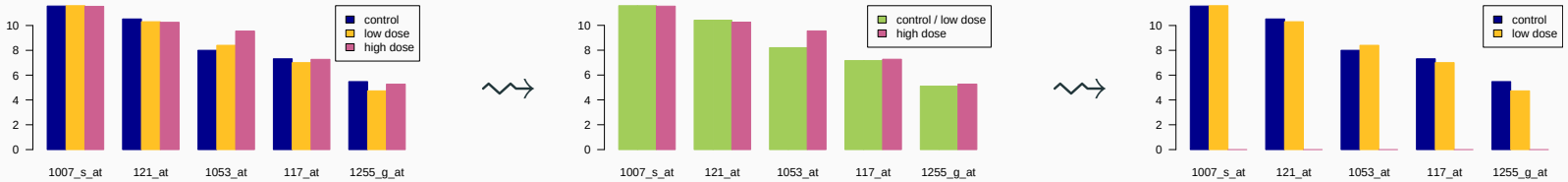
Data from Coser et al 2003 via R Geoquery package (data set GDS2324)

Gene dosage data

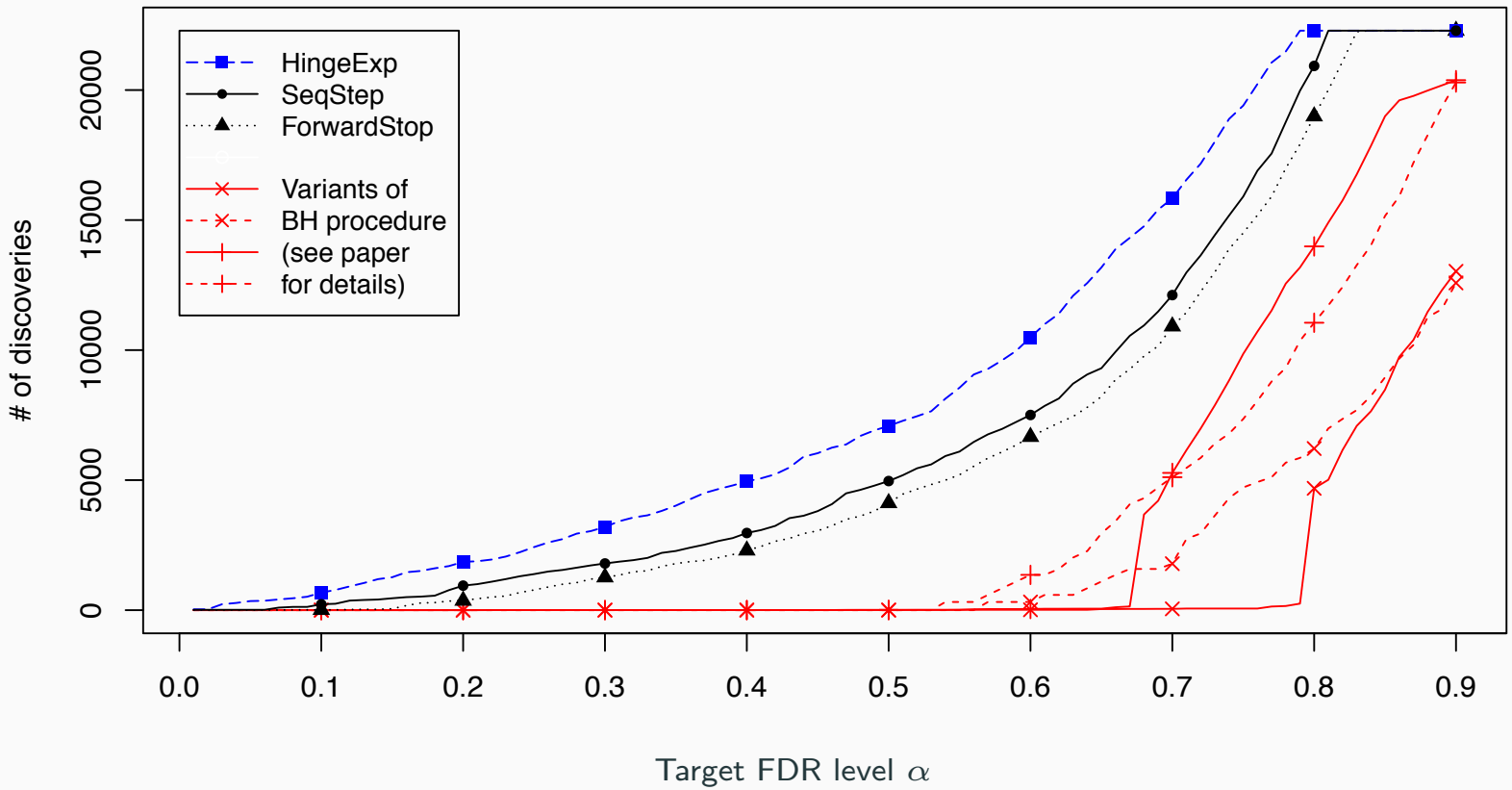
- Standard approach w/o high dose data:
  1. Two-sample test for control vs. low dose
  2. Then correct for multiple comparisons (BH & variants)



- Our approach:
  1. Rank genes by comparing high dose vs. control/low dose
  2. Run accumulation test to compare control vs. low dose



Gene dosage data

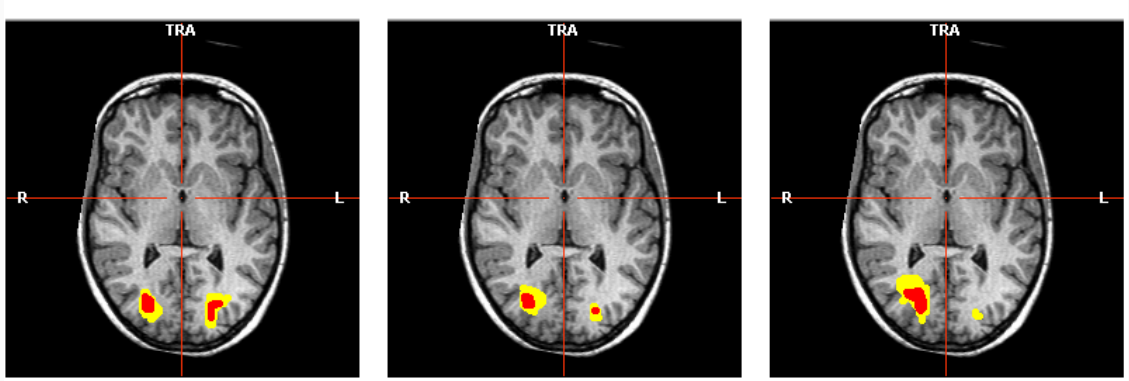


Outline

1. Accumulation tests: testing a ranked list of hypotheses
    - Joint work with Ang Li

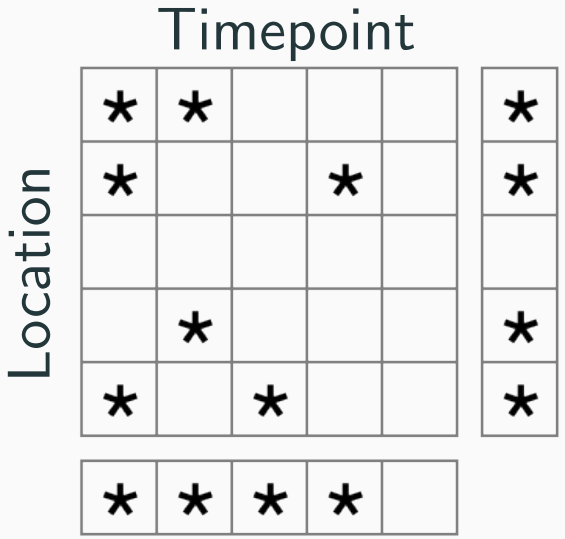
2. The p-filter: FDR control across groups
    - Joint work with Aaditya Ramdas

Structured set of hypotheses



Time 1 Time 2 Time 3

Hypotheses:



Structured set of hypotheses

- n hypotheses with p-values $P_1, \ldots, P_n$
- M "layers" $=$ partitions of the hypotheses
  (e.g. entries, rows, columns in our array)
- Goal: select set $\widehat{S}$ of discoveries such that FDR is bounded simultaneously for layer $1, 2, \ldots, M$.

Structured set of hypotheses

Where do the groupings come from?

- Natural structure in the set of hypotheses
- Regression setting:
  Clusters / correlations within the features;
  Hierarchical structure (e.g. due to interaction terms)

Multilayer FDR

How to define FDR for the mth layer?

- Partition $[n] = A_1^m \cup \cdots \cup A_{G_m}^m$
- Nulls $\mathcal{H}_m^0 = \{g : A_g^m \subseteq \mathcal{H}^0\}$
- Selected set $\widehat{S}_m = \{g : A_g^m \cap \widehat{S} \neq \varnothing\}$

- FDR control: $\mathbb{E}\left[\frac{|\mathcal{H}_m^0 \cap \widehat{S}_m|}{|\widehat{S}_m|}\right] \leq \alpha_m$?

Multilayer FDR

A naive method:

- For the mth layer,
  - — Calculate Simes p-values

$$P_1^m, \ldots, P_{G_m}^m$$

(P_g^m tests whether group A_g^m is all nulls)

  - — Run BH with threshold α_m on this list
    $\rightsquigarrow$ reject groups with $P_g^m \leq$ adaptive threshold t_m
- Problem: results might not be consistent across the M layers

Multilayer FDR

$$\alpha_{\text{indiv}} = 0.1$$
$$\alpha_{\text{group}} = 0.2$$

Multilayer FDR

$$\alpha_{\text{indiv}} = 0.1$$
$$\alpha_{\text{group}} = 0.2$$

Multilayer FDR

$$\alpha_{\text{indiv}} = 0.1$$
$$\alpha_{\text{group}} = 0.2$$

Multilayer FDR

The p-filter:

- $\widehat{S}(t_1, \ldots, t_m) =$ set of discoveries at thresholds $t_1, \ldots, t_M$:

  P_i is selected, if it belongs to a selected group in all M layers

- Now estimate FDP's for $\widehat{S}(t_1, \ldots, t_m)$, in each layer:

$$\widehat{\text{FDP}}_m = \frac{t_m \cdot G_m}{|\widehat{S}_m(t_1, \ldots, t_m)|} \begin{array}{l} \leftarrow \text{approx. } \# \text{ false discoveries} \\ \leftarrow \# \text{ discoveries} \end{array}$$

- Choose t_m's adaptively: maximize t_m's s.t. $\widehat{\text{FDP}}_m \leq \alpha_m \; \forall \; m$.

Theoretical results

Theorem 1

This maximum is well-defined and can be computed efficiently.

Algorithm:

- Initialize thresholds $t_1 = \alpha_1, \ldots, t_M = \alpha_M$
- Cycle through layers $1, \ldots, M$:
  - Check if $\widehat{\text{FDP}}_m$ is low enough:

$$\frac{t_m \cdot G_m}{|\widehat{S}_m(t_1, \ldots, t_M)|} \leq \alpha_m \ ?$$

  - If not, reduce t_m until $\widehat{\text{FDP}}_m$ is $\leq \alpha_m$
- ... until there are no more changes.

Theoretical results

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

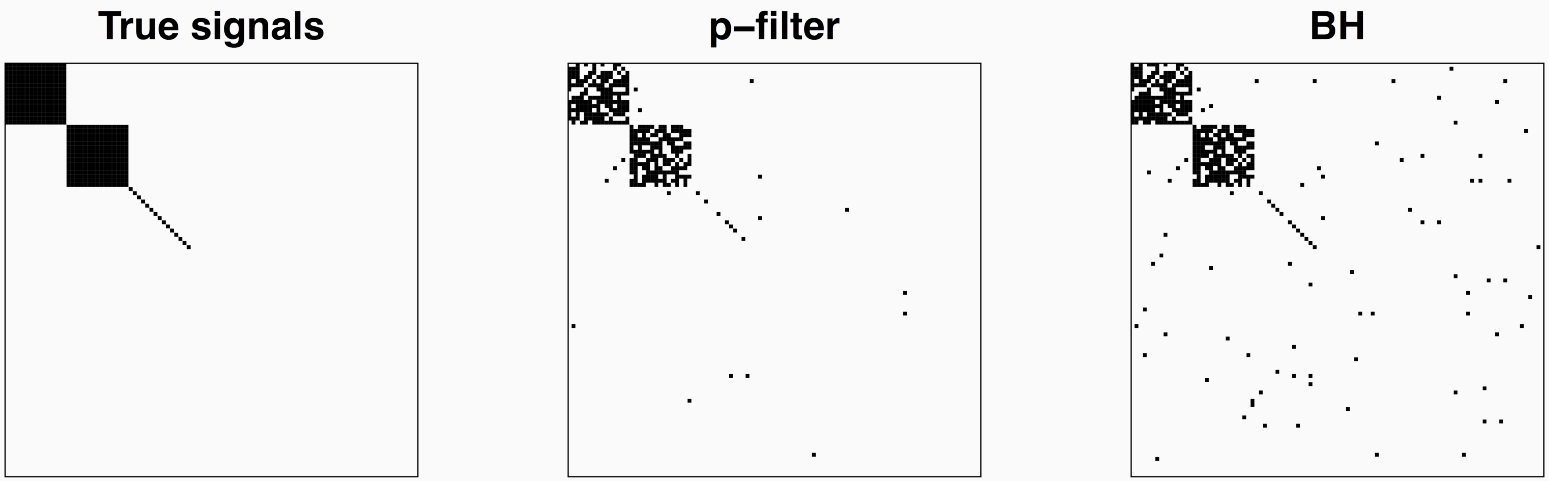
This procedure controls FDR for all layers:

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

Key lemma: If $f(P)$ is a decreasing function of P, then

$$\mathbb{E}\left[\frac{\mathbb{1}\{P_i \leq f(P)\}}{f(P)}\right] \leq 1.$$

Simulation results



Layers: entries; rows; columns.

Target FDR: $\alpha_{\text{entries}} = \alpha_{\text{rows}} = \alpha_{\text{columns}} = 0.2$

Future work

- Connection between ordered testing & online testing?
- Create data-adaptive clusters?
- An ordered testing approach for grouped hypotheses?

Thank you!

Accumulation tests (w/ Ang Li):
`http://www.stat.uchicago.edu/~rina/accumulationtests.html`

Multi-FDR (w/ Aaditya Ramdas):
`http://www.stat.uchicago.edu/~rina/pfilter.html`