Controlling False Discovery Rate Privately

Weijie Su

University of Pennsylvania

NIPS, Barcelona, December 9, 2016

Joint work with Cynthia Dwork and Li Zhang

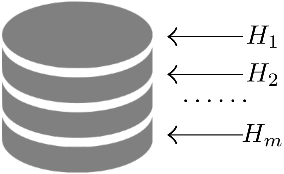
Living in the Big Data world

Privacy loss

Privacy loss

- Second Netflix challenge canceled
- AOL search data leak
- Inference presence of individual from minor allele frequencies [Homer et al '08]

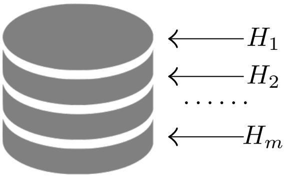
This talk: privacy-preserving multiple testing



A hypothesis H could be

- Is the SNP associated with diabetes?
- Does the drug affect autism?

This talk: privacy-preserving multiple testing



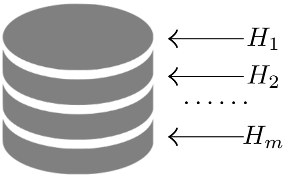
A hypothesis H could be

- Is the SNP associated with diabetes?
- Does the drug affect autism?

Goal

- Preserve privacy
- Control false discovery rate (FDR)

This talk: privacy-preserving multiple testing



A hypothesis H could be

- Is the SNP associated with diabetes?
- Does the drug affect autism?

Goal

- Preserve privacy
- Control false discovery rate (FDR)

Application

- Genome-wide association studies
- A/B testing

Outline

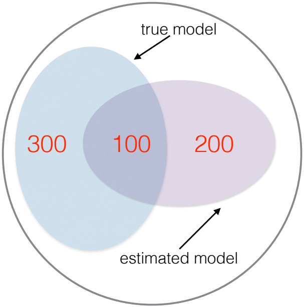
1. Warm-ups
   - FDR and BHq procedure
   - Differential privacy
2. Introducing PrivateBHq
3. Proof of FDR control

Two types of errors

	Not reject	Reject	Total
Null is true	True negative	False positive	m_0
Null is false	False negative	True positive	m_1
Total			m

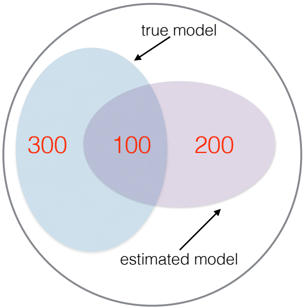
False discovery rate (FDR)

$$\text{FDR} := \mathbb{E}\left[\frac{\#\text{false discoveries}}{\#\text{discoveries}}\right]$$



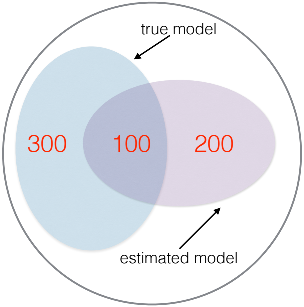
False discovery rate (FDR)

$$\text{FDR} := \mathbb{E}\left[\frac{\#\text{false discoveries}}{\#\text{discoveries}}\right] = \frac{200}{100+200}$$



False discovery rate (FDR)

$$\text{FDR} := \mathbb{E}\left[\frac{\#\text{false discoveries}}{\#\text{discoveries}}\right] = \frac{200}{100+200}$$



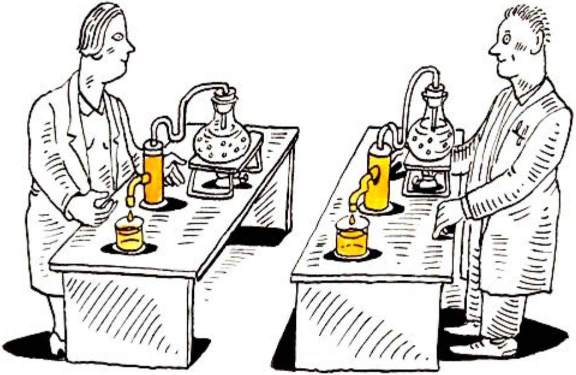
- Wish FDR $\leq q$ (often $q = 0.05, 0.1$)
- Proposed by Benjamini and Hochberg '95
- 35,490 citations as of yesterday

Why FDR?

Why FDR?

FDR addresses reproducibility

FDR addresses reproducibility



How to control FDR?

p-values of hypotheses

p-value

The probability of finding the observed, or more extreme, results when the null hypothesis of a study question is true

- Uniform in $[0, 1]$ (or stochastically larger) under true null

p-values of hypotheses

p-value

The probability of finding the observed, or more extreme, results when the null hypothesis of a study question is true

- Uniform in $[0, 1]$ (or stochastically larger) under true null

H_0: the drug does not lower blood pressure

p-values of hypotheses

p-value

The probability of finding the observed, or more extreme, results when the null hypothesis of a study question is true

- Uniform in $[0, 1]$ (or stochastically larger) under true null

H_0: the drug does not lower blood pressure

- If $p = 0.5$, no evidence

p-values of hypotheses

p-value

The probability of finding the observed, or more extreme, results when the null hypothesis of a study question is true

- Uniform in $[0, 1]$ (or stochastically larger) under true null

H_0: the drug does not lower blood pressure

- If $p = 0.5$, no evidence
- If $p = 0.01$, there is evidence!

p-values of hypotheses

p-value

The probability of finding the observed, or more extreme, results when the null hypothesis of a study question is true

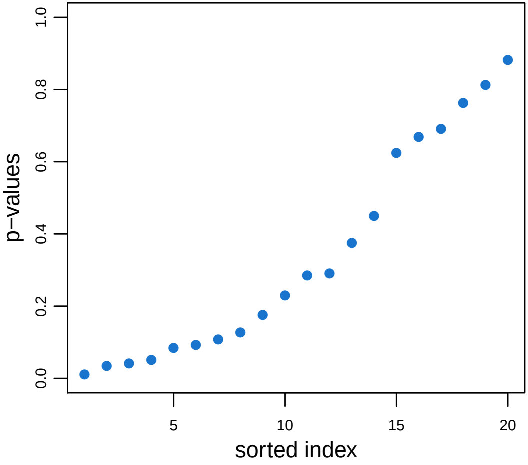
- Uniform in $[0, 1]$ (or stochastically larger) under true null

H_0: the drug does not lower blood pressure

- If $p = 0.5$, no evidence
- If $p = 0.01$, there is evidence?

Benjamini–Hochberg procedure (BHq)

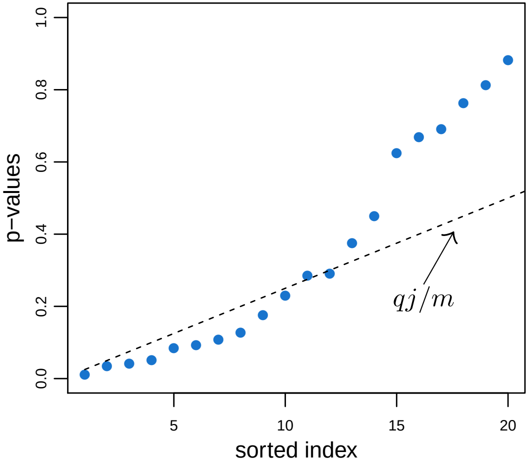
Let $p_1, p_2, \ldots, p_m$ be p-values of m hypotheses



- Sort $p_{(1)} \leq \cdots \leq p_{(m)}$

Benjamini–Hochberg procedure (BHq)

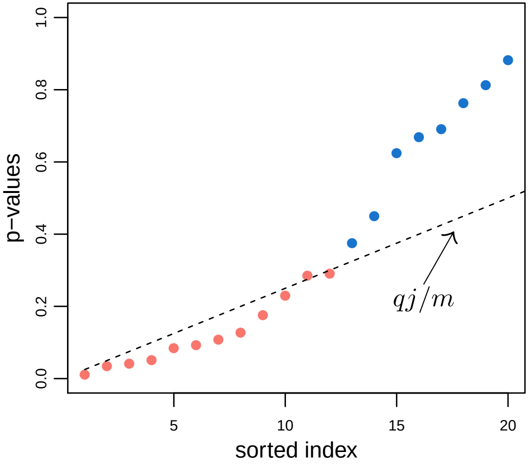
Let $p_1, p_2, \ldots, p_m$ be p-values of m hypotheses



- Sort $p_{(1)} \leq \cdots \leq p_{(m)}$
- Draw rank-dependent threshold qj/m

Benjamini–Hochberg procedure (BHq)

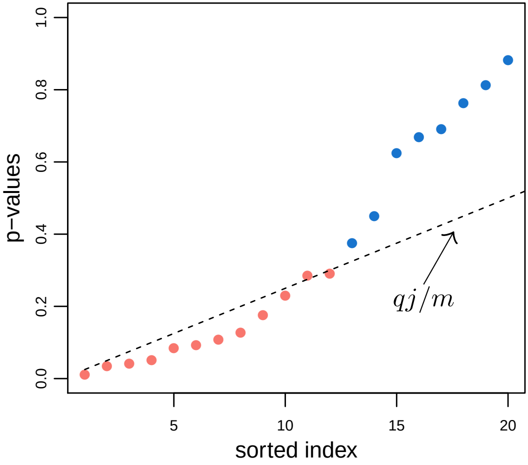
Let $p_1, p_2, \ldots, p_m$ be p-values of m hypotheses



- Sort $p_{(1)} \leq \cdots \leq p_{(m)}$
- Draw rank-dependent threshold qj/m
- Reject hypotheses below cutoffs

Benjamini–Hochberg procedure (BHq)

Let $p_1, p_2, \ldots, p_m$ be p-values of m hypotheses

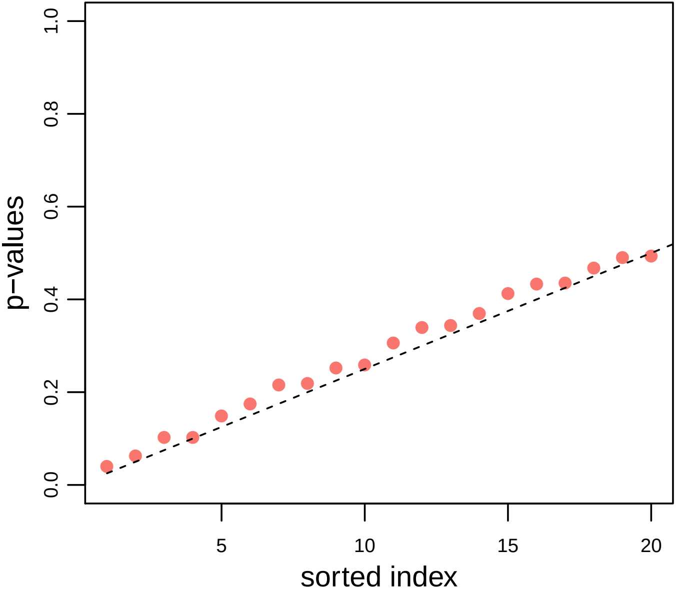


- Sort $p_{(1)} \leq \cdots \leq p_{(m)}$
- Draw rank-dependent threshold qj/m
- Reject hypotheses below cutoffs
- Under independence $\mathrm{FDR} \leq q$

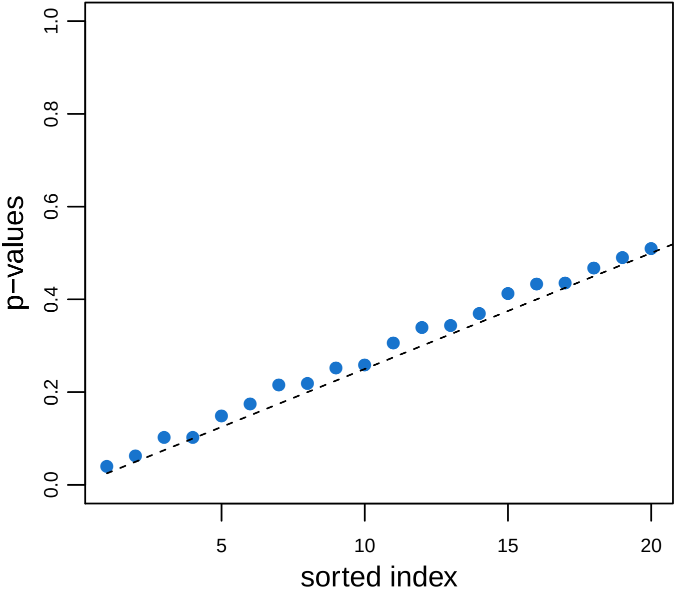
What is privacy?

- My response had little impact on released results
- Any adversary cannot learn much information about me based on released results
- Anonymity may not work
- Is the Benjamini-Hochberg procedure (BH) privacy-preserving?

BHq is sensitive to perturbations



BHq is sensitive to perturbations



A concrete foundation of privacy

Let $\mathcal{M}$ be a (random) data-releasing mechanism

Differential privacy (Dwork, McSherry, Nissim, Smith '06)

$\mathcal{M}$ is called (ϵ, δ)-differentially private if for all databases D and D' differing with one individual, and all $S \subset \mathrm{Range}(\mathcal{M})$,

$$\mathbb{P}(\mathcal{M}(D) \in S) \le \mathrm{e}^{\epsilon}\, \mathbb{P}(\mathcal{M}(D') \in S) + \delta$$

A concrete foundation of privacy

Let $\mathcal{M}$ be a (random) data-releasing mechanism

Differential privacy (Dwork, McSherry, Nissim, Smith '06)

$\mathcal{M}$ is called (ϵ, δ)-differentially private if for all databases D and D' differing with one individual, and all $S \subset \mathrm{Range}(\mathcal{M})$,

$$\mathbb{P}(\mathcal{M}(D) \in S) \le \mathrm{e}^{\epsilon}\, \mathbb{P}(\mathcal{M}(D') \in S) + \delta$$

- Probability space is over the randomness of $\mathcal{M}$

A concrete foundation of privacy

Let $\mathcal{M}$ be a (random) data-releasing mechanism

Differential privacy (Dwork, McSherry, Nissim, Smith '06)

$\mathcal{M}$ is called (ϵ, δ)-differentially private if for all databases D and D' differing with one individual, and all $S \subset \mathrm{Range}(\mathcal{M})$,

$$\mathbb{P}(\mathcal{M}(D) \in S) \leq \mathrm{e}^{\epsilon}\, \mathbb{P}(\mathcal{M}(D') \in S) + \delta$$

- Probability space is over the randomness of $\mathcal{M}$
- If $\delta = 0$ (pure privacy),

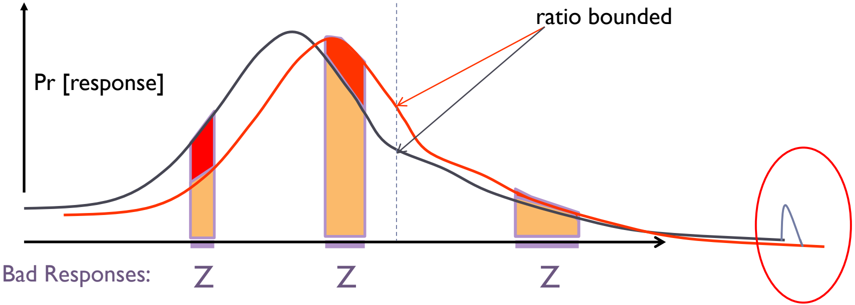
$$\mathrm{e}^{-\epsilon} \leq \frac{\mathbb{P}(\mathcal{M}(D) \in S)}{\mathbb{P}(\mathcal{M}(D') \in S)} \leq \mathrm{e}^{\epsilon}$$

A concrete foundation of privacy

Differential privacy (Dwork, McSherry, Nissim, Smith '06)

For all neighboring databases D and D',

$$\mathbb{P}(\mathcal{M}(D) \in S) \leq e^{\epsilon}\, \mathbb{P}(\mathcal{M}(D') \in S) + \delta$$



An addition to a vast literature

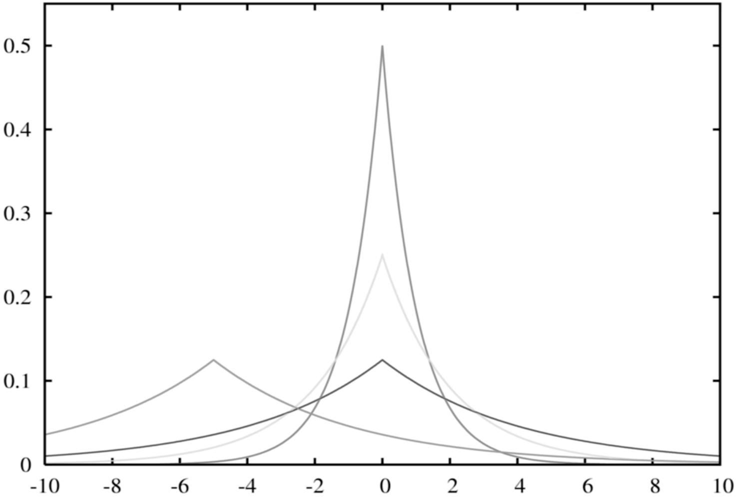
- Counts, linear queries, histograms, contingency tables
- Location and spread
- Dimension reduction (PCA, SVD), clustering
- Support vector machine
- Sparse regression, Lasso, logistic regression
- Gradient descent
- Boosting, multiplicative weights
- Combinatorial optimization, mechanism design
- Kalman filtering
- Statistical queries learning model, PAC learning

An addition to a vast literature

- Counts, linear queries, histograms, contingency tables
- Location and spread
- Dimension reduction (PCA, SVD), clustering
- Support vector machine
- Sparse regression, Lasso, logistic regression
- Gradient descent
- Boosting, multiplicative weights
- Combinatorial optimization, mechanism design
- Kalman filtering
- Statistical queries learning model, PAC learning
- FDR control

Laplace noise

$\mathrm{Lap}(b)$ has density $\exp(-|x|/b)/2b$



Achieving $(\epsilon, 0)$-differential privacy: a vignette

How many members of the House of Representatives voted for Trump?

- Sensitivity is 1
- Add symmetric noise $\mathrm{Lap}(\frac{1}{\epsilon})$ to the counts

Achieving $(\epsilon, 0)$-differential privacy: a vignette

How many members of the House of Representatives voted for Trump?

- Sensitivity is 1
- Add symmetric noise $\mathrm{Lap}(\frac{1}{\epsilon})$ to the counts

How many albums of Taylor Swift are bought in total by people in this room?

- Sensitivity is 5
- Add symmetric noise $\mathrm{Lap}(\frac{5}{\epsilon})$ to the counts

Outline

1. Warm-ups
   - FDR and BHq procedure
   - Differential privacy
2. Introducing PrivateBHq
3. Proof of FDR control

Sensitivity of p-values

- Additive noise can kill signals when p-values are small
- Solution: take logarithm of p-values

Sensitivity of p-values

- Additive noise can kill signals when p-values are small
- Solution: take logarithm of p-values

Databases D and D' are adjacent.

Definition

Tuples $(p_1(D), \ldots, p_m(D))$ and $(p_1(D'), \ldots, p_m(D'))$ are called (η, ν)-multiplicatively sensitive if, for all i,

- either $p_i(D), p_i(D') < \nu$, or
- $\mathrm{e}^{-\eta} p_i(D) \leq p_i'(D') \leq \mathrm{e}^{\eta} p_i(D)$

- $\pi_i = \log \max\{p_i(D), \nu\}$ has sensitivity η

Examples of multiplicatively Sensitive p-values

iid $\xi_1, \dots, \xi_n$, taking 1 with probability of α and 0 otherwise. T is the sum. To test $H_0 : \alpha \leq \frac{1}{2}$ against $H_1 : \alpha > \frac{1}{2}$:

$$p(D) = \sum_{i=T}^{n} \frac{1}{2^n} \binom{n}{i}.$$

Assume $m = n^C$. Then we can take $\nu = m^{-2}$ and $\eta = n^{-\frac{1}{2}+o(1)}$

Building blocks of PrivateBHq

Private Min

a.k.a. Report Noisy Min

Algorithm 1: Private Min

Input: $\pi_1, \cdots, \pi_m$

1: **for** $i = 1$ to m **do**

2: set $\pi_i^{\otimes} = \pi_i + g_i$ where g_i is i.i.d. $\mathrm{Lap}(\eta\sqrt{10k\log(1/\delta)}/\epsilon)$

3: **end for**

4: return $(i^\star = \operatorname{argmin} \pi_i^{\otimes}, \ \pi^\star = \pi_{i^\star} + g)$ where $g \sim \mathrm{Lap}(\eta\sqrt{10k\log(1/\delta)}/\epsilon)$

- Private Min is $(2\epsilon/\sqrt{10k\log(1/\delta)}, 0)$-differentially privacy
- Less noise [Raskhodnikova and Smith '16]

Pre-selection by peeling

Algorithm 2: Peeling

Input: $\pi_1, \cdots, \pi_m$ and k

for $j = 1$ to k **do**
    run Private Min
    remove selected $\pi_{i^\star}$
end for
report k selected pairs $(i,\ \tilde{\pi}_i)$

Pre-selection by peeling

Algorithm 2: Peeling

Input: $\pi_1, \cdots, \pi_m$ and k

1: **for** $j = 1$ to k **do**
2: run Private Min
3: remove selected $\pi_{i^\star}$
4: **end for**
5: report k selected pairs $(i,\ \tilde{\pi}_i)$

Lemma

$\mathrm{peeling}(k)$ *is* (ϵ, δ)*-differentially private*

- A simple application of Advanced Composition Theorem [Dwork, Rothblum, and Vadhan '10]

Finally, PrivateBHq

Algorithm 3: PrivateBHq

Input: (η, ν)-sensitive p-values $p_1, \cdots, p_m,\ k \geq 1$ and ϵ, δ
Output: a set of up to k rejected hypotheses

1: set $\pi_i = \log(\max\{p_i, \nu\})$
2: apply $\mathrm{peeling}(k)$ to $\pi_1, \ldots, \pi_m$
3: apply BHq to $y_1, \ldots, y_k$ with cutoffs $\alpha_j = \log(qj/m + \nu) + \eta\Delta$, where $\Delta = (1 + o(1))\sqrt{k \log(1/\delta)} \log m/\epsilon$

Finally, PrivateBHq

Algorithm 3: PrivateBHq

Input: (η, ν)-sensitive p-values $p_1, \cdots, p_m$, $k \geq 1$ and ϵ, δ
Output: a set of up to k rejected hypotheses

1: set $\pi_i = \log(\max\{p_i, \nu\})$
2: apply $\mathrm{peeling}(k)$ to $\pi_1, \ldots, \pi_m$
3: apply BHq to $y_1, \ldots, y_k$ with cutoffs $\alpha_j = \log(qj/m + \nu) + \eta\Delta$, where $\Delta = (1 + o(1))\sqrt{k \log(1/\delta)} \log m/\epsilon$

Theorem (Dwork, S., and Zhang)

The PrivateBHq is (ϵ, δ)-differentially private

Outline

1. Warm-ups
   - FDR and BHq procedure
   - Differential privacy
2. Introducing PrivateBHq
3. Proof of FDR control

New techniques required

- Smallest p-values may not be selected
- Difficult to specify the joint distribution of selected p-values
- Destroys crucial properties for proving FDR control

Compliant procedures

Definition

A procedure is called compliant with $\{q_j\}_{j=1}^m$ if all the R rejected p-values are below q_R

Compliant procedures

Definition

A procedure is called compliant with $\{q_j\}_{j=1}^{m}$ if all the R rejected p-values are below q_R

- Self-consistency condition [Blanchard and Roquain '08]
- Step-up and step-down BHqs are $\{jq/m\}$-compliant
- So are the generalized step-up-step-down procedures [Tamhane, Liu, and Dunnett '98; Sarkar 02']
- How about the PrivateBHq?

PrivateBHq is compliant

Lemma

Given (η, ν)-sensitive p-values with $\nu = o(1/m)$, then with probability $1 - o(1)$, the private FDR-controlling algorithm is compliant with $\{jq'/m\}$, where $q' = (1 + o(1))\mathrm{e}^{\eta\Delta} \cdot q$

Compliance + IWS = FDR control

Definition

A set of test statistics are called to satisfy the *independence within a subset* $\mathcal{I}_0$ (IWS on $\mathcal{I}_0$), if the test statistics from $\mathcal{I}_0$ are jointly independent.

Compliance + IWS = FDR control

Definition

A set of test statistics are called to satisfy the *independence within a subset* $\mathcal{I}_0$ (IWS on $\mathcal{I}_0$), if the test statistics from $\mathcal{I}_0$ are jointly independent.

Theorem

Suppose the test statistics satisfies IWS on the subset of true null hypotheses. Then any procedure compliant with the BHq critical values qj/m *obeys*

$$\text{FDR} \le q\log(1/q) + Cq$$
$$\text{FDR}_2 \le Cq$$
$$\text{FDR}_k \le \left(1 + 2/\sqrt{qk}\right) q.$$

- $\text{FDR}_k := \mathbb{E}\left[\frac{V}{R}; V \ge k\right]$
- $C \approx 2.7$

Compliance + IWS = FDR control

Theorem

IWS on the subset of true nulls + compliance with the BHq critical values qj/m *give*

$$\mathsf{FDR} \le q\log(1/q) + Cq$$

$$\mathsf{FDR}_2 \le Cq$$

$$\mathsf{FDR}_k \le \left(1 + 2/\sqrt{qk}\right) q.$$

- Arbitrary correlations between true null and false null test statistics

Compliance + IWS = FDR control

Theorem

IWS on the subset of true nulls + compliance with the BHq critical values qj/m *give*

$$\mathsf{FDR} \leq q\log(1/q) + Cq$$

$$\mathsf{FDR}_2 \leq Cq$$

$$\mathsf{FDR}_k \leq \left(1 + 2/\sqrt{qk}\right) q.$$

- Arbitrary correlations between true null and false null test statistics
- Can be even adversarial!

Compliance + IWS = FDR control

Theorem

IWS on the subset of true nulls + compliance with the BHq critical values qj/m *give*

$$\begin{aligned} \mathsf{FDR} &\le q\log(1/q) + Cq \\ \mathsf{FDR}_2 &\le Cq \\ \mathsf{FDR}_k &\le \left(1 + 2/\sqrt{qk}\right) q. \end{aligned}$$

- Arbitrary correlations between true null and false null test statistics
- Can be even adversarial!
- Explains partially why BHq is so robust

Compliance + IWS = FDR control

Theorem

IWS on the subset of true nulls + compliance with the BHq critical values qj/m *give*

$$\mathsf{FDR} \le q\log(1/q) + Cq$$
$$\mathsf{FDR}_2 \le Cq$$
$$\mathsf{FDR}_k \le \left(1 + 2/\sqrt{qk}\right) q.$$

- Arbitrary correlations between true null and false null test statistics
- Can be even adversarial!
- Explains partially why BHq is so robust
- If $V \to \infty$ with probability tending to one, then FDR $\le q + o(1)$

Proof Sketch

An upper bound on FDP

Let $p_{i_1}, \ldots, p_{i_R}$ be those rejected, among which $p^0_{(1)} \leq \cdots \leq p^0_{(V)}$ are from *true nulls*.

An upper bound on FDP

Let $p_{i_1}, \ldots, p_{i_R}$ be those rejected, among which $p^0_{(1)} \leq \cdots \leq p^0_{(V)}$ are from *true nulls*. Compliance requires

$$p^0_{(V)} \leq \max_{1 \leq j \leq R} p_{i_j} \leq \alpha_R = qR/m$$

An upper bound on FDP

Let $p_{i_1}, \ldots, p_{i_R}$ be those rejected, among which $p^0_{(1)} \le \cdots \le p^0_{(V)}$ are from *true nulls*. Compliance requires

$$p^0_{(V)} \le \max_{1 \le j \le R} p_{i_j} \le \alpha_R = qR/m$$

Hence

$$\begin{aligned}
& R \ge \lceil m p^0_{(V)}/q \rceil \\
\Rightarrow & \frac{V}{\max\{R, 1\}} \le \frac{V}{\lceil m p^0_{(V)}/q \rceil} \\
\Rightarrow & \mathrm{FDP} \le \max_{2 \le j \le m_0} \frac{j}{\lceil m p^0_{(j)}/q \rceil} + \min\left\{ \frac{1}{\lceil m p^0_{(1)}/q \rceil}, 1 \right\}
\end{aligned}$$

- m_0 is the total number of true nulls

Bounding the two terms

Lemma

- $\mathbb{E} \max_{2 \le j \le m_0} \frac{j}{\lceil m p^0_{(j)} / q \rceil} \le C_1 q$
- $\mathbb{E} \min \left\{ \frac{1}{\lceil m p^0_{(1)} / q \rceil}, 1 \right\} \le q \log \frac{1}{q} + C_2 q$

for some absolute constants C_1 and C_2

Bounding the two terms

Lemma

- $\mathbb{E} \max_{2 \le j \le m_0} \frac{j}{\lceil m p^0_{(j)} / q \rceil} \le C_1 q$
- $\mathbb{E} \min \left\{ \frac{1}{\lceil m p^0_{(1)} / q \rceil}, 1 \right\} \le q \log \frac{1}{q} + C_2 q$

for some absolute constants C_1 and C_2

- Assume $m_0 = m$

Bounding the two terms

Lemma

- $\mathbb{E} \max_{2 \le j \le m_0} \frac{j}{\lceil m p^0_{(j)} / q \rceil} \le C_1 q$
- $\mathbb{E} \min \left\{ \frac{1}{\lceil m p^0_{(1)} / q \rceil}, 1 \right\} \le q \log \frac{1}{q} + C_2 q$

for some absolute constants C_1 and C_2

- Assume $m_0 = m$
- Assume all true null p-values are iid uniform on $[0, 1]$

Bounding the two terms

Lemma

- $\mathbb{E} \max_{2 \le j \le m} \frac{j}{\lceil m U_{(j)}/q \rceil} \le C_1 q$
- $\mathbb{E} \min \left\{ \frac{1}{\lceil m U_{(1)}/q \rceil}, 1 \right\} \le q \log \frac{1}{q} + C_2 q$

for some absolute constants C_1 and C_2

- Assume $m_0 = m$
- Assume all true null p-values are iid uniform on $[0, 1]$
- Let $U_1, U_2, \ldots, U_m$ be iid and uniform on $[0, 1]$

Using Rényi's representation

Wish to prove

$$\mathbb{E} \max_{2 \le j \le m} \frac{j}{\lceil mU_{(j)}/q \rceil} \le C_1 q$$

Using Rényi's representation

Wish to prove

$$\mathbb{E} \max_{2 \le j \le m} \frac{j}{\lceil mU_{(j)}/q \rceil} \le C_1 q$$

Let $\xi_1, \ldots, \xi_{m+1}$ be iid exponential random variables

$$(U_{(1)}, U_{(2)}, \ldots, U_{(m)}) \stackrel{d}{=} \left(\frac{T_1}{T_{m+1}}, \frac{T_2}{T_{m+1}}, \ldots, \frac{T_m}{T_{m+1}} \right)$$

- $T_j = \xi_1 + \cdots + \xi_j$

Using Rényi's representation

Wish to prove

$$\mathbb{E} \max_{2 \le j \le m} \frac{j}{\lceil mU_{(j)}/q \rceil} \le C_1 q$$

Let $\xi_1, \ldots, \xi_{m+1}$ be iid exponential random variables

$$(U_{(1)}, U_{(2)}, \ldots, U_{(m)}) \stackrel{d}{=} \left(\frac{T_1}{T_{m+1}}, \frac{T_2}{T_{m+1}}, \ldots, \frac{T_m}{T_{m+1}} \right)$$

- $T_j = \xi_1 + \cdots + \xi_j$
- $\displaystyle \frac{j}{\lceil mU_{(j)}/q \rceil} \le \frac{qj}{mU_{(j)}} = \frac{q}{m} \cdot \frac{jT_{m+1}}{T_j} \equiv \frac{q}{m} \cdot W_j$
- $W_j \equiv jT_{m+1}/T_j$

W_j is a backward submartingale

Wish to prove

$$\mathbb{E} \max_{2 \leq j \leq m} \frac{W_j}{m} \leq C_1$$

Submartingale definition

$$\mathbb{E}(W_j | T_{j+1}, \ldots, T_{m+1}) \geq W_{j+1}$$

W_j is a backward submartingale

Wish to prove

$$\mathbb{E} \max_{2 \le j \le m} \frac{W_j}{m} \le C_1$$

Submartingale definition

$$\mathbb{E}(W_j | T_{j+1}, \ldots, T_{m+1}) \ge W_{j+1}$$

By martingale theory

$$\mathbb{E} \max_{2 \le j \le m} \frac{W_j}{m} \le (1 - \mathrm{e}^{-1})^{-1} \left[1 + \mathbb{E} \left(\frac{W_2}{m} \log \frac{W_2}{m}; \frac{W_2}{m} \ge 1 \right) \right]$$

W_j is a backward submartingale

Wish to prove

$$\mathbb{E} \max_{2 \le j \le m} \frac{W_j}{m} \le C_1$$

Submartingale definition

$$\mathbb{E}(W_j | T_{j+1}, \ldots, T_{m+1}) \ge W_{j+1}$$

By martingale theory

$$\begin{aligned}
\mathbb{E} \max_{2 \le j \le m} \frac{W_j}{m} &\le (1 - \mathrm{e}^{-1})^{-1} \left[1 + \mathbb{E} \left(\frac{W_2}{m} \log \frac{W_2}{m} ; \frac{W_2}{m} \ge 1 \right) \right] \\
&\le (1 - \mathrm{e}^{-1})^{-1} \left[1 + \mathbb{E} \left(\frac{2}{mU_{(2)}} \log \frac{2}{mU_{(2)}} ; \frac{2}{mU_{(2)}} \ge 1 \right) \right] \\
&\le C_1
\end{aligned}$$

Summary

Take-home message

- FDR addresses reproducibility
- Differential privacy is a rigorous definition
- Privatize BH by adding noise in peeling
- A bonus: Compliance with IWS gives FDR control

Take-home message

- FDR addresses reproducibility
- Differential privacy is a rigorous definition
- Privatize BH by adding noise in peeling
- A bonus: Compliance with IWS gives FDR control

Thank You!