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

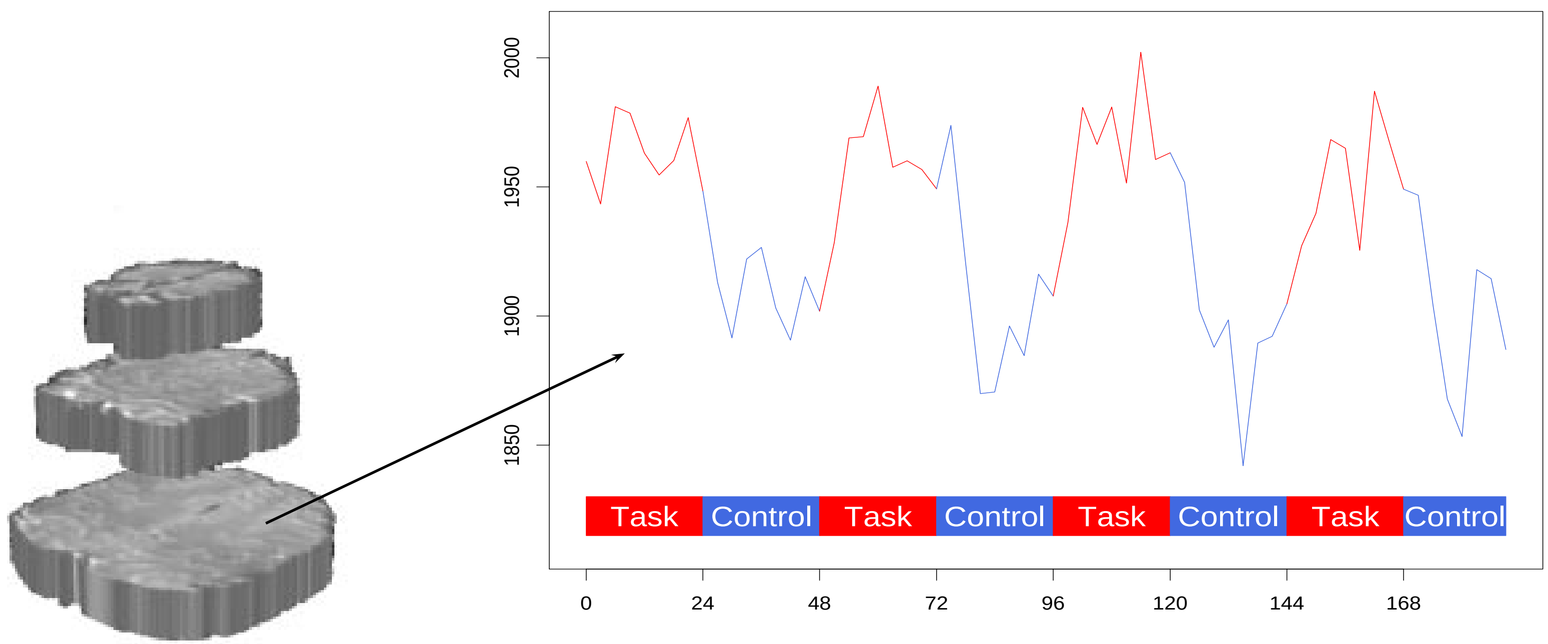
Christopher R. Genovese
Department of Statistics
Carnegie Mellon University

joint work with Larry Wasserman

This work partially supported by NSF Grant SES 9866147.

Motivating Example #1: fMRI

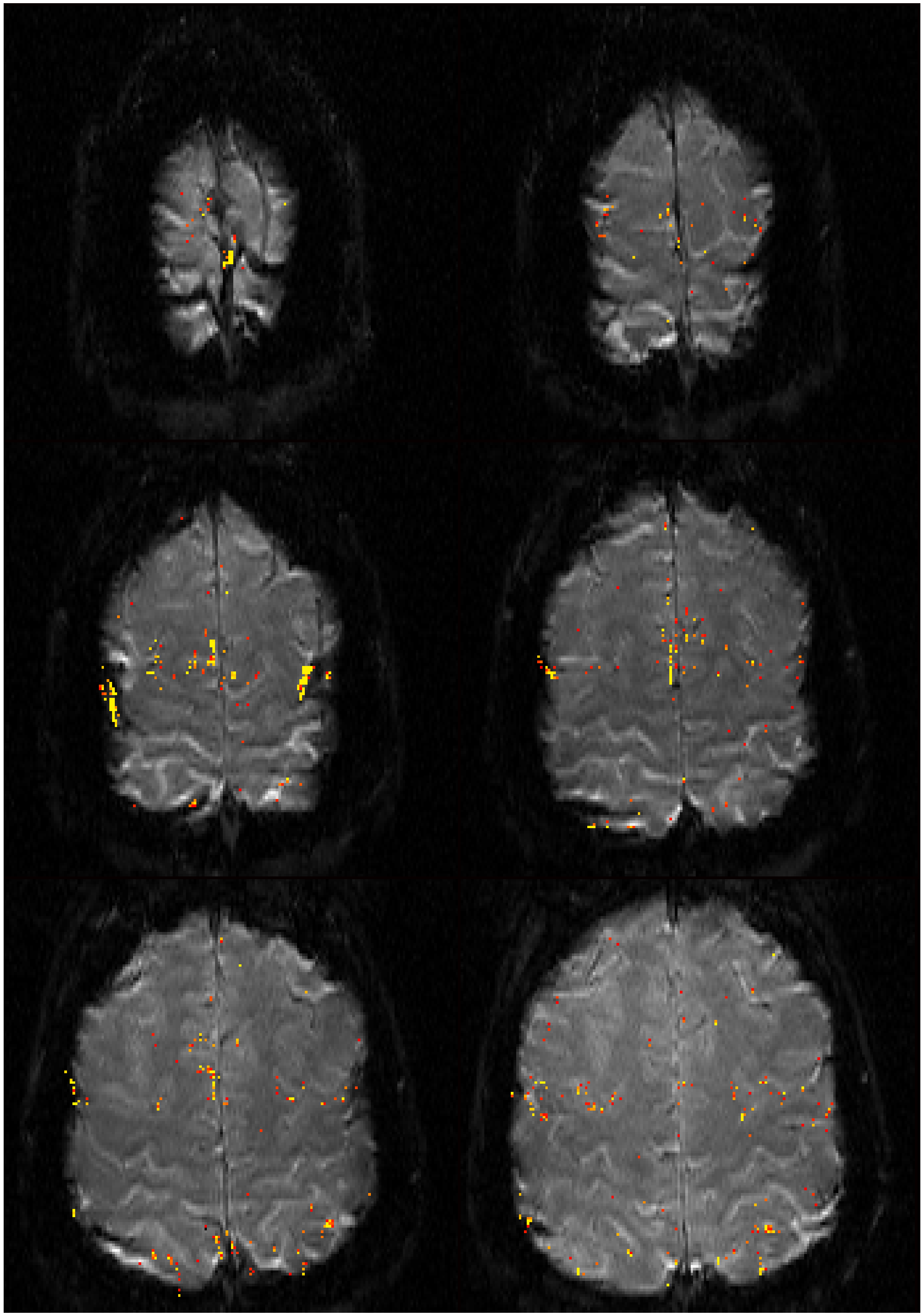
- fMRI Data: Time series of 3-d images acquired while subject performs specified tasks.



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

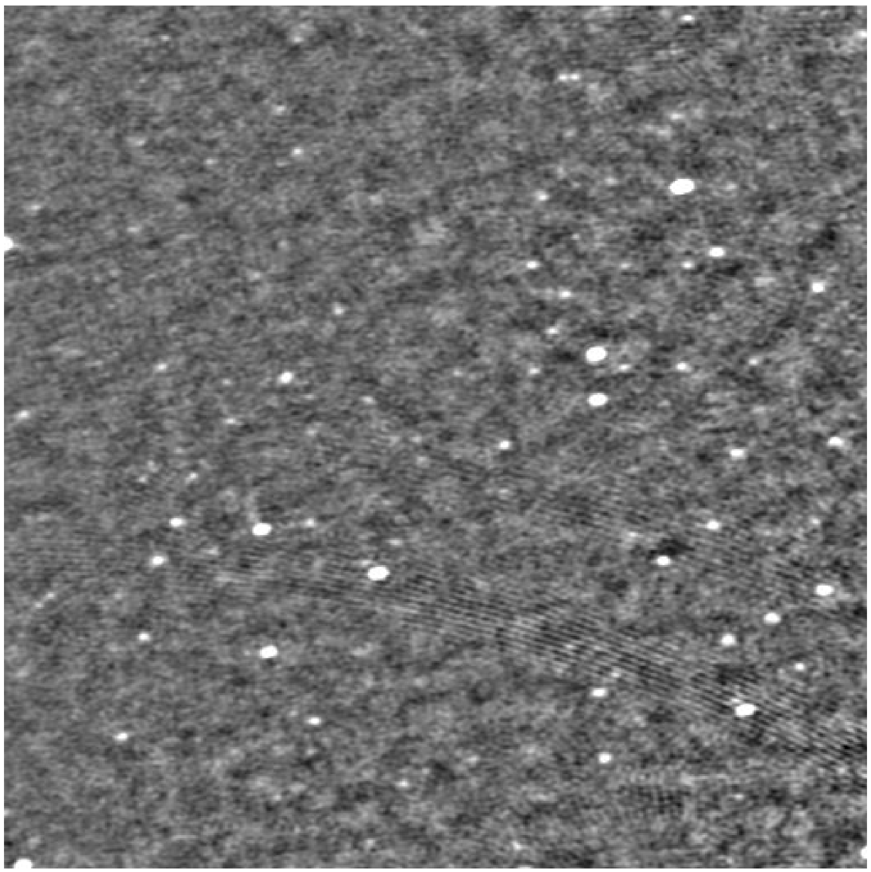
fMRI (cont'd)

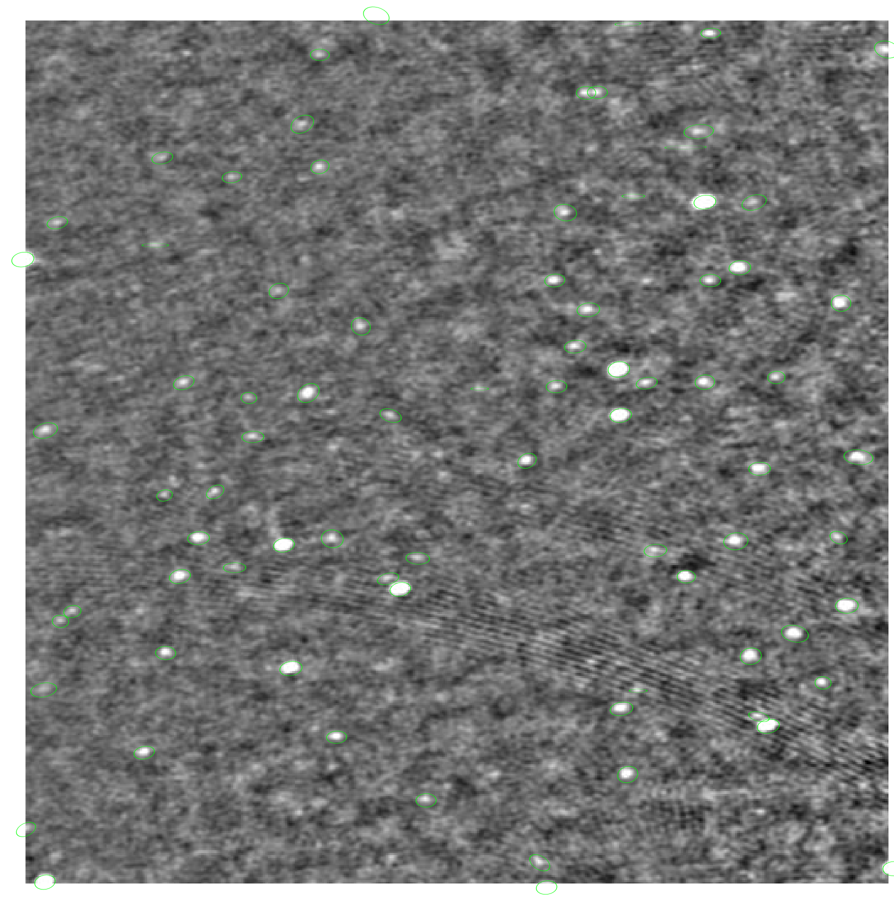
Perform hypothesis tests at many thousands of volume elements to identify loci of activation.



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	2	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	2	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.

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.

  This binds the separate decision rules together.

Multiple Testing (cont'd)

- Traditional methods seek strong control of familywise Type I error (FWER).
  - Weak Control: If all nulls true, $\mathrm{P}\left\{N_{1|0} > 0\right\} \leq \alpha$.
  - Strong Control: Corresponding statement holds for any subset of tests for which all nulls are true.

  For example, Bonferroni correction provides strong control but is quite conservative.

- 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

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

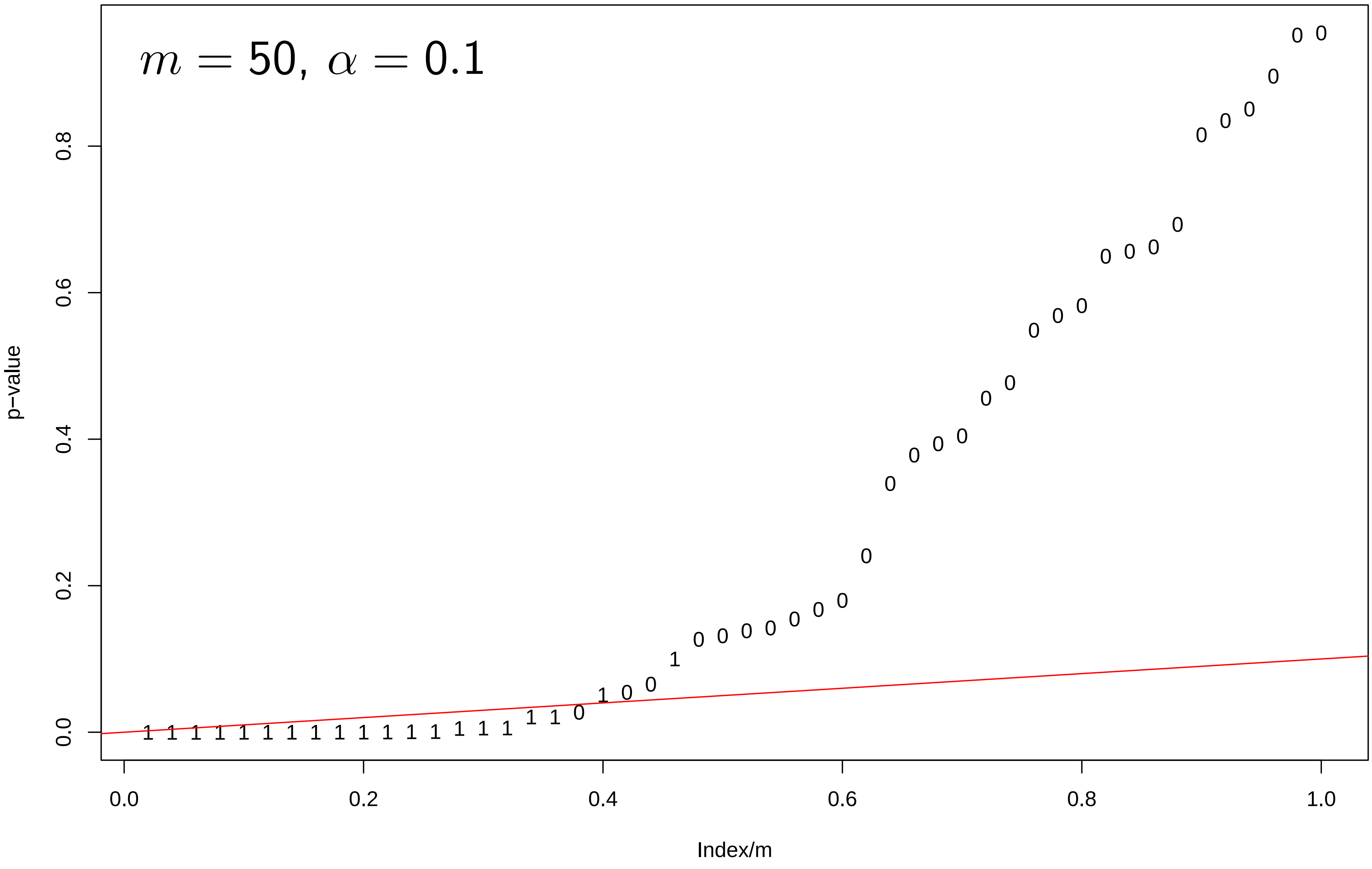
- Benjamini & Hochberg (1995) define a sequential p-value procedure that controls *expected* FDR.

  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
0.0
0.2
0.4
0.6
0.8
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

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

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

- 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!)

Road Map

1. Preliminaries

   – Considering both types of errors: The False Nondiscovery Rate (FNR)

   – Models for realized FDR and FNR

   – FDR and FNR as stochastic processes

2. Understanding BH

   – Re-express BH procedure as plug-in estimator

   – Asymptotic behavior of BH

   – Improving the power – more general plug-ins

   – Asymptotic risk comparisons

3. Extensions to BH

   – Confidence thresholds for FDR

   – Estimating the a and F

   – Conditional risk

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)

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

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.

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 $[0,1]^m \longrightarrow [0,1]$, where the null hypotheses are rejected in all those tests for which $P_i \leq T(P^m)$. 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	$T_{\mathrm{Reg}}(P^m) = \sup\{t\colon \widehat{\mathsf{P}}\{H_1{=}1 \mid P_1{=}t\} > 1/2\}$

FDR and FNR as Stochastic Processes

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

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

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

- 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.

FDR and FNR as Stochastic Processes (cont'd)

- Both these processes converge to Gaussian processes outside a neighborhood of 0 and 1 respectively.
- For example, define

$$Z_m(t) = \sqrt{m}\,(\mathsf{FDR}(t) - Q(t))\,, \quad \delta \le t \le 1,$$

  where $0 < \delta < 1$ and $Q(t) = (1-a)U/G$.

- Let Z be a mean 0 Gaussian process on $[\delta, 1]$ with covariance kernel

$$K(s,t) = a(1-a)\,\frac{(1-a)stF(s \wedge t)\; + \; aF(s)F(t)(s \wedge t)}{G^2(s)\,G^2(t)}.$$

- Then, $Z_m \rightsquigarrow Z$.

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

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

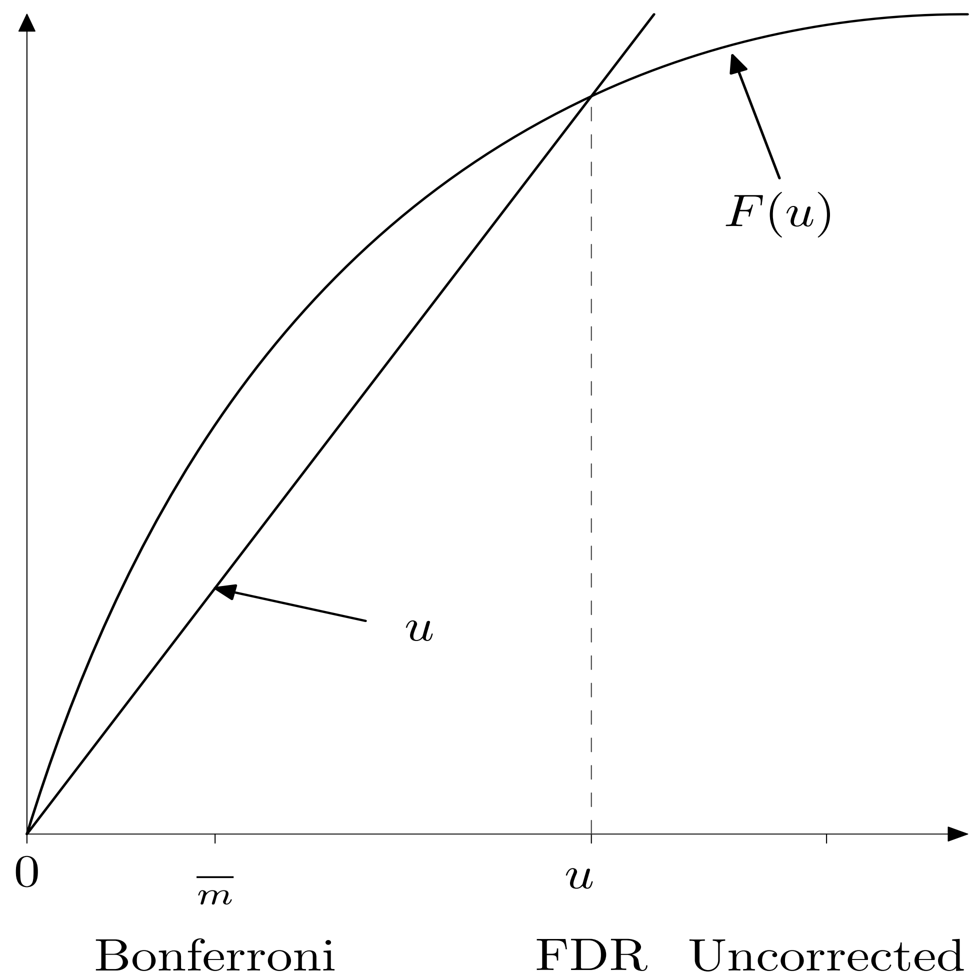
- We can think of this as a plug-in procedure for estimating

$$\begin{aligned} u^*(a, F) &= \operatorname{arg\,max}\left\{t\colon\ G(t) = \frac{t}{\alpha}\right\} \\ &= \operatorname{arg\,max}\left\{t\colon\ F(t) = \beta t\right\}, \end{aligned}$$

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

Asymptotic Behavior of BH Procedure

This yields the following picture:



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

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

  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 sufficiently far from U.

Operating Characteristics of the BH Method

- Define the misclassification risk of a procedure T by

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

  This is the average fraction of errors of both types.

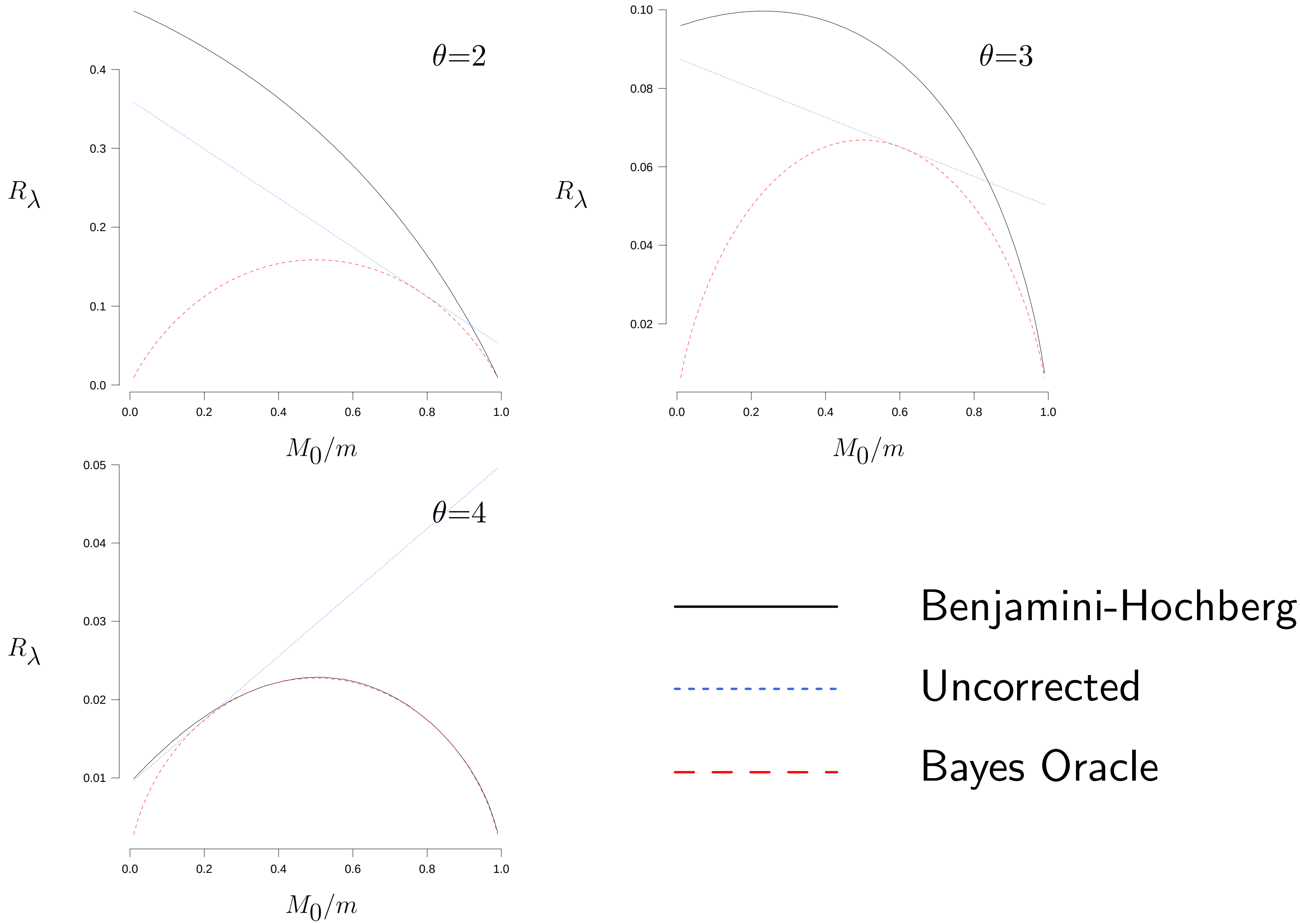
- Then $R_M(T_{\mathrm{BH}}) \sim R(a, F)$ as $m \to \infty$, where

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

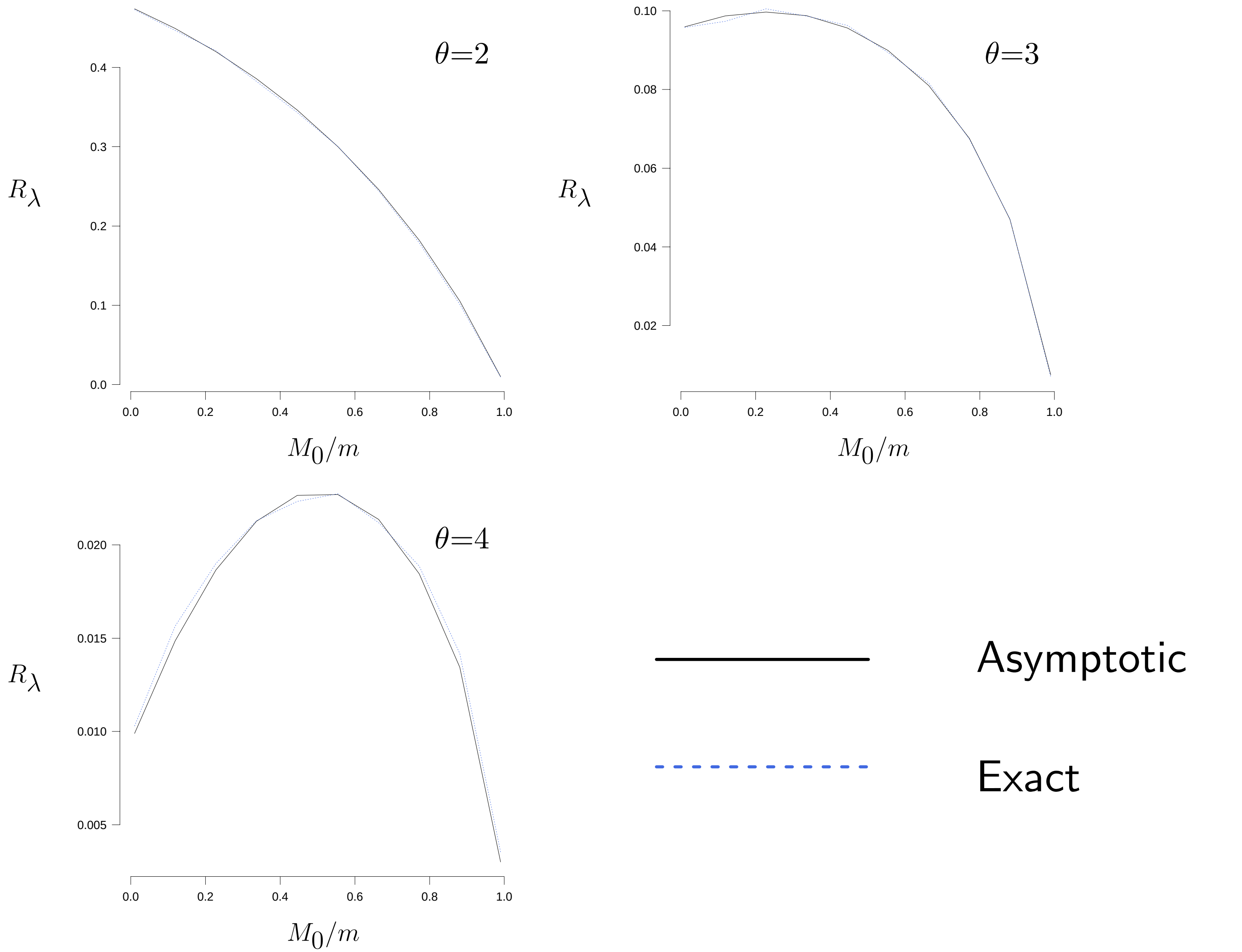
- Compare this to Uncorrected and Bonferroni and the Bayes' oracle rule $T_{\mathrm{BO}}(P^m) = b$ where b solves $f(b) = (1-a)/a$.

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

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



Check Approximation Accuracy $m = 100$



Extension 1: 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

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

  Unfortunately, this fails.

- The problem is an additional bias term:

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

  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 $\Omega_c^*(t) = \sum_i I\{P_i^* \le t\}(1 - H_i^* - c)$.
  4. Use threshold defined by

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

- Then,

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

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, $\mathsf{P}_G\Big\{\mathsf{FDR}(T_C) \leq c\Big\} \geq 1 - \alpha + o(1)$.
- 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

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

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

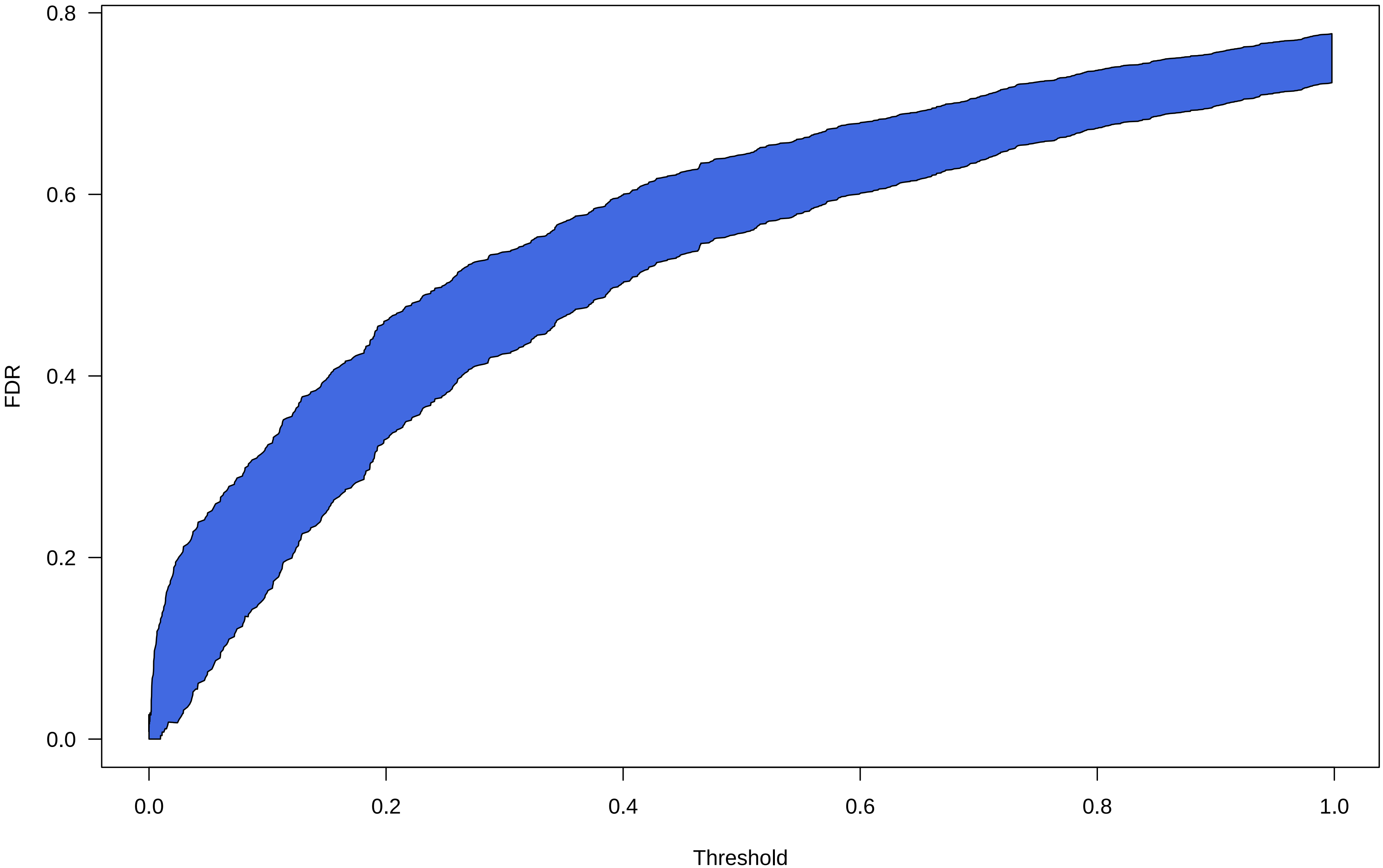
- Then, if $\beta = 1-\sqrt{1-\alpha}$, $\mathsf{P}_G\Big\{(H^m, P^m) \in \mathcal{U}\Big\} \ge 1-\alpha$ and

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

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.



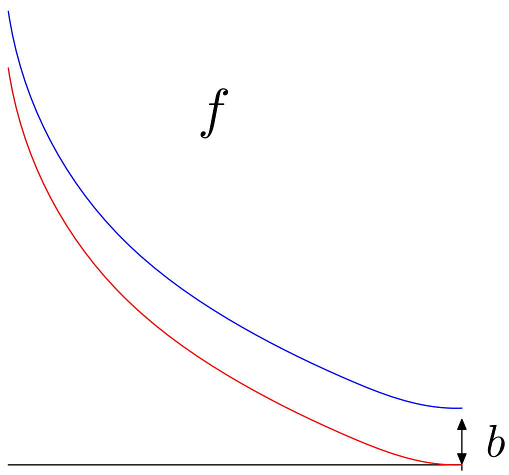
Extension 2: Estimating a and F

- We need a good estimate of a for plug-in estimates,

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

  that approximate the optimal threshold.

- We need good estimates of a and F for confidence thresholds.
- 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 $\hat{a}_{\mathrm{LCM}} = \underline{a} + O_P(m^{-1/3})$.
  - When F smooth enough, get $\hat{a}_{\mathrm{S}} = \underline{a} + O_P(m^{-2/5})$.
  - Estimate F by: $\widehat{F}_m = \arg\min_{H \in \mathcal{F}} \|\widehat{G} - (1 - \hat{a})U - \hat{a}H\|_\infty$. Consistent for F_0 if $\hat{a}$ consistent for $\underline{a}$.

Extension 3: Conditional Risk

- It is intuitively appealing (cf. Kiefer, 1977) to assess the performance of a procedure conditionally given the ordered p-values.
- When conditioning, we need only consider the $m + 1$ procedures $T_{(r)}(P^m) = P_{(r)}$ for $r = 0, \ldots, m$.
- Under the conditional model, once $P_{()}^m$ is observed, only the randomness in the labelling of the true classifications remains.
- Consider a parametric family $\mathcal{F} = \{F_\theta \colon \theta \in \Theta\}$ of alternative p-value distributions.

  Then, (M_0, θ) becomes the unknown parameter.
  Begin by treating this as known.

Conditional Risk (cont'd)

- Define a conditional risk for $\lambda \geq 0$ by

$$R_\lambda(r; M_0, \theta \mid P_{()}^m) = \mathsf{E}_{M_0,\theta}\left[\mathsf{FNR}(P_{(r)}) + \lambda\,\mathsf{FDR}(P_{(r)}) \,\middle|\, P_{()}^m\right],$$

  where M_0 and r are in $\{0, \ldots, m\}$ and $\theta \in \Theta$.

- Here λ determines the balance between the two error types.

  It also serves as a Lagrange multiplier for the optimization problem:

$$r_* = \arg\min_{0 \leq r \leq m} \mathsf{E}_{M_0,\theta}(\mathsf{FNR}(P_{(r)}) \mid P_{()}^m)$$

$$\text{subject to}$$

$$\mathsf{E}_{M_0,\theta}(\mathsf{FDR}(P_{(r)}) \mid P_{()}^m) \leq \alpha.$$

Conditional Risk (cont'd)

- This problem can be solved exactly:
  - Closed form for conditional distribution of FDR and FNR based on expressions for

$$\mathsf{P}_{M_0,\theta}\left\{N_{1|0} = k \mid P^m_{()}\right\} \quad \text{and} \quad \mathsf{E}_{M_0,\theta}(N_{1|0} \mid P^m_{()})$$

    derived via generating function methods.
  - Find R_λ-minimizer explicitly.
  - Select λ to satisfy the constraint.

- Remark: The R_λ minimizing conditional procedure also minimizes the unconditional R_λ risk, but the constrained optimization problem is harder to solve unconditionally.

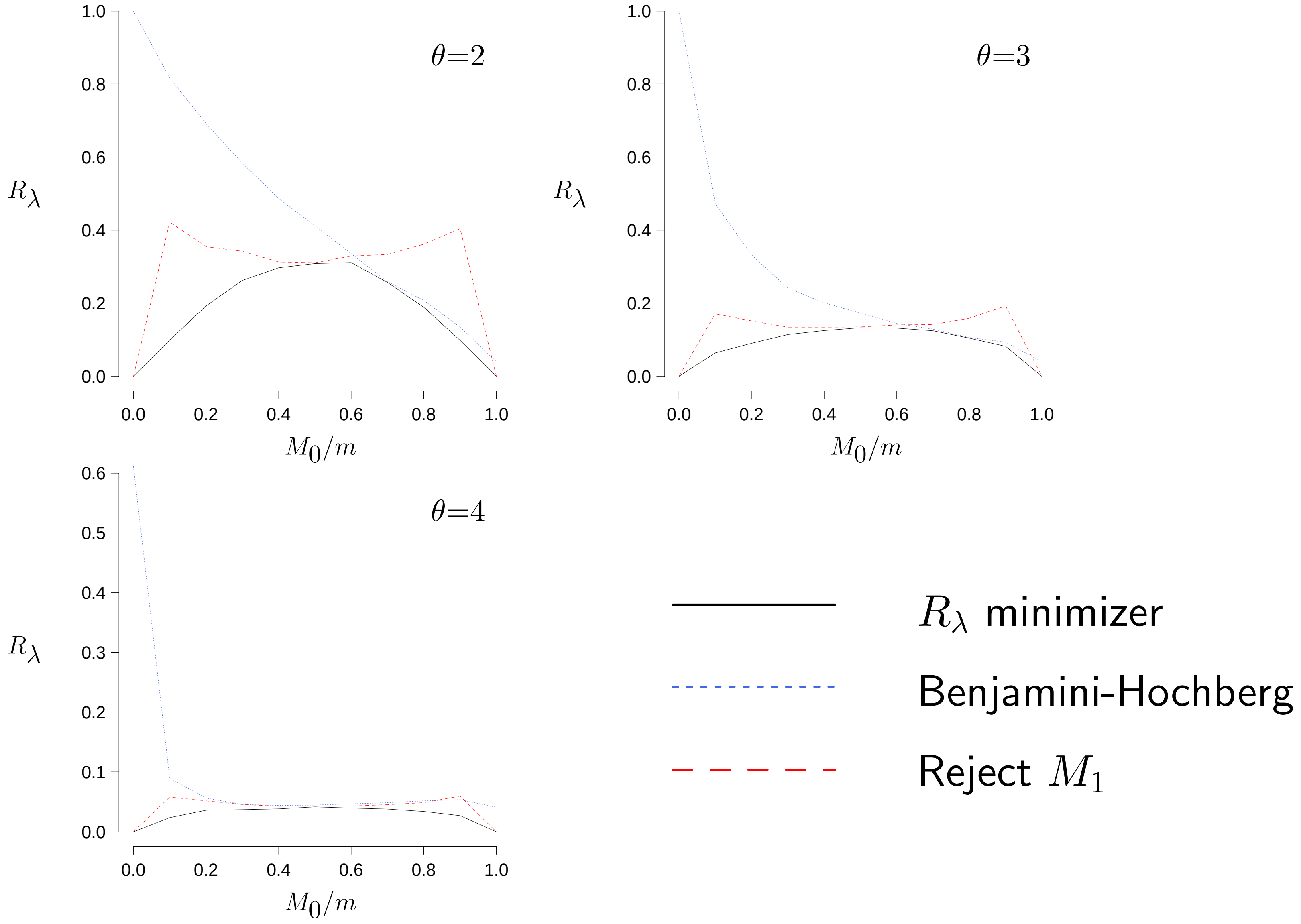
Conditional Risk (cont'd)

- For M_0 unknown case, R_λ dominated by extremes, $M_0 = 0$ or $M_0 = m$.
- One approach: minimize conditional Bayes risk based on R_λ

$$R_\lambda(r;\theta \mid P_{()}^m) = \sum_{m_0=0}^{m} R_\lambda(r; m_0, \theta \mid P_{()}^m)\, p_\theta(m_0 \mid P_{()}^m),$$

  where $p_\theta(m_0 \mid P_{()}^m)$ is derived from a specified (e.g., Uniform) prior on $\{0, \ldots, m\}$.
- This minimizes the unconditional Bayes risk.

Normal$\langle\theta, 1\rangle$ Model, $m = 100$, $\alpha = 0.05$



Bayesian FDR

- These conditional results yield the posterior distribution of FDR and FNR (and related quantities).

  No simulation necessary: can compute full posterior directly.

- Suggests the procedure $T_{\mathrm{Bayes}}(P^m) = P_{(r_*)}$, where

$$
\begin{aligned}
r_* = \arg \min_{0 \le r \le m} \ & \mathsf{E}(\mathsf{FNR}(P_{(r)}) \mid P^m_{()}) \\
\text{subject to} \ & \\
& \mathsf{E}(\mathsf{FDR}(P_{(r)}) \mid P^m_{()}) \le \alpha.
\end{aligned}
$$

- This procedure has good asympotic frequentist performance.

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.

Recurring Notation

$m, M_0, N_{1\|0}$	# of tests, true nulls, false discoveries
a	Mixture weight on $a$lternative
$H^m = (H_1, \ldots, H_m)$	Unobserved true classifications
$P^m = (P_1, \ldots, P_m)$	Observed p-values
$P^m_{()} = (P_{(1)}, \ldots, P_{(m)})$	Sorted p-values (define $P_{(0)} \equiv 0$)
U	CDF of Uniform$\langle 0, 1 \rangle$
F, f	Alternative CDF and density
$G = (1-a)U + aF$	Marginal CDF of P_i (mixture model)
$\widehat{G}$	Estimate of G (e.g., empirical CDF of P^m)
$\epsilon_n(\beta) = \sqrt{\frac{1}{2n} \log\left(\frac{2}{\beta}\right)}$	DKW bound $1-\beta$ quantile of $\\|\widehat{G}_n - G\\|_\infty$