Controlling the False Discovery Rate: Understanding and Extending the Benjamini-Hochberg Method

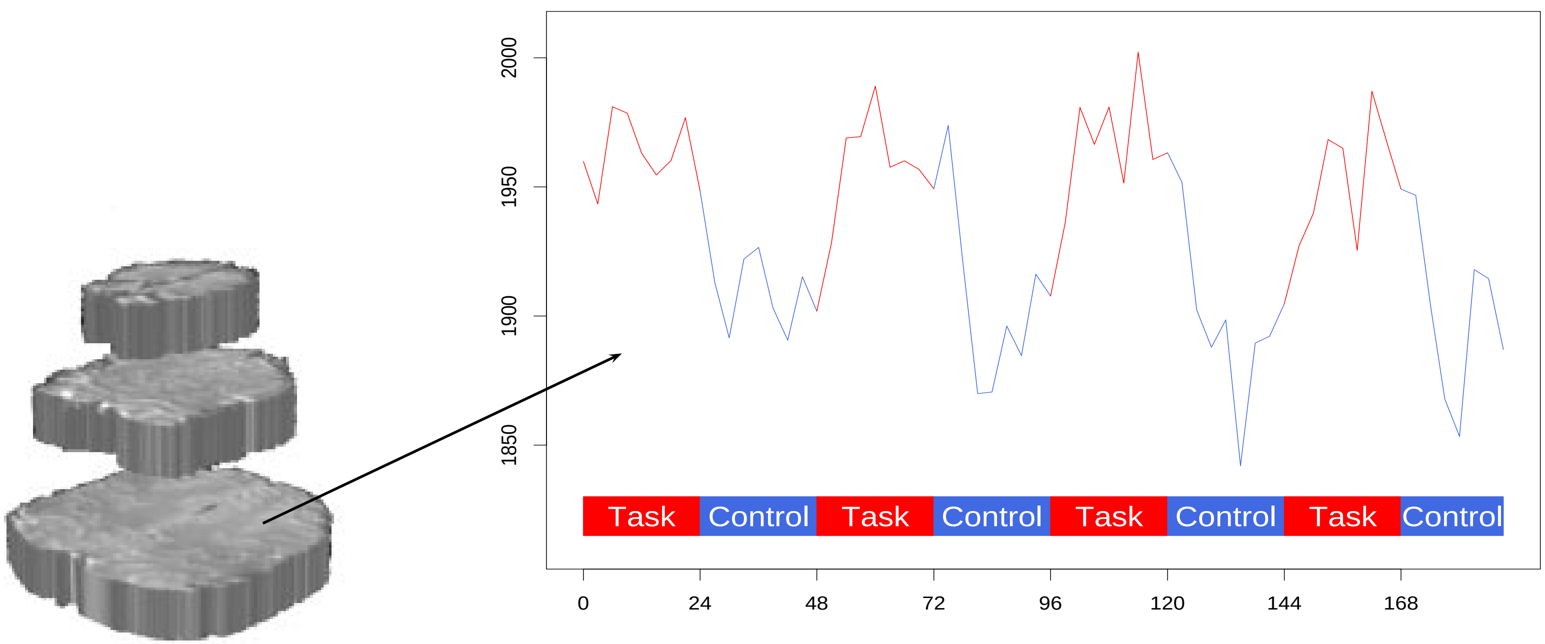
Christopher R. Genovese
Department of Statistics
Carnegie Mellon University

joint work with Larry Wasserman

This work partially supported by NSF Grant SES 9866147.

Motivating Example #1: fMRI

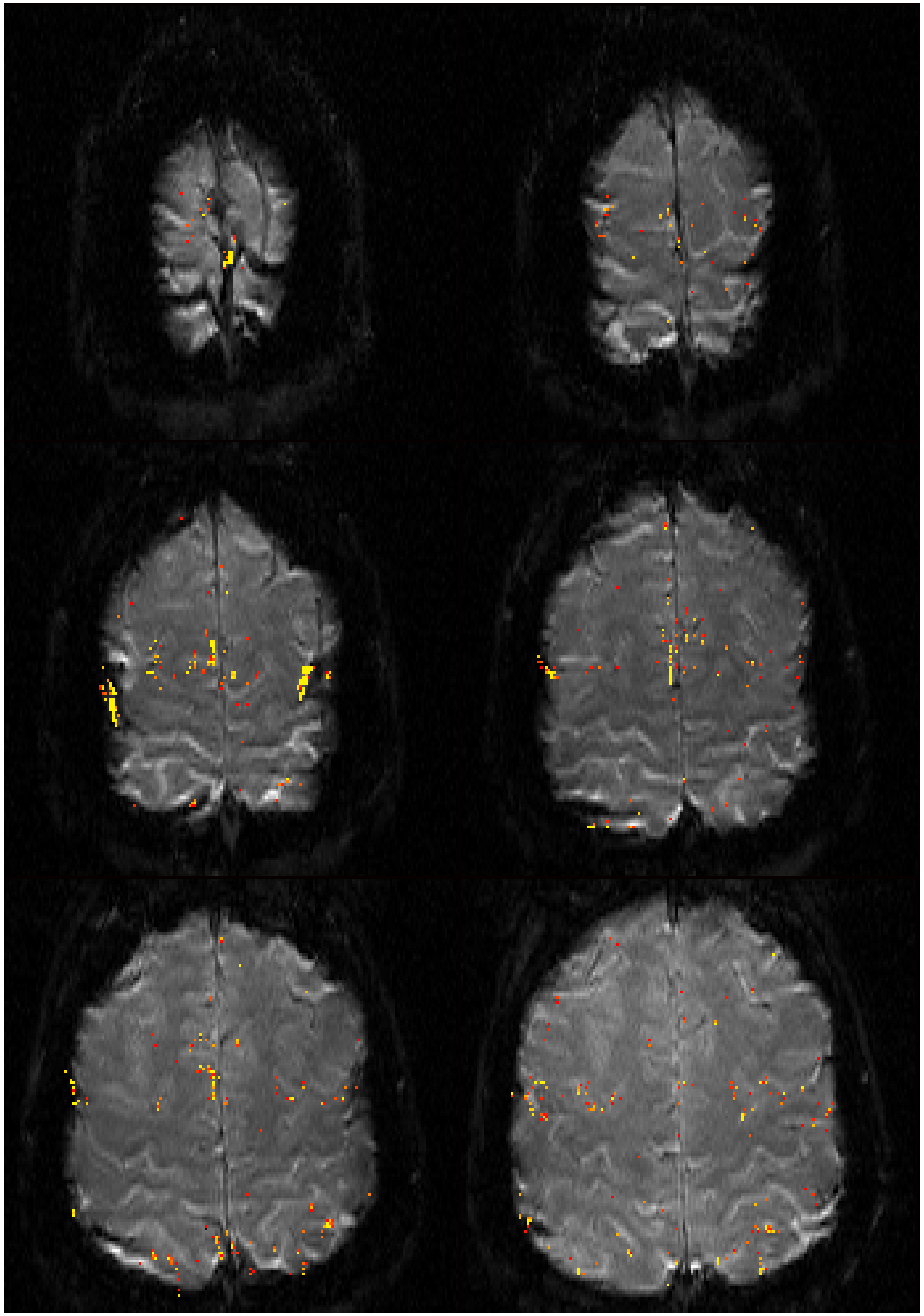
- fMRI Data: Time series of 3-d images acquired while subject performs specified tasks.



- Goal: Characterize task-related signal changes caused (indirectly) by neural activity. [See, for example, Genovese (2000), *JASA* 95, 691.]

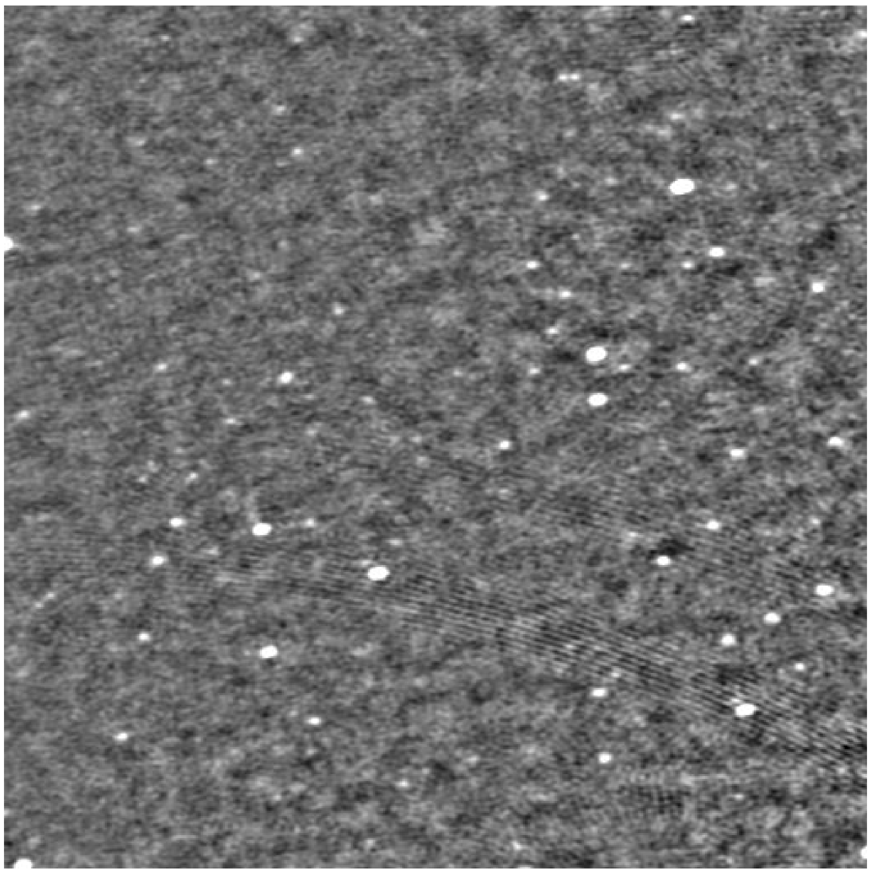
fMRI (cont'd)

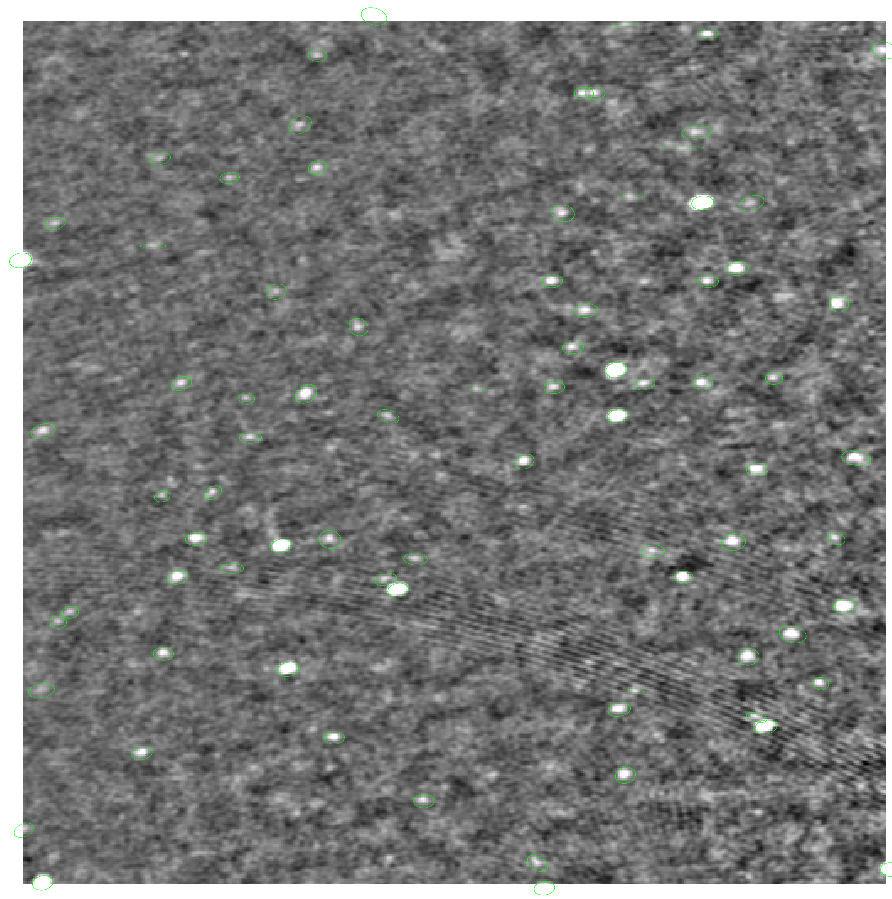
Perform hypothesis tests at many thousands of volume elements to identify loci of activation.



Motivating Example #2: Source Detection

- Interferometric radio telescope observations processed into digital image of the sky in radio frequencies.
- Signal at each pixel is a mixture of source and background signals.





Motivating Example #3: DNA Microarrays

- New technologies allow measurement of gene expression for thousands of genes simultaneously.

Condition 1

Gene	Subject 1	Subject 2	Subject 3	...
1	X_{111}	X_{121}	X_{131}	...
2	X_{211}	X_{221}	X_{231}	...
3	$\vdots$	$\vdots$	$\vdots$	...
4				
5				
6				
$\vdots$				

Condition 2

Gene	Subject 1	Subject 2	Subject 3	...
1	X_{112}	X_{122}	X_{132}	...
2	X_{212}	X_{222}	X_{232}	...
3	$\vdots$	$\vdots$	$\vdots$	...
4				
5				
6				
$\vdots$				

- Goal: Identify genes associated with differences among conditions.
- Typical analysis: hypothesis test at each gene.

Road Map

1. What is the False Discovery Rate?

   Preliminaries

2. Why does the BH method work?

   BH as a plug-in estimator

3. How does the BH method perform?

   Operating Characteristics

4. Can BH be made more powerful?

   Plug-in Procedures

5. What are the implications for inference?

   Confidence Thresholds

6. How does dependence among the p-values affect the results?

   Dealing with Dependence

The Multiple Testing Problem

- Perform m simultaneous hypothesis tests.

  Classify results as follows:

	H_0 Retained	H_0 Rejected	Total
H_0 True	$N_{0\|0}$	$N_{1\|0}$	M_0
H_0 False	$N_{0\|1}$	$N_{1\|1}$	M_1
Total	$m - R$	R	m

  Only R is observed here.

- Assess outcome through combined error measure.
- Traditional methods seek strong control of familywise Type I error.
- Can power be improved while maintaining control over a meaningful measure of error? Enter Benjamini & Hochberg . . .

FDR and the BH Procedure

- Define the *realized* False Discovery Rate (FDR) by

$$\text{FDR} = \begin{cases} \dfrac{N_{1|0}}{R} & \text{if } R > 0, \\ 0, & \text{if } R = 0. \end{cases}$$

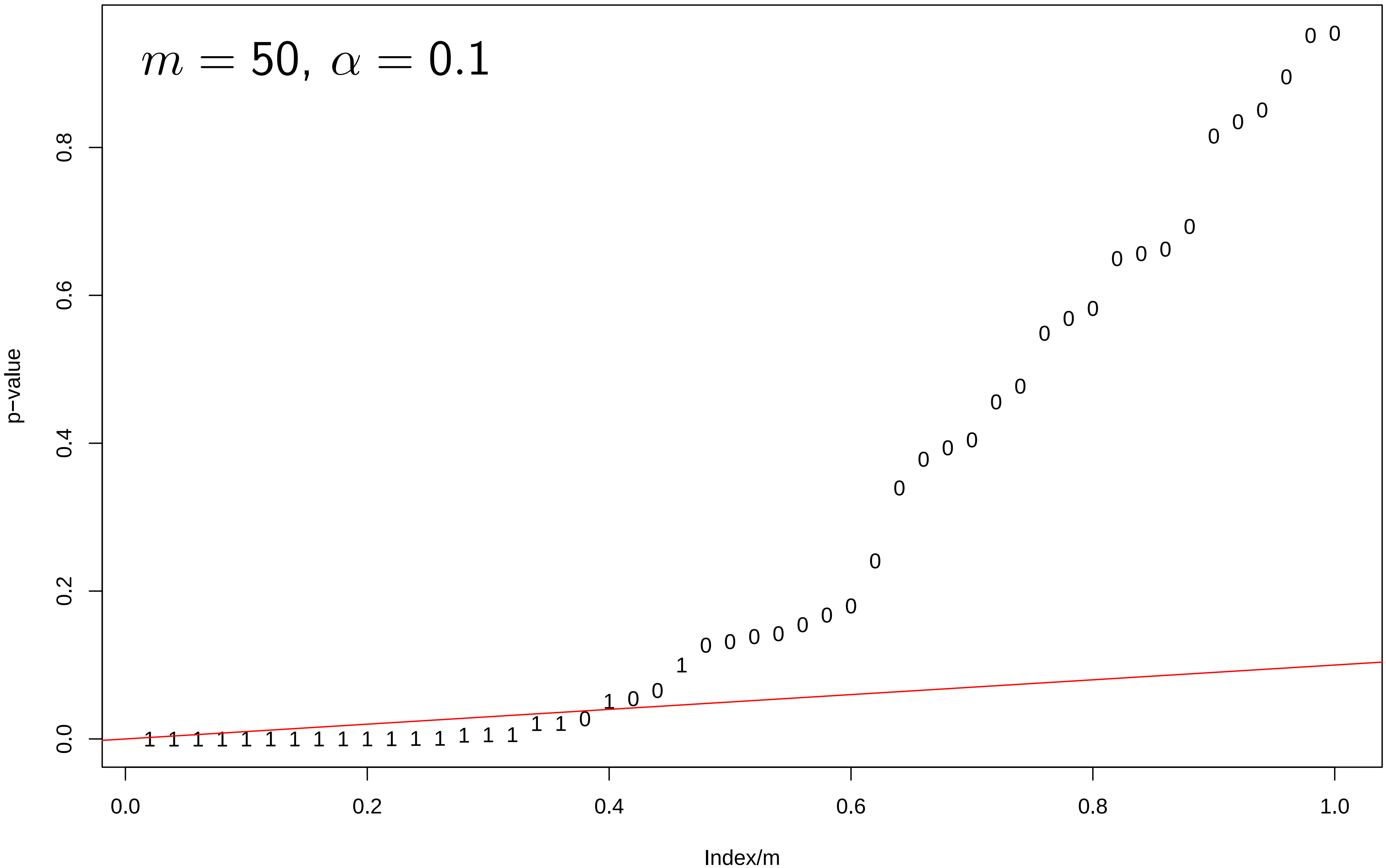
- Benjamini & Hochberg (1995) define a sequential p-value procedure that controls *expected* FDR.

  Specifically, the BH procedure guarantees

$$\text{E(FDR)} \leq \frac{M_0}{m}\alpha \leq \alpha$$

  for a pre-specified $0 < \alpha < 1$.

  (The first inequality is an equality in the continuous case.)


$m = 50,\ \alpha = 0.1$
p−value
Index/m
0.0
0.2
0.4
0.6
0.8
1.0

- The BH procedure for p-values $P_1, \ldots, P_m$:

  0. Select $0 < \alpha < 1$.

  1. Define $P_{(0)} \equiv 0$ and

  $$R_{\mathrm{BH}} = \max\left\{0 \le i \le m\colon\ P_{(i)} \le \alpha \frac{i}{m}\right\}.$$

  2. Reject H_0 for every test where $P_j \le P_{(R_{\mathrm{BH}})}$.

- Several variant procedures also control E(FDR).

- Bound on E(FDR) holds if p-values are independent or positively dependent (Benjamini & Yekutieli, 2001). Storey (2001) shows it holds under a possibly weaker condition.

- By replacing α with $\alpha / \sum_{i=1}^{m} 1/i$, control E(FDR) at level α for any joint distribution on the p-values. (Very conservative!)

Recent Work on FDR

Benjamini & Hochberg (1995)
Benjamini & Liu (1999)
Benjamini & Hochberg (2000)
Benjamini & Yekutieli (2001)

Storey (2001a,b)
Efron, et al. (2001)
Storey & Tibshirani (2001)
Tusher, Tibshirani, Chu (2001)

Abromovich, et al. (2000)

Genovese & Wasserman (2001a,b)
Genovese, Lazar, & Nichols (2002)

See also technical reports 735, 737, 747, 752, 754
at `http://lib.stat.cmu.edu/www/cmu-stats/tr/`.

Basic Models

- Let $H_i = 0$ (or 1) if the i^{th} null hypothesis is true (or false). These are unobserved.
- Let P_i be the i^{th} p-value.
- We assume that $(P_1, H_1), \ldots, (P_m, H_m)$ are independent with $P_i \mid \{H_i = 0\} \sim \text{Uniform}\langle 0, 1\rangle$, and $P_i \mid \{H_i = 1\} \sim F \in \mathcal{F}$, a class of alternative p-value distributions.
  - Under the *conditional model*, $H_1, \ldots, H_m$ are fixed, unknown.
  - Under the *mixture model*, we assume each $H_i \sim \text{Bernoulli}\langle a\rangle$.
- Define $M_0 = \sum_i (1 - H_i)$ and $M_1 = \sum_i H_i = m - M_0$. Under the *mixture model*, $M_1 \sim \text{Binomial}\langle m, a\rangle$. Under the *conditional model*, these are fixed.

Basic Models (cont'd)

- Typical examples:
  - Parametric family: $\mathcal{F}_\Theta = \{F_\theta\colon\ \theta \in \Theta\}$
  - Concave, continuous distributions

$$\mathcal{F}_C = \{F\colon\ F \text{ concave, continuous cdf with } F \geq U\}.$$

- Remark: The assumption of the mixture model does not require the same alternative for each test. For example, suppose that when the null is false

$$P_i \mid \Psi_i = \psi \sim F_\psi$$
$$\Psi_i \sim L$$

Then, $F = \int F_\psi \, dL(\psi)$.

Multiple Testing Procedures

- A multiple testing procedure T is a map $[0,1]^m \to [0,1]$, where the null hypotheses are rejected in all those tests for which $P_i \le T(P^m)$. Often call T a *threshold*.
- Examples:

Uncorrected testing	$T_{\mathrm{U}}(P^m) = \alpha$
Bonferroni	$T_{\mathrm{B}}(P^m) = \alpha/m$
Fixed threshold at t	$T_t(P^m) = t$
First r	$T_{(r)}(P^m) = P_{(r)}$
Benjamini-Hochberg	$T_{\mathrm{BH}}(P^m) = P_{(R_{\mathrm{BH}})}$ or $\sup\{t\colon \widehat{G}(t) = t/\alpha\}$
Oracle	$T_{\mathrm{O}}(P^m) = \sup\{t\colon G(t) = (1-a)t/\alpha\}$
Plug-In	$T_{\mathrm{PI}}(P^m) = \sup\{t\colon \widehat{G}(t) = (1-\widehat{a})t/\alpha\}$
Regression Classifier	$T_{\mathrm{Reg}}(P^m) = \sup\{t\colon \widehat{\mathsf{P}}\{H_1{=}1 \mid P_1{=}t\} > 1/2\}$

The False Nondiscovery Rate

- Controlling FDR alone only deals with Type I errors.
- Define the *realized* False Nondiscovery Rate as follows:

$$\text{FNR} = \begin{cases} \dfrac{N_{0|1}}{m - R} & \text{if } R < m, \\ 0 & \text{if } R = m. \end{cases}$$

  This is the proportion of false non-rejections among those tests whose null hypothesis is not rejected.
- Idea: Combine FDR and FNR in assessment of procedures.

FDR and FNR as Stochastic Processes

- Define the realized FDR and FNR processes, respectively, by

$$\mathrm{FDR}(t) \equiv \mathrm{FDR}(t; P^m, H^m) = \frac{\sum_i 1\left\{P_i \le t\right\}(1 - H_i)}{\sum_i 1\left\{P_i \le t\right\} + \prod_i 1\left\{P_i > t\right\}}$$

$$\mathrm{FNR}(t) \equiv \mathrm{FNR}(t; P^m, H^m) = \frac{\sum_i 1\left\{P_i > t\right\} H_i}{\sum_i 1\left\{P_i > t\right\} + \prod_i 1\left\{P_i \le t\right\}}.$$

- For procedure T, the realized FDR and FNR are obtained by evaluating these processes at $T(P^m)$.
- Inherent difficulty: The processes and the threshold both depend on the observed data.

Next Question . . .

1. What is the False Discovery Rate?

2. Why does the BH method work?

3. How does the BH method perform?

4. Can BH be made more powerful?

5. What are the implications for inference?

6. How does dependence among the p-values affect the results?

BH as a Plug-in Procedure

- Let $\widehat{G}$ be the empirical cdf of P^m under the mixture model. Ignoring ties, $\widehat{G}(P_{(i)}) = i/m$, so BH equivalent to

$$T_{\mathrm{BH}}(P^m) = \max\left\{t\colon\ \widehat{G}(t) = \frac{t}{\alpha}\right\}.$$

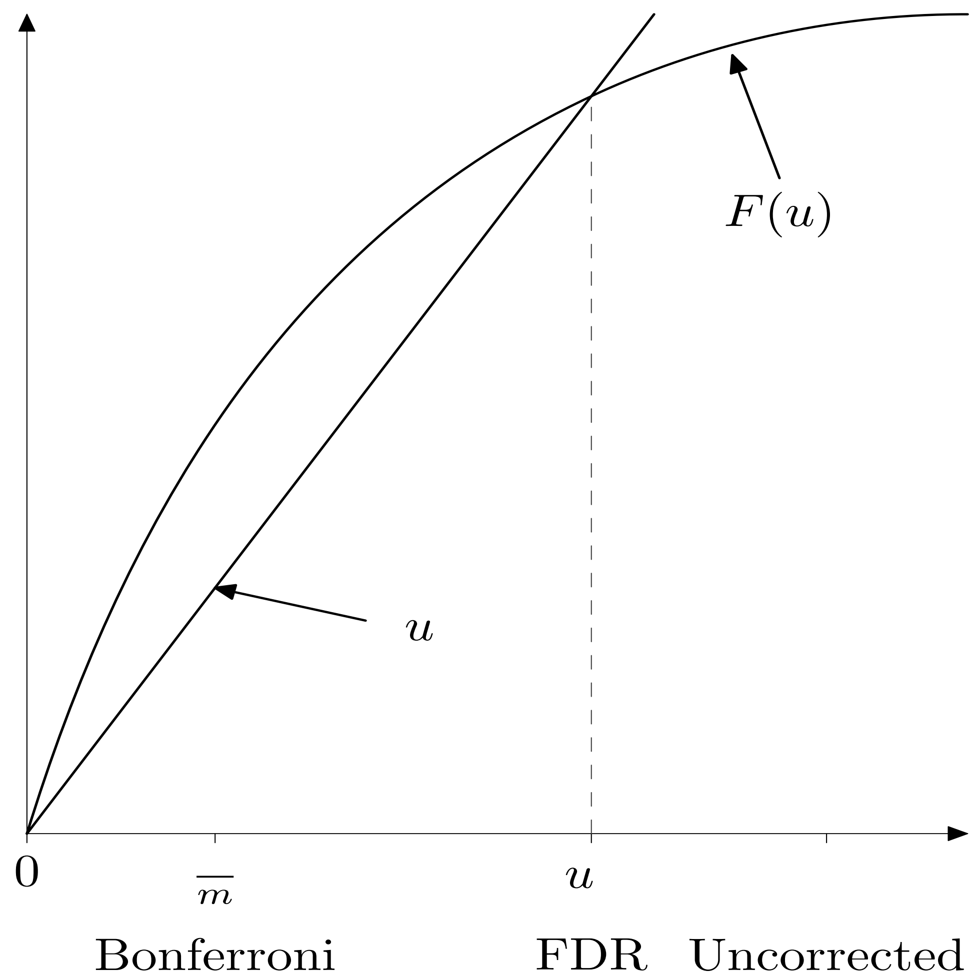
- We can think of this as a plug-in procedure for estimating

$$u^*(a, F) = \max\left\{t\colon\ G(t) = \frac{t}{\alpha}\right\} = \max\left\{t\colon\ F(t) = \beta t\right\},$$

where $\beta = (1 - \alpha + \alpha a)/\alpha a$.

Asymptotic Behavior of BH Procedure

This yields the following picture:



Next Question . . .

1. What is the False Discovery Rate?
2. Why does the BH method work?
3. How does the BH method perform?
4. Can BH be made more powerful?
5. What are the implications for inference?
6. How does dependence among the p-values affect the results?

Operating Characteristics of the BH Method

- Define the misclassification risk of a procedure T by

$$R_M(T) = \frac{1}{m}\sum_{i=1}^{m} \mathsf{E}\left|1\left\{P_i \le T(P^m)\right\} - H_i\right|.$$

  This is the average fraction of errors of both types.

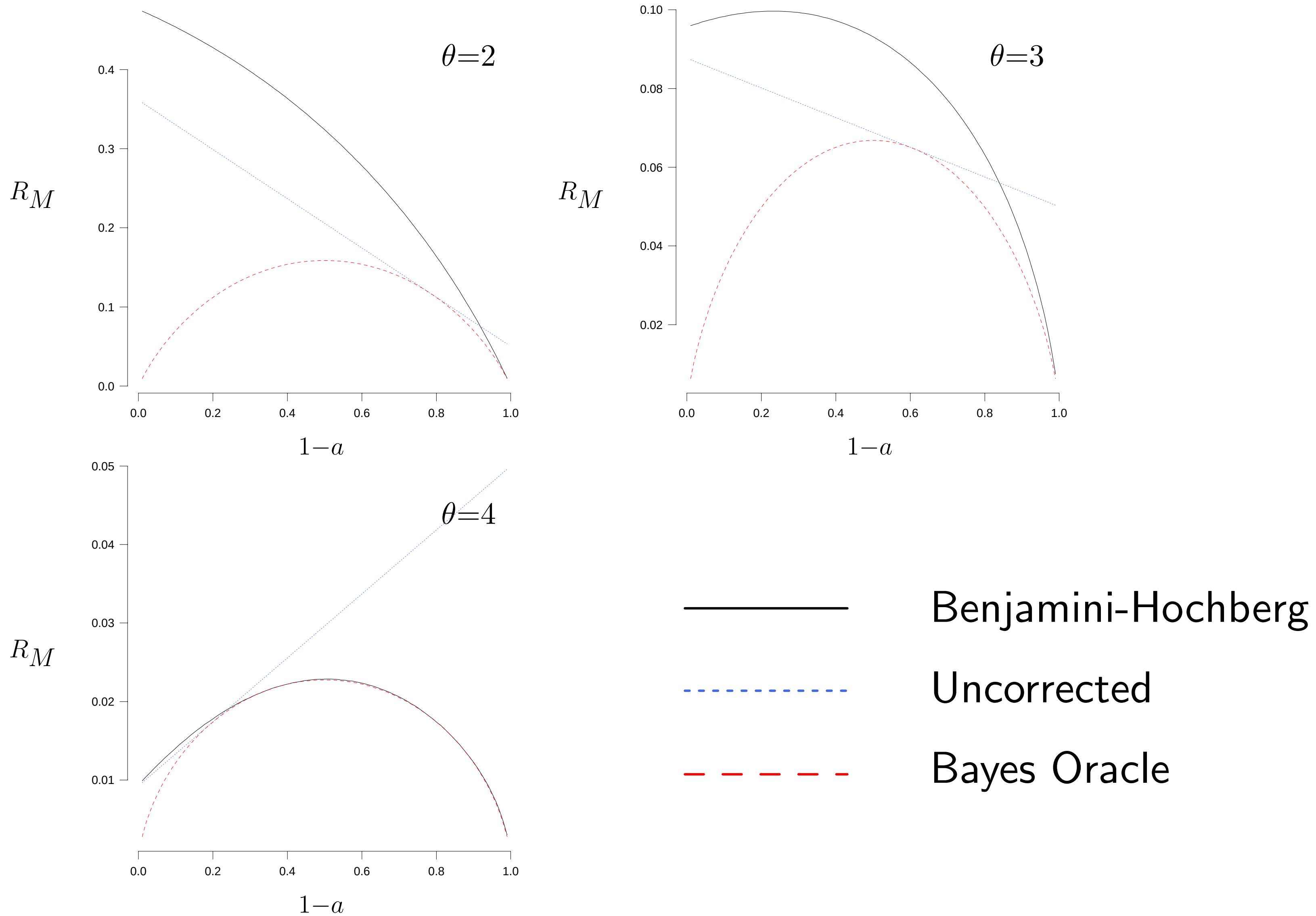
- Then $R_M(T_{\rm BH}) \sim R(a, F)$ as $m \to \infty$, where

$$R(a,F) = (1-a)u^* + a(1-F(u^*)) = (1-a)u^* + a(1-\beta u^*).$$

- Compare this to Uncorrected and Bonferroni and the Bayes' oracle rule $T_{\rm BO}(P^m) = b$ where b solves $f(b) = (1-a)/a$.

$$\begin{aligned}
R_M(T_{\rm U}) &= (1-a)\,\alpha \;\; + a\,(1 - F(\alpha)) \\
R_M(T_{\rm B}) &= (1-a)\,\frac{\alpha}{m} + a\left(1 - F\left(\frac{\alpha}{m}\right)\right) \\
R_M(T_{\rm BO}) &= (1-a)\,b \;\; + a\,(1 - F(b)).
\end{aligned}$$

Normal$\langle\theta, 1\rangle$ Model, $\alpha = 0.05$



Next Question . . .

1. What is the False Discovery Rate?

2. Why does the BH method work?

3. How does the BH method perform?

4. Can BH be made more powerful?

5. What are the implications for inference?

6. How does dependence among the p-values affect the results?

Optimal Thresholds

- Under the mixture model and in the continuous case,

$$\mathsf{E}(\mathsf{FDR}(T_{\mathrm{BH}}(P^m))) = (1-a)\alpha.$$

- The BH procedure overcontrols E(FDR) and thus will not in general minimize E(FNR).
- This suggests using T_{PI}, the plug-in estimator for

$$t^*(a, F) = \max\left\{t\colon\ G(t) = \frac{(1-a)t}{\alpha}\right\}$$
$$= \max\left\{t\colon\ F(t) = (\beta - 1/\alpha)t\right\},$$

  where $\beta - 1/\alpha = (1-a)(1-\alpha)/a\alpha$.
- Note that $t^* \geq u^*$.

Optimal Thresholds (cont'd)

- For each $0 \le t \le 1$,

$$\mathsf{E}(\mathsf{FDR}(t)) = \frac{(1-a)\,t}{G(t)} + O\left((1-t)^m\right)$$

$$\mathsf{E}(\mathsf{FNR}(t)) = a\,\frac{1-F(t)}{1-G(t)} + O\left((a+(1-a)t)^m\right).$$

- Ignoring $O()$ terms and choosing t to minimize E(FNR(t)) subject to E(FDR(t)) $\le \alpha$, yields $t^*(a, F)$ as the optimal threshold.
- Can the potential improvement in power be achieved when estimating t^*? Yes, if $F \neq U$.

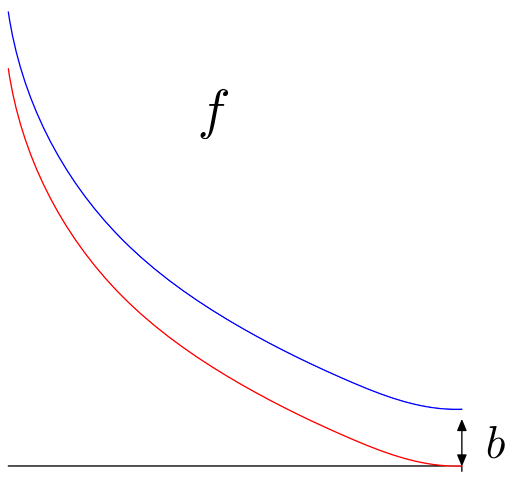
Plug-in Procedures

- The procedure T_{PI} is a plug-in estimator of the optimal $t^*(a, F)$

$$T_{\mathrm{PI}}(P^m) = \max\left\{t \colon\ \widehat{G}(t) = \frac{(1-\widehat{a})t}{\alpha}\right\}.$$

  We need good estimates of G and a to make this work. Later, we will also need good estimates of F.

- Identifiability and Purity



If $\min f = b > 0$, can write $F = (1-b)U + bF_0$,
$\mathcal{O}_G = \{(\widetilde{a}, \widetilde{F}) : \ \widetilde{F} \in \mathcal{F}, G = (1-\widetilde{a})U + \widetilde{a}\widetilde{F}\}$
may contain more than one element.

If $f = F'$ is decreasing with $f(1) = 0$, then (a, F) is identifiable.

Estimating a and F (cont'd)

- In general, let $\underline{a} \leq a$ be the smallest mixing weight in the orbit. $a - \underline{a}$ is typically small. For example, $a - \underline{a} = ae^{-n\theta^2/2}$ in the two-sided test of $\theta = 0$ versus $\theta \neq 0$ in the Normal$\langle \theta, 1 \rangle$ model.
- Parametric Case: (a, θ) typically identifiable; use MLE.
- Non-parametric case:
  - Derived a $1 - \beta$ one-sided conf. int. for $\underline{a}$ and thus a.
  - When F concave, get $\widehat{a}_{\mathrm{LCM}} = \underline{a} + O_P(m^{-1/3})$.
  - When F smooth enough, get $\widehat{a}_{\mathrm{S}} = \underline{a} + O_P(m^{-2/5})$.
  - Estimate F by: $\widehat{F}_m = \arg\min_{H \in \mathcal{F}} \|\widehat{G} - (1 - \widehat{a})U - \widehat{a}H\|_\infty$. Consistent for F_0 if $\widehat{a}$ consistent for $\underline{a}$.

Next Question . . .

1. What is the False Discovery Rate?

2. Why does the BH method work?

3. How does the BH method perform?

4. Can BH be made more powerful?

5. What are the implications for inference?

6. How does dependence among the p-values affect the results?

Confidence Thresholds

- In practice, it would be useful to have a procedure T_C that guarantees

$$\mathsf{P}_G\Big\{\mathsf{FDR}(T_C) > c\Big\} \le \alpha$$

for some specified c and α.

We call this a $(1-\alpha, c)$ *confidence threshold procedure*.

- Three approaches: (i) an asymptotic Bootstrap threshold, (ii) an asymptotic closed-form threshold, and (iii) an exact (small-sample) threshold requiring numerical search.

- Here, I'll discuss the case where a is known.

  In general, all of this works using a consistent estimate of $\underline{a}$, but this introduces additional complexity.

Bootstrap Confidence Thresholds

- First guess: Choose T such that

$$\mathsf{P}_{\widehat{G}}\Big\{\mathsf{FDR}^*(T) \leq c\Big\} \geq 1 - \alpha.$$

  Unfortunately, this fails.

- The problem is an additional bias term:

$$\begin{aligned} 1 - \alpha &= \mathsf{P}_{\widehat{G}}\Big\{\mathsf{FDR}^*(T) \leq c\Big\} \\ &\approx \mathsf{P}_{G}\Big\{\mathsf{FDR}(T) \leq c + (Q(T) - \widehat{Q}(T))\Big\} \\ &\neq \mathsf{P}_{G}\Big\{\mathsf{FDR}(T) \leq c\Big\}, \end{aligned}$$

  where $Q = (1-a)U/G$ and $\widehat{Q} = (1-a)U/\widehat{G}$.

Bootstrap Confidence Thresholds (cont'd)

- Let $\beta = \alpha/2$ and $\epsilon_m \equiv \epsilon_m(\beta) = \sqrt{\frac{1}{2m} \log\left(\frac{2}{\beta}\right)}$.
- Procedure

  1. Draw $H_1^* \ldots, H_m^*$ iid Bernoulli$\langle a \rangle$
  2. Draw $P_i^* | H_i^*$ from $(1 - H_i^*)U + H_i^* \widehat{F}$.
  3. Define $\Omega_c^*(t) = \sum_i I\{P_i^* \leq t\}(1 - H_i^* - c)$.
  4. Use threshold defined by

$$T_C = \max\left\{t\colon \mathsf{P}_{\widehat{G}}\left\{\Omega_c^*(t) \leq -c\,\epsilon_m\right\} \geq 1 - \beta\right\}.$$

- Then,

$$\mathsf{P}_G\left\{\mathsf{FDR}(T_C) \leq c\right\} \geq 1 - \alpha + O\left(\frac{1}{\sqrt{m}}\right).$$

Closed-Form Asymptotic Confidence Thresholds

- Let t_0 solve $G(t_0) = (1-a)t_0/c$ and let $\widehat{t}_0$ denote an estimate of t_0 based on $\widehat{G}$.
- Let
$$T_C = \widehat{t}_0 + \frac{\widehat{\Delta}_{m,\alpha}}{\sqrt{m}},$$
where $\widehat{\Delta}$ is a complicated expression that depends on a density estimate of $g = G'$.
- Then, $\mathsf{P}_G\Big\{\mathsf{FDR}(T_C) \le c\Big\} \ge 1 - \alpha + o(1)$.
- This requires no bootstrapping but does require density estimation.

  This is analogous to the situation faced when estimating the standard error of a median.

Exact Confidence Thresholds

- Let $\mathcal{M}_\beta$ be a $1-\beta$ confidence set for M_0, derived from the Binomial$\langle m, 1-a\rangle$.
- Define

$$S(t; h^m, p^m) = \frac{\sum_i 1\left\{p_i \le t\right\}(1-h_i)}{\sum_i (1-h_i)},$$

$$\mathcal{U}_\beta(p^m) = \left\{ h^m \colon \sum_i (1-h_i) \in \mathcal{M}_\beta \text{ and } \|S(\cdot; h^m, p^m) - U\|_\infty \le \epsilon_{m_0}(\beta) \right\},$$

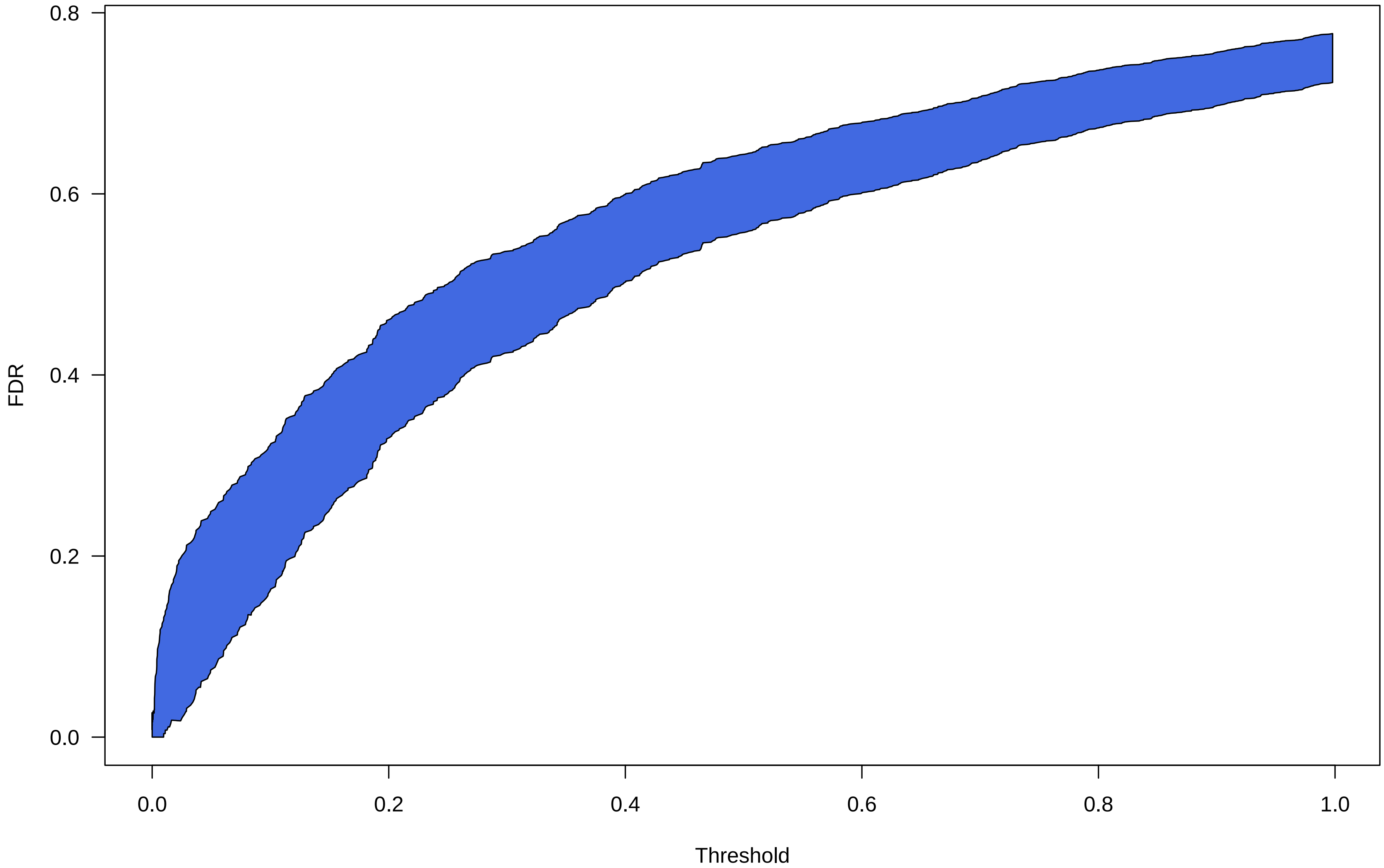
where $m_0 = \sum_i (1-h_i)$.
- If $\beta = 1 - \sqrt{1-\alpha}$, then $\mathsf{P}_G\Big\{H^m \in \mathcal{U}_\beta(P^m)\Big\} \ge 1-\alpha$ and

$$T_C = \sup\Big\{t : \ \mathsf{FDR}(t; h^m, P^m) \le c \text{ and } h^m : h^m \in \mathcal{U}_\beta(P^m)\Big\}$$

is a $(1-\alpha, c)$ confidence threshold procedure.
That is, $\mathsf{P}_G\Big\{\mathsf{FDR}(T_C) \le c\Big\} \ge 1-\alpha$.

Exact Confidence Thresholds (cont'd)

$\mathcal{U}$ yields a confidence envelope for FDR(t) sample paths.



Next Question . . .

1. What is the False Discovery Rate?
2. Why does the BH method work?
3. How does the BH method perform?
4. Can BH be made more powerful?
5. What are the implications for inference?
6. How does dependence among the p-values affect the results?

Dealing with Dependence

- Most of the foregoing assumed independence among the p-values. This rarely holds.
- Although standard BH works under "positive dependence", which often seems reasonable as with fMRI data.
- Yet, whatever form the dependence takes, BH is increasingly conservative as correlation increases.

  Hence, for example, spatial pre-smoothing of fMRI data is not recommended prior to BH.
- Two other approaches in my current work:
  - Local dependence and blocked correction
  - Incorporating estimated covariance into generalized plug-in procedure

Take-Home Points

- Realized versus Expected FDR
- Considering both FDR and FNR yields greater power
- Multiple testing problem is transformed to an estimation problem.
- Must control FDR and FNR as stochastic processes.

  In general, the threshold and the FDR are coupled, and these correlations can have a large effect.
- Results can be improved under dependence