New Approaches to False Discovery Control in Multiple Testing

Christopher R. Genovese

Department of Statistics

Carnegie Mellon University

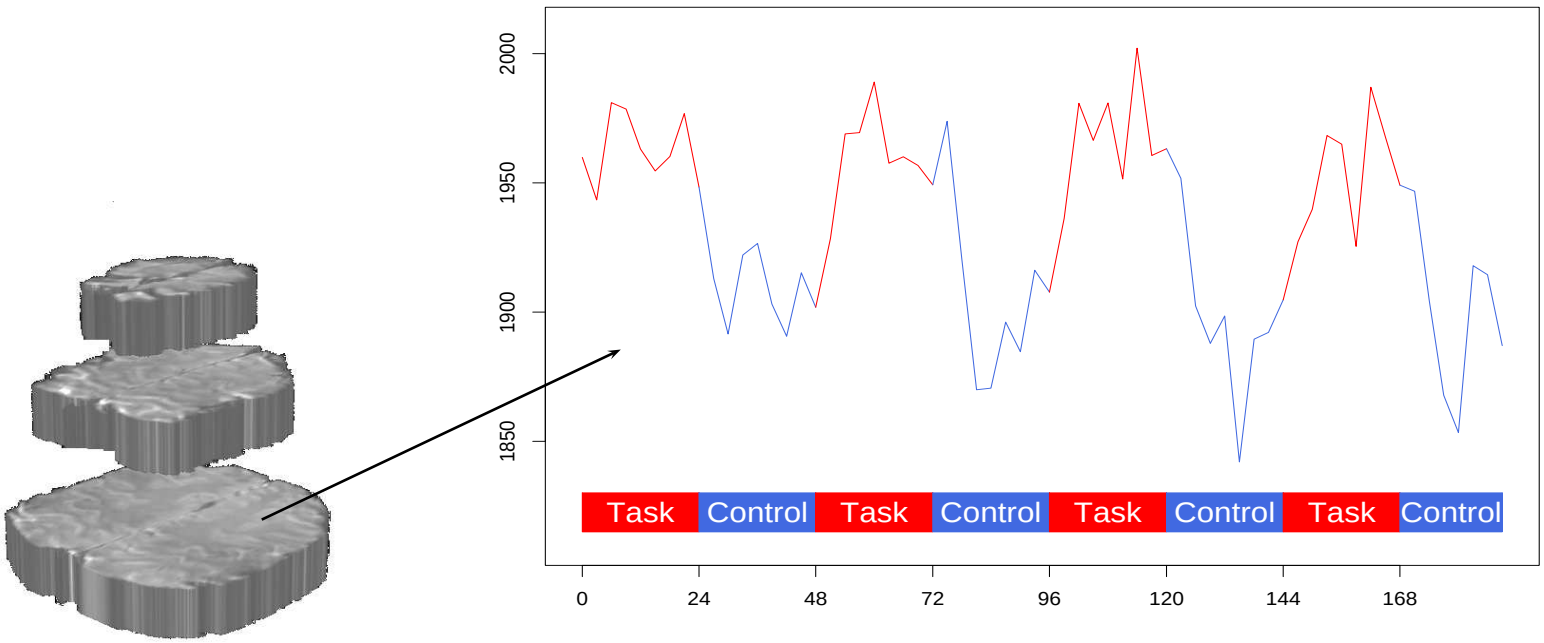
http://www.stat.cmu.edu/~genovese/

joint work with Larry Wasserman

This work partially supported by NSF Grants SES 9866147 and NIH Grant 1 R01 NS047493-01.

Motivating Example #1: fMRI

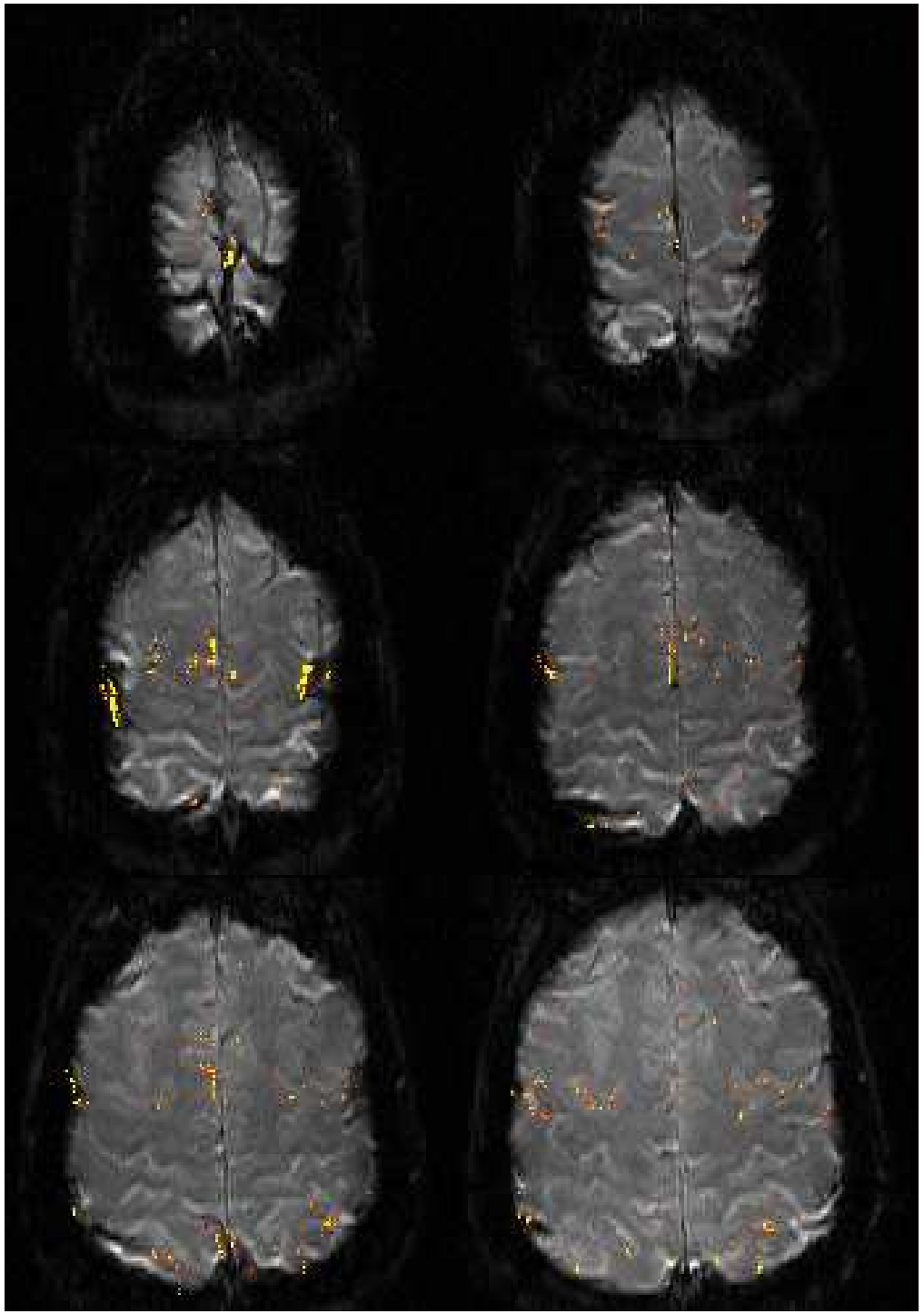
- fMRI Data: Time series of 3-d images acquired while subject performs specified tasks.



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

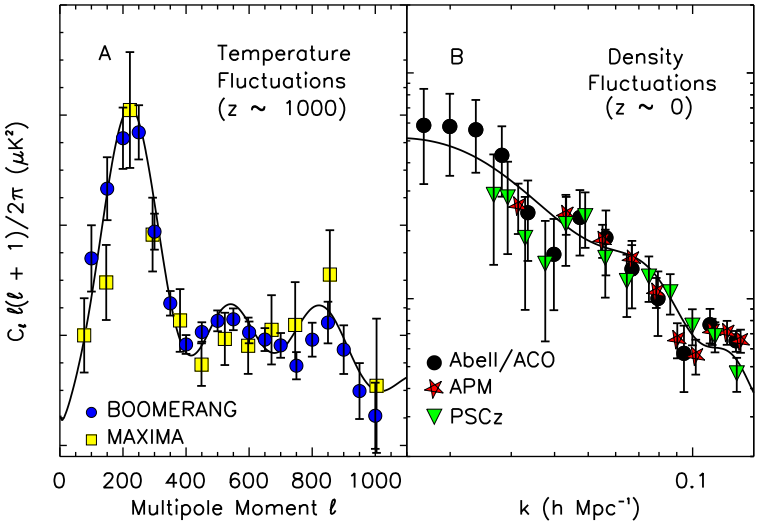
fMRI (cont'd)

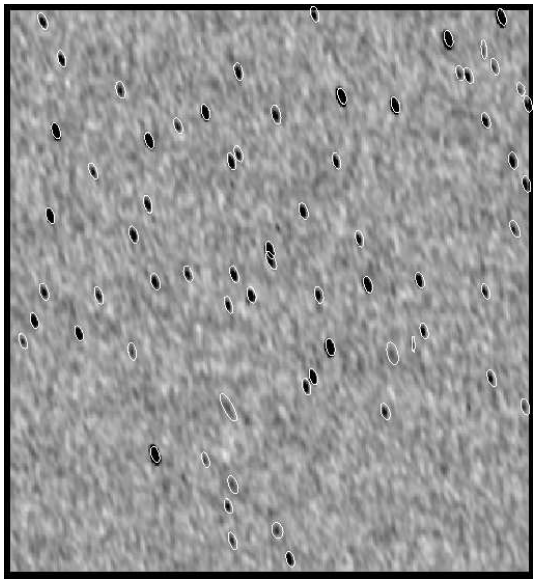
Perform hypothesis tests at many thousands of volume elements to identify loci of activation.

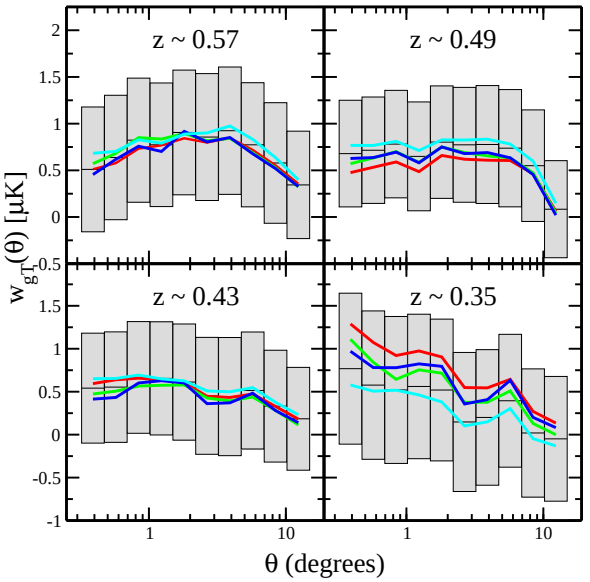


Motivating Example #2: Cosmology

- Baryon wiggles (Miller, Nichol, Batuski 2001)
- Radio Source Detection (Hopkins et al. 2002)
- Dark Energy (Scranton et al. 2003)

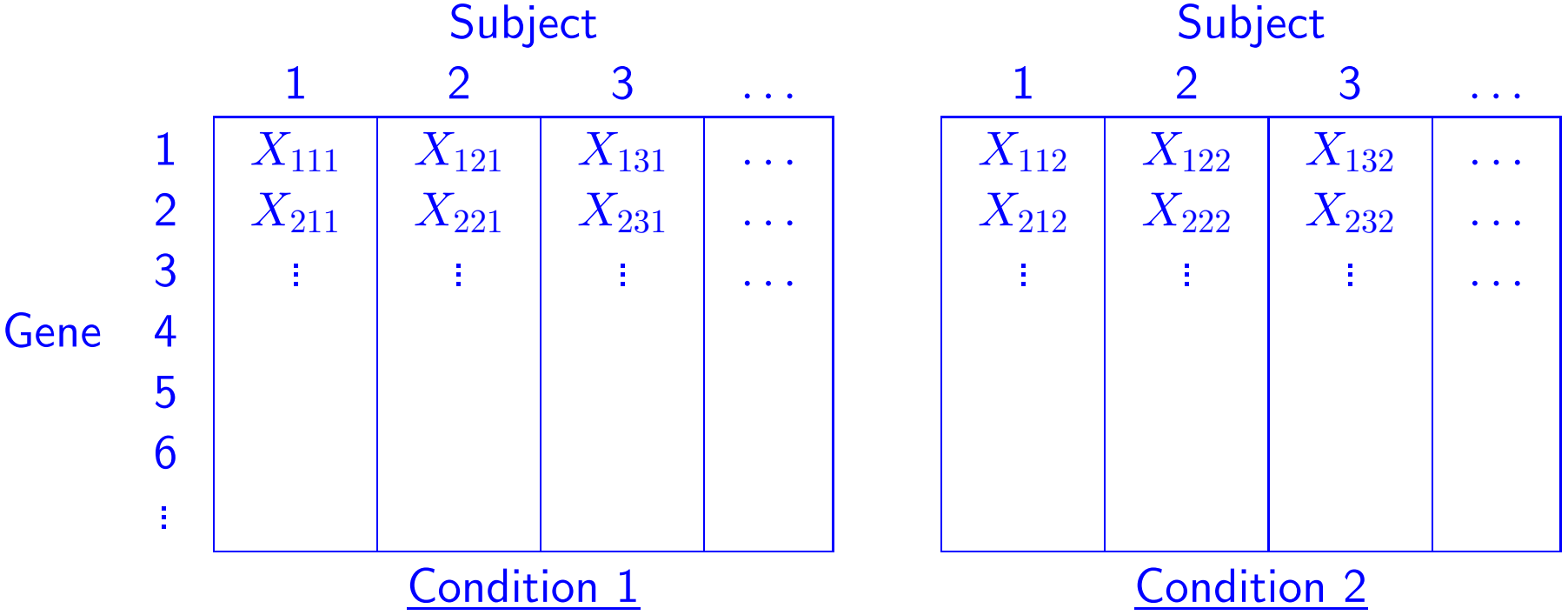






Motivating Example #3: DNA Microarrays

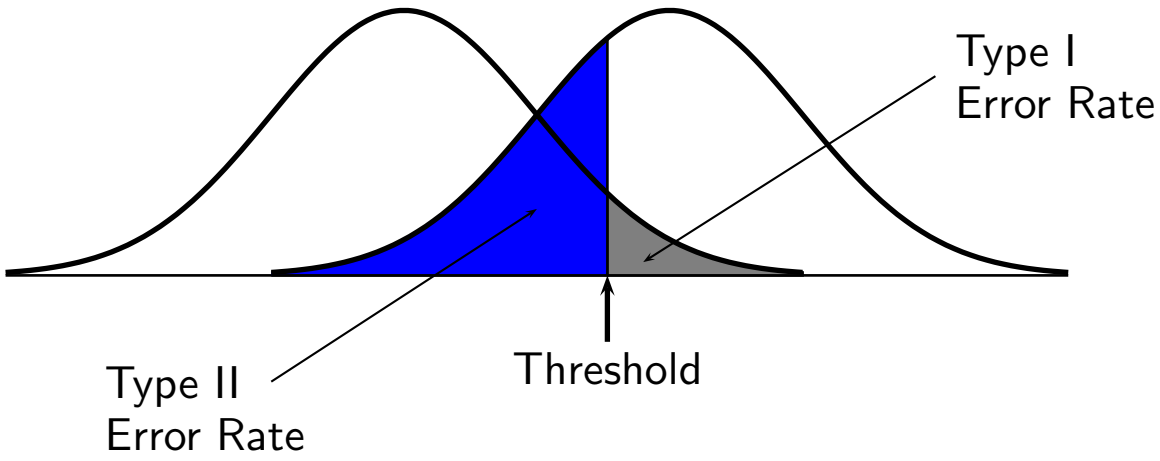
- New technologies allow measurement of gene expression for thousands of genes simultaneously.



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

One Test, One Threshold

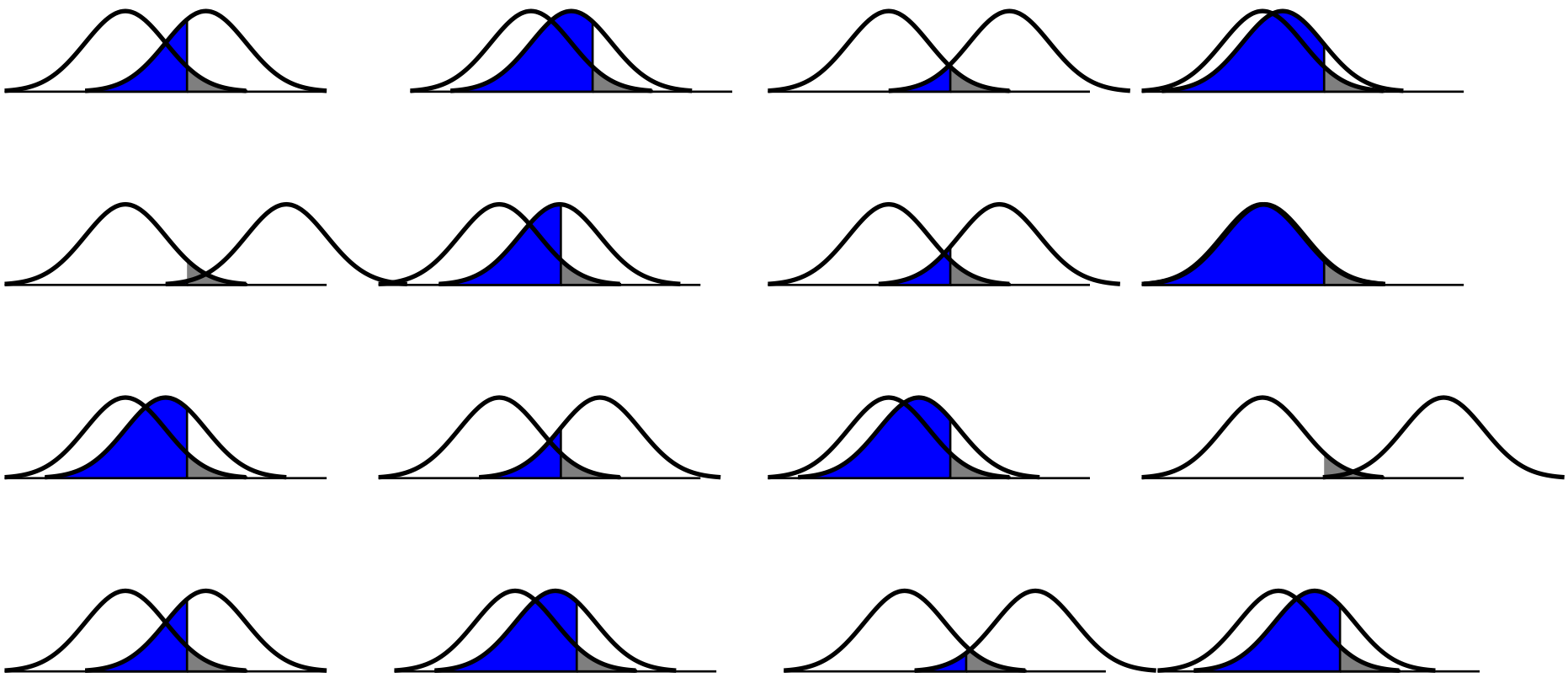
With a single hypothesis test, we choose a rejection threshold to control the Type I error rate,



while achieving a desirable Type II error rate for relevant alternatives.

Many Tests, One Threshold

With multiple tests, the problem is more complicated

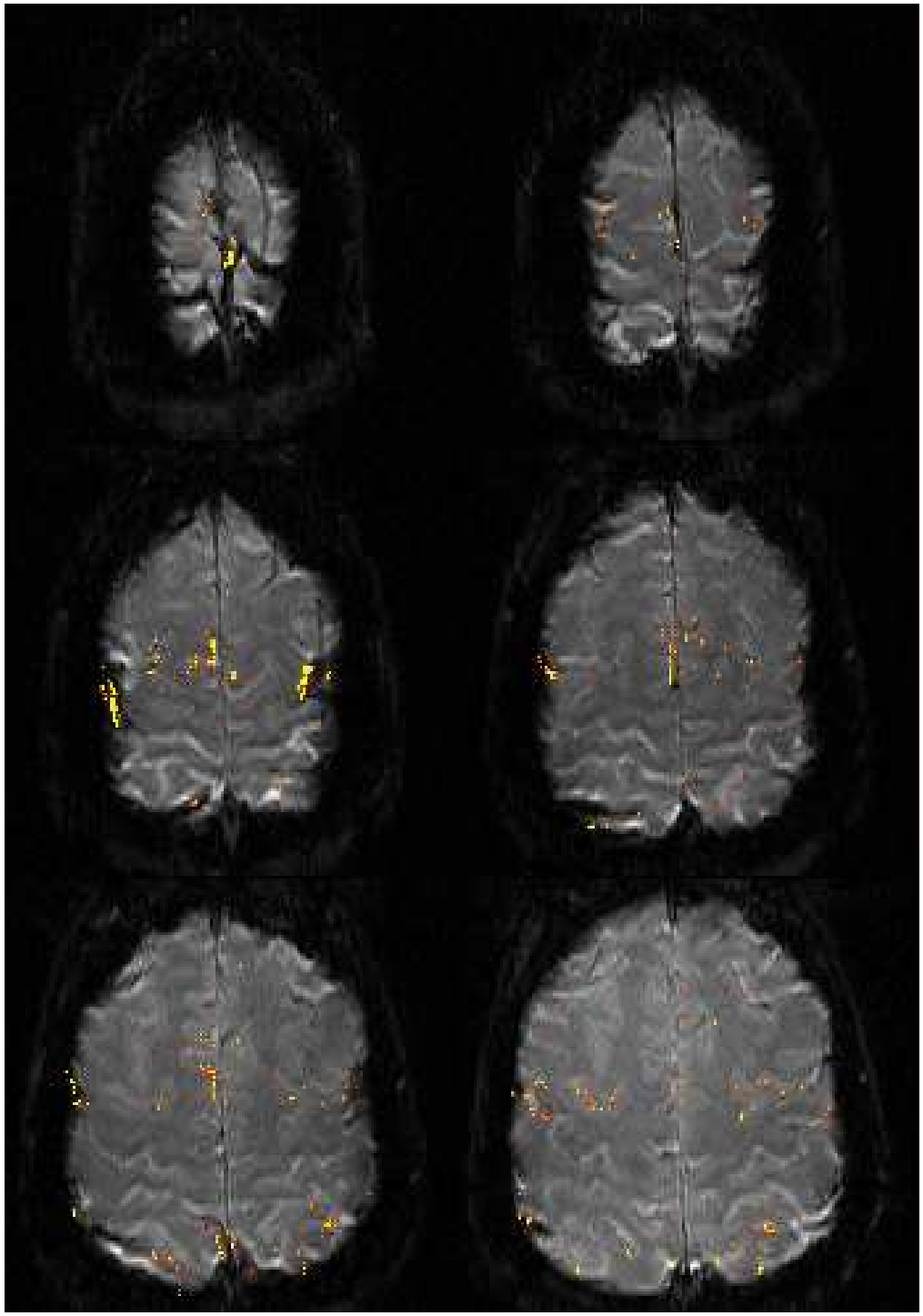


Each test has possible Type I and Type II errors, and there are many possible ways to combine them. The probability of a Type I error grows with the number of tests.

Many, Many Tests

It has become quite common in applications to perform *many thousands, even millions*, of simultaneous hypothesis tests.

Power is critical in these applications because the most interesting effects are usually at the edge of detection.



Plan

1. The Multiple Testing Problem
   - Error Criteria and Power
   - False Discovery Control and the BH Method
2. Why BH Works
   - A Useful Model
   - A Stochastic Process Perspective
   - Performance Characteristics
3. Toward False Discovery Control: Variations on BH
   - Improving Power
   - Dependence
   - Alternative Formulations
4. Exceedance Control and Random Fields

Plan

1. The Multiple Testing Problem
   - Error Criteria and Power
   - False Discovery Control and the BH Method
2. Why BH Works
   - A Useful Model
   - A Stochastic Process Perspective
   - Performance Characteristics
3. Toward False Discovery Control: Variations on BH
   - Improving Power
   - Dependence
   - Alternative Formulations
4. Exceedance Control and Random Fields

The Multiple Testing Problem

- Perform m simultaneous hypothesis tests with a common procedure.
- For any given procedure, classify the results as follows:

	H_0 Retained	H_0 Rejected	Total
H_0 True	TN	FD	T_0
H_0 False	FN	TD	T_1
Total	N	D	m

Mnemonics: T/F = True/False, D/N = Discovery/Nondiscovery

All quantities except m, D, and N are *unobserved*.

- The problem is to choose a procedure that balances the competing demands of sensitivity and specificity.

How to Choose a Threshold?

- Control Per-Comparison Type I Error (PCER)
  - a.k.a. "uncorrected testing," many type I errors
  - Gives $\mathbb{P}\{FD_i > 0\} \leq \alpha$ marginally for all $1 \leq i \leq m$

- Control Familywise Type I Error (FWER)
  - e.g.: Bonferroni: use per-comparison significance level α/m
  - Guarantees $\mathbb{P}\{FD > 0\} \leq \alpha$

- Control False Discovery Rate (FDR)
  - first defined by Benjamini & Hochberg (BH, 1995, 2000)
  - Guarantees FDR $\equiv \mathbb{E}\left(\frac{FD}{D}\right) \leq \alpha$

- ...

A Practical Problem

- While guarantee of FWER-control is appealing, the resulting thresholds often suffer from low power.

  In practice, this tends to wipe out evidence of the most interesting effects.

- FDR control offers a way to increase power while maintaining some principled bound on error.

  It is based on the assessment that

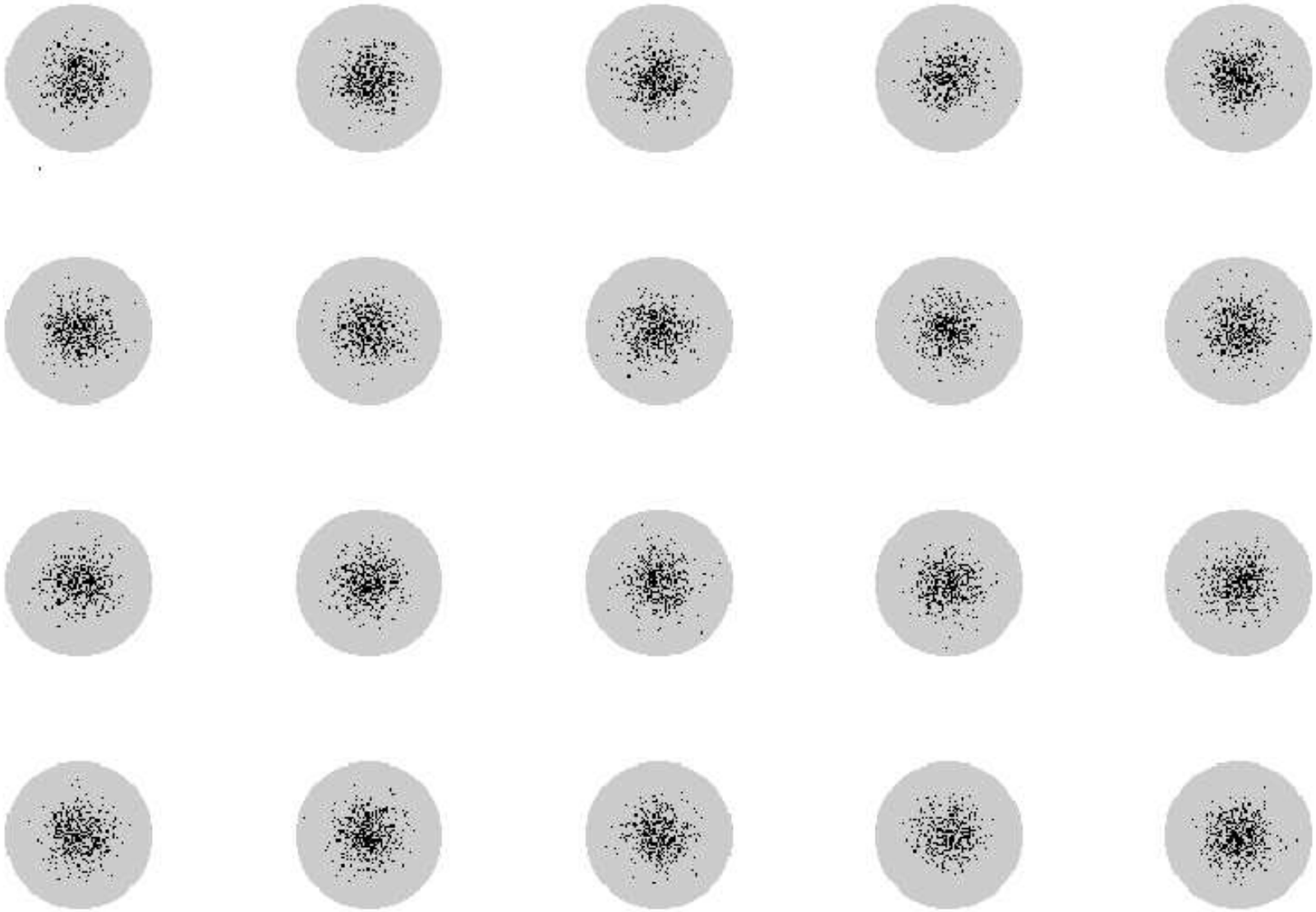
  4 false discoveries out of 10 rejected null hypotheses

  is a more serious error than

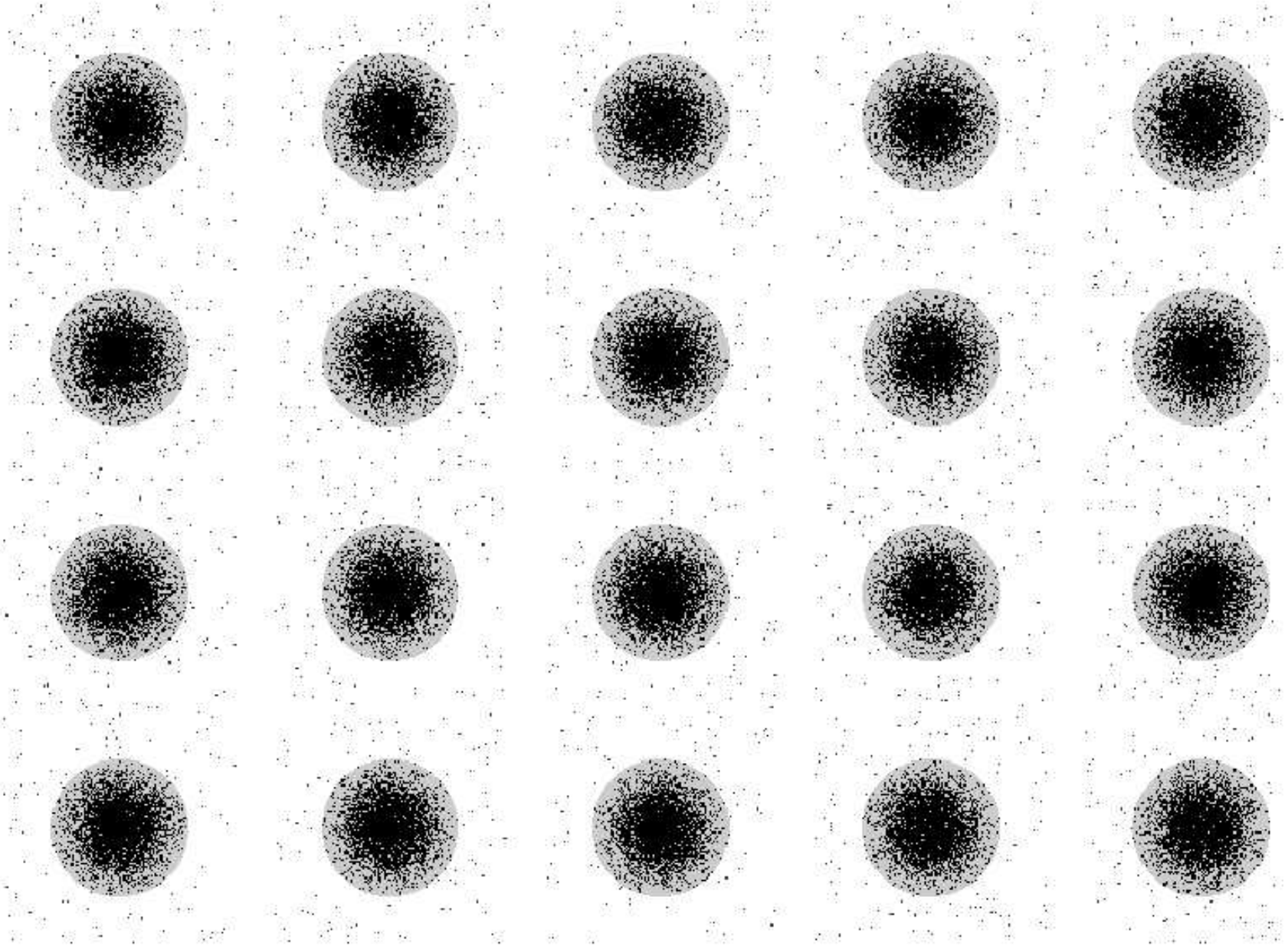
  20 false discoveries out of 100 rejected null hypotheses.

- A simple illustration ...

FWER Control

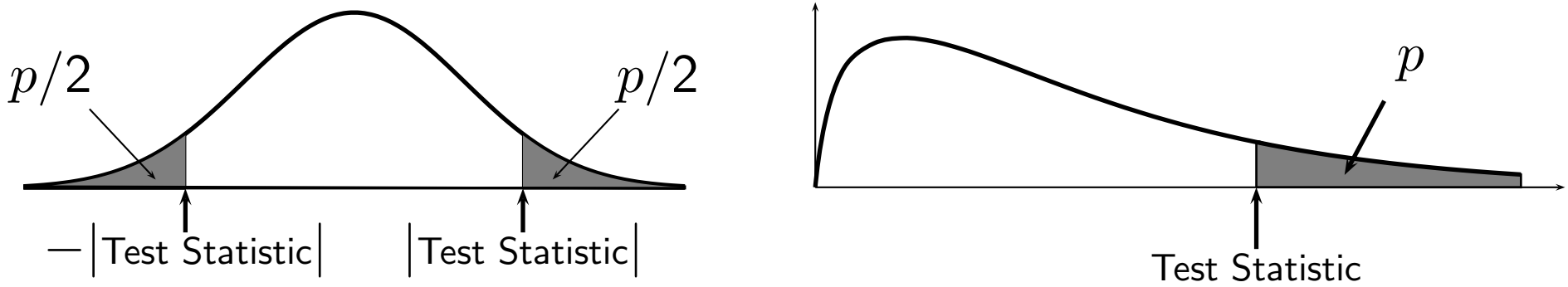


FDR Control



Recurring Notation

- Define p-values $P^m = (P_1, \ldots, P_m)$ for the m tests.



- Let $P_{(0)} \equiv 0$ and order the p-values

$$P_{(0)} = 0 < P_{(1)} < \cdots < P_{(m)}.$$

- Define hypothesis indicators $H^m = (H_1, \ldots, H_m)$, where $H_i = 0$ when the ith null hypothesis is true and $H_i = 1$ when the ith alternative is true.
- A multiple testing threshold T is a map $[0,1]^m \to [0,1]$, where we reject each null hypothesis with $P_i \leq T(P^m)$.

The False Discovery Rate

- Define the False Discovery Proportion (FDP) to be the (unobserved) *proportion of false discoveries among total rejections*.

  As a function of threshold t (and implicitly P^m and H^m), write this as

$$\text{FDP}(t) = \frac{\sum_i 1\Big\{P_i \leq t\Big\}\,(1 - H_i)}{\sum_i 1\Big\{P_i \leq t\Big\} + 1\Big\{\text{all } P_i > t\Big\}} = \frac{\#\text{False Discoveries}}{\#\text{Discoveries}}$$

- The False Discovery Rate (FDR) for a multiple testing threshold T is defined as the expected FDP using that procedure:

$$\text{FDR} = \mathbb{E}\Big(\text{FDP}(T)\Big)\,.$$

Aside: The False *Non*-Discovery Rate

- We can define a dual quantity to the FDR, the False Nondiscovery Rate (FNR).
- Begin with the False Nondiscovery Proprotion (FNP): the proportion of missed discoveries among those tests for which the null is retained.

$$\text{FNP}(t) = \frac{\sum_i 1\{P_i > t\} H_i}{\sum_i 1\{P_i > t\} + 1\{\text{all } P_i \leq t\}} = \frac{\#\text{False Nondiscoveries}}{\#\text{Nondiscoveries}}$$

- Then, the False Nondiscovery Rate (FNR) is given by

$$\text{FNR} = \mathbb{E}\left(\text{FNP}(T)\right).$$

The Benjamini-Hochberg Procedure

- Benjamini and Hochberg (BH, 1995) introduced the FDR and show that a procedure of Eklund and Simes controls it.
- The BH threshold is defined for pre-specified $0 < \alpha < 1$ as

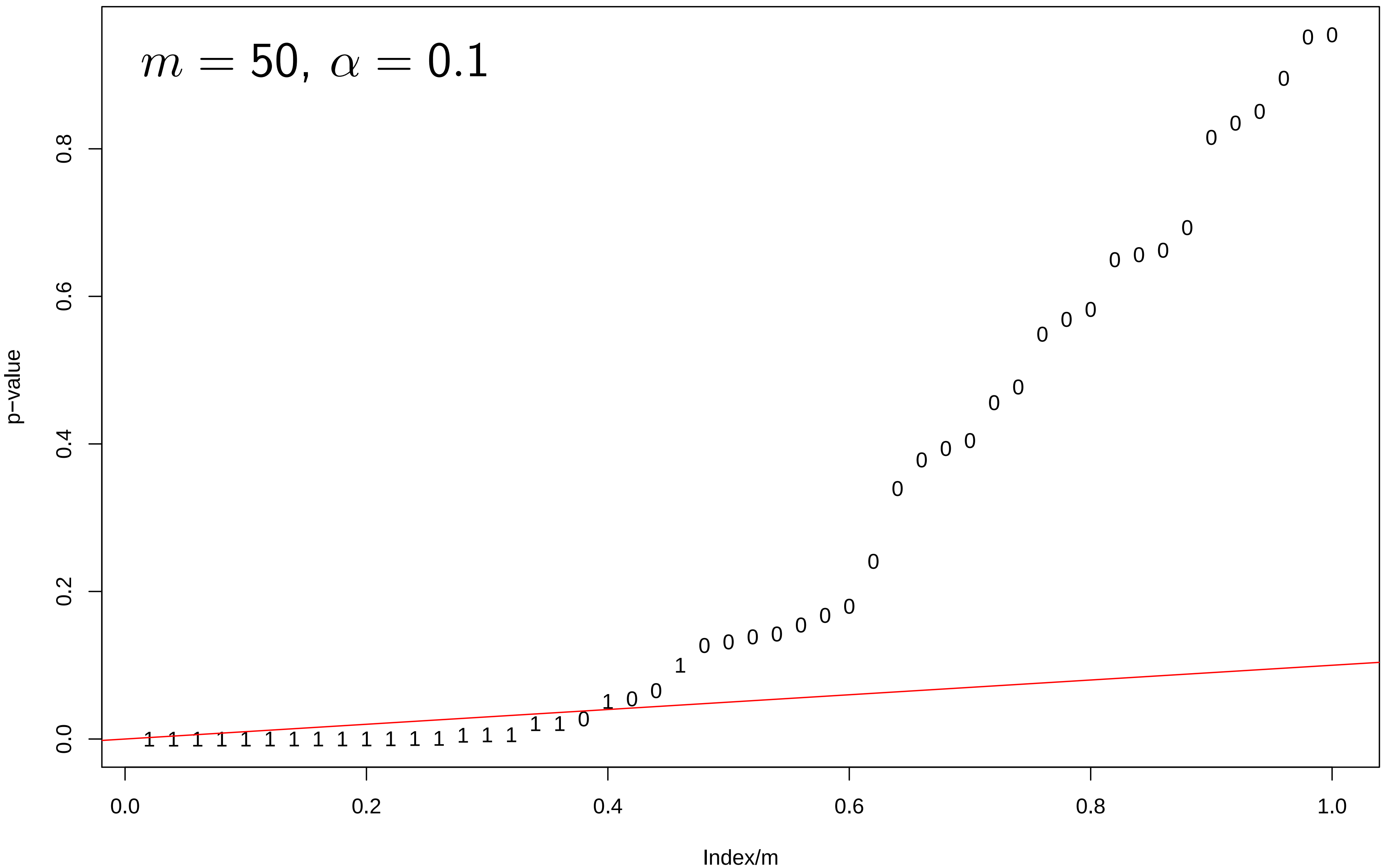
$$T_{\mathrm{BH}} = \max \left\{ P_{(i)} \colon \; P_{(i)} \leq \alpha \frac{i}{m},\ 0 \leq i \leq m \right\}.$$

- BH (1995) proved (for independent tests) that using this procedure guarantees – *for any alternative distributions* – that

$$\mathrm{FDR} \equiv \mathbb{E}\left(\mathrm{FDP}(T_{\mathrm{BH}})\right) \leq \frac{T_0}{m}\,\alpha,$$

and equality holds with continuous test statistics.

The Benjamini-Hochberg Procedure (cont'd)



Plan

1. The Multiple Testing Problem
   - Error Criteria and Power
   - False Discovery Control and the BH Method
2. Why BH Works
   - A Useful Model
   - A Stochastic Process Perspective
   - Performance Characteristics
3. Toward False Discovery Control: Variations on BH
   - Improving Power
   - Dependence
   - Alternative Formulations
4. Exceedance Control and Random Fields

A Useful Mixture Model

- The following model is helpful for understanding and analyzing BH and its variants:

$$
\begin{aligned}
H_1, \ldots, H_m &\text{ iid Bernoulli}\langle a \rangle \\
\Xi_1, \ldots, \Xi_m &\text{ iid } \mathcal{L}_{\mathcal{F}} \\
P_i \mid H_i = 0, \Xi_i = \xi_i &\sim \text{Uniform}\langle 0, 1 \rangle \\
P_i \mid H_i = 1, \Xi_i = \xi_i &\sim \xi_i.
\end{aligned}
$$

  where $\mathcal{L}_{\mathcal{F}}$ denotes a probability distribution on a class $\mathcal{F}$ of distributions on $[0, 1]$.

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

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

A Useful Mixture Model (cont'd)

- Under this model, the m p-values $P^m = (P_1, \ldots, P_m)$ are *marginally* IID from

$$G = (1 - a)U + aF,$$

where: 1. $0 \leq a \leq 1$ is the frequency of alternatives,

2. U is the Uniform$\langle 0, 1 \rangle$ cdf, and

3. $F = \int \xi \, d\mathcal{L}_{\mathcal{F}}(\xi)$ is a distribution on [0,1].

- The marginal alternative distribution F comes up again and again, but its use does not preclude having different alternatives for different tests.
- Although the model posits IID Bernoulli$\langle a \rangle$ H_is, all the theory carries through with fixed H_is as well.

BH Revisited

Let's use this model to understand FDR and BH.

At any fixed threshold t, we have

$$\begin{aligned}
\mathsf{FDR}(t) &= \mathbb{E}\,\frac{\sum_i 1\{P_i \le t\}(1-H_i)}{\sum_i 1\{P_i \le t\} + 1\{\text{all } P_i > t\}} \\
&\approx \frac{\mathbb{E}\,\frac{1}{m}\sum_i 1\{P_i \le t\}(1-H_i)}{\mathbb{E}\,\frac{1}{m}\sum_i 1\{P_i \le t\} + \frac{1}{m}\mathbb{P}\{\text{all } P_i > t\}} \\
&= \frac{(1-a)t}{G(t) + \frac{1}{m}(1-G(t))^m} \approx \frac{(1-a)t}{G(t)}.
\end{aligned}$$

BH Revisited (cont'd)

Now, let

$$\hat{G}_m(t) = \frac{1}{m} \sum_i 1\Big\{P_i \le t\Big\}$$

be the empirical cdf of P^m.

In the continuous case, we can ignore ties, so $\hat{G}_m(P_{(i)}) = \frac{i}{m}$.

BH is thus equivalent to the following:

$$\begin{aligned} T_{\mathrm{BH}}(P^m) &= \sup\left\{t\colon \ t \le \alpha \hat{G}_m(t)\right\} \\ &= \sup\left\{t\colon \ \hat{G}_m(t) = \frac{t}{\alpha}\right\} \\ &= \sup\left\{t\colon \ \frac{t}{\hat{G}_m(t)} = \alpha\right\}. \end{aligned}$$

BH Revisited (cont'd)

One can think of this in two ways.

First, the BH procedure equates estimated FDR to the target α.

This estimator,

$$\widehat{\mathsf{FDR}}(t) = \frac{t}{\widehat{G}_m(t)},$$

uses $\widehat{G}_m$ in place of G and $\widehat{a} \equiv 0$ in place of a.

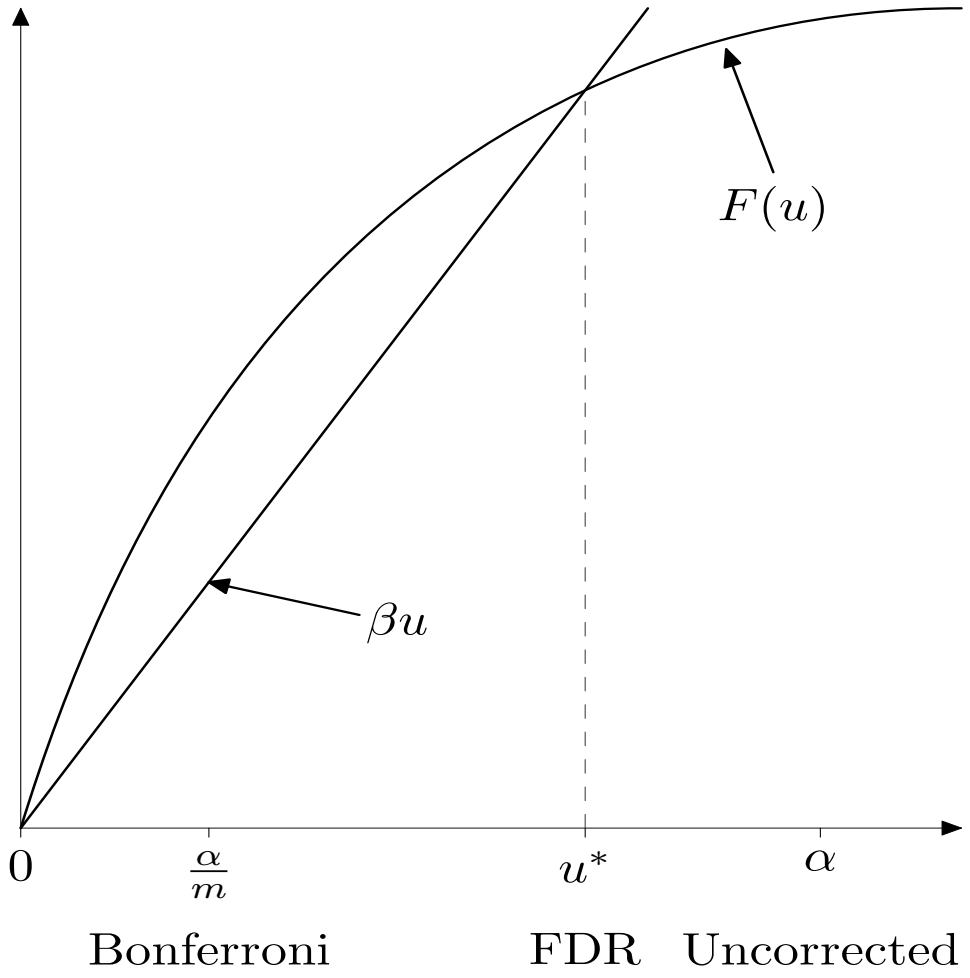
Second, the BH threshold is a plug-in estimator of

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

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

Asymptotic Behavior of BH Procedure

This yields the following picture:



BH Performance

- BH generally gives more power than FWER control and fewer Type I errors than uncorrected testing.

- BH performs best in very sparse cases ($T_0 \approx m$).

  For example, under the mixture model and in the continuous case,

  $$\mathbb{E}(\text{FDP}(T_{\text{BH}})) = (1 - a)\alpha.$$

  The BH procedure thus *overcontrols* FDR and thus will not in general minimize FNR.

- Power can be improved in non-sparse cases by more complicated adaptive procedures.

BH Performance (cont'd)

- When all m null hypotheses are true, BH is equivalent to FWER control.

- The BH FDR bound holds for certain classes of dependent tests, as we will see.

  In practice, it is quite hard to "break".

- $D \cdot \alpha$ need not bound the number of false discoveries.

  This is a common misconception for end users.

Operating Characteristics of the BH Method

- Define the misclassification risk of a procedure T by

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

  This is the average fraction of errors of both types.

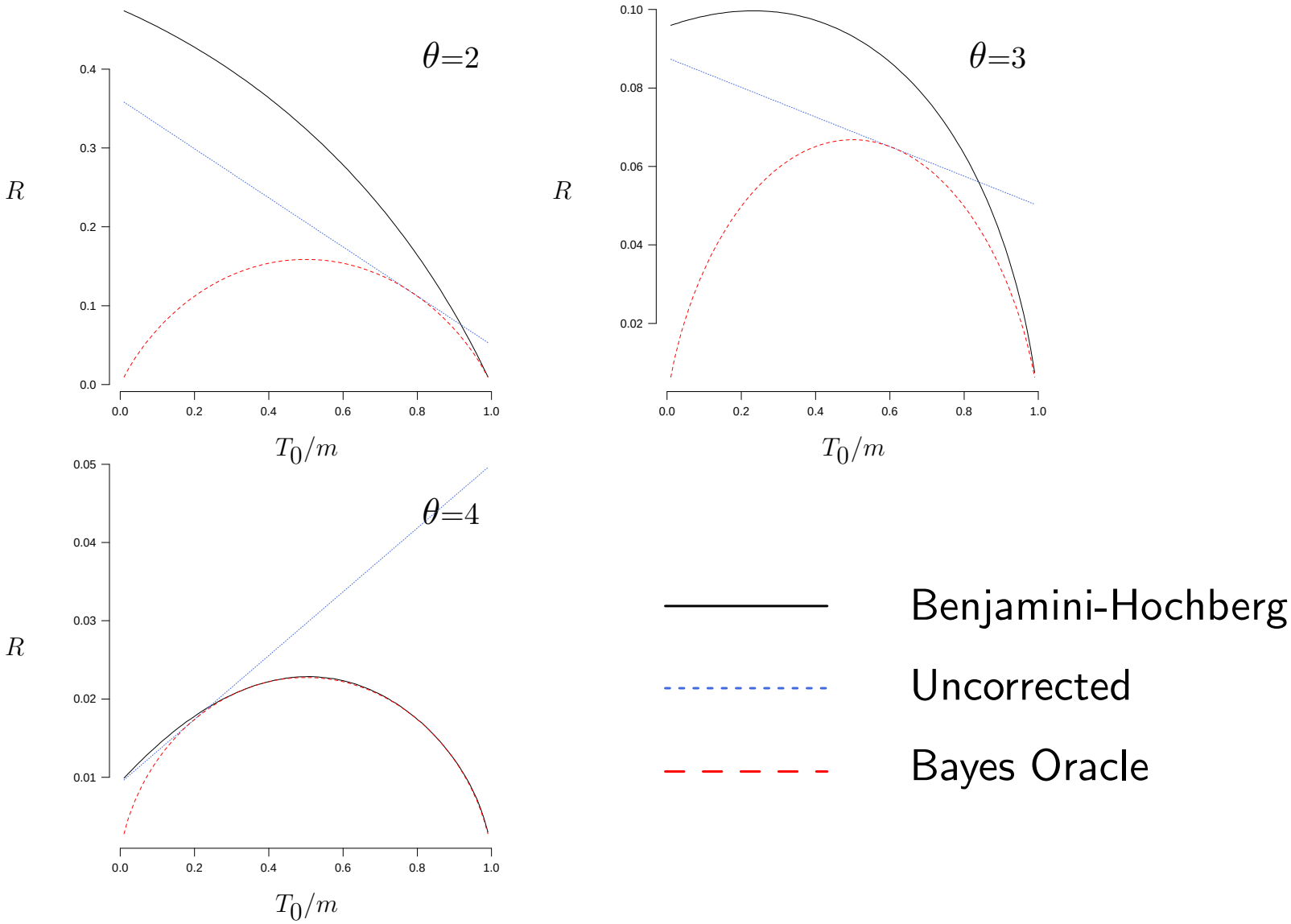
- Then $R_M(T_{\mathrm{BH}}) \sim R(a, F)$ as $m \to \infty$, where

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

- 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, $\alpha = 0.05$



FDP and FNP as Stochastic Processes

- Both the FDP(t) and FNP(t) stochastic processes converge to Gaussian processes outside a neighborhood of 0 and 1 respectively.
- For example, define

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

  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, $Z_m \rightsquigarrow Z$.

Plan

1. The Multiple Testing Problem
   - Error Criteria and Power
   - False Discovery Control and the BH Method
2. Why BH Works
   - A Useful Model
   - A Stochastic Process Perspective
   - Performance Characteristics
3. Toward False Discovery Control: Variations on BH
   - Improving Power
   - Dependence
   - Alternative Formulations
4. Exceedance Control and Random Fields

Optimal Thresholds

- The equality

$$\mathbb{E}(\mathsf{FDP}(T_{\mathrm{BH}})) = (1-a)\alpha$$

  implies that if we knew a, we could improve power by applying BH at level $\alpha/(1-a)$.

- This suggests using T_{PI}, the plug-in estimator for

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

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

- Note that $t^* \geq u^*$.

Optimal Thresholds (cont'd)

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

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

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

- Ignore $O()$ terms and choose t to minimize $\mathbb{E}(\mathsf{FNP}(t))$ subject to $\mathbb{E}(\mathsf{FDP}(t)) \leq \alpha$.

  This yields $t^*(a, G)$ as the optimal threshold.

- Genovese and Wasserman (2002) show that

$$\mathbb{E}(\mathsf{FDP}(t^*(\widehat{a}, \widehat{G}))) \leq \alpha + O(m^{-1/2})$$

  under weak conditions on $\widehat{a}$.

Improving Power

- In practice, the main difficulty here is finding a good estimator of $1 - a$, or alternatively, a good estimator of T_0.

  Part of the challenge is guaranteeing FDR control with the increased variability induced by the estimator.

- Adaptive estimators for improving power in FWER-controlling methods go back to Schweder and Spjotvol (1982) and Hochberg and Benjamini (1990).

- Recent approaches in the context of FDR have come from Benjamini and Hochberg (2000), Efron et al. (2001), Storey (2002), Genovese and Wasserman (2002), Storey et al. (2003), and Benjamini, Krieger, and Yekutieli (BKY, 2004).

Improving Power (cont'd)

- Benjamini, Krieger, and Yekutieli (BKY, 2004) give a comprehensive numerical comparison of adaptive procedures and introduce new procedures, with an elegant proof of FDR control.
- Their two stage method is as follows:
  - Use BH at level β_1. Let r_1 be the number of rejected null hypotheses.
  - If $r_1 = 0$, stop.
  - Otherwise, let $\widehat{T}_0 = m - r_1$.
  - Use BH at level $\alpha' = \beta_2 m / \widehat{T}_0$.
- The initial procedure takes $\beta_1 = \beta_2 = \alpha/(1+\alpha)$, but they also have success with $\beta_1 = \alpha$ and $\beta_2 = \alpha/(1+\alpha)$.
- This method has good power and remains valid under certain kinds of dependence, as we will see.

Improving Power (cont'd)

- Genovese, Roeder, and Wasserman (2004) give an alternative way to increase power: a priori p-value weighting.
- For instance, if we define prior weights $W_1, \ldots, W_m > 0$, we can define a weighted-BH (wBH) threshold

$$T_{\text{wBH}} = \sup\{t : \ \widehat{R}(t) \leq \alpha\},$$

where

$$\widehat{R}(t) = \frac{t \sum_{i=1}^m W_i}{\sum_{i=1}^m 1\Big\{P_i \leq W_i t\Big\}} = \frac{t\overline{W}}{\widehat{D}(t)}.$$

where $\widehat{D}$ is the EDF of the $Q_i = P_i/W_i$. When all $W \equiv 1$, get BH.

- Main result: if weights are positively associated with the null being false, power improves (unless already very near 1).

  Even weights are poorly chosen, power is only reduced slightly, as long as weightsare not too large.

Dependence

- BH procedure still controls FDR at nominal level for some dependent tests (Benjamini and Yekutieli, 2001).

  In particular, this holds under *positive regression dependence on a subset*.

- Under general dependence structure, the BH method controls FDR at level

$$\alpha \frac{T_0}{m} \sum_{i=1}^{m} \frac{1}{i}.$$

  Distribution-free procedure: Apply BH at level $\alpha / \sum_{i=1}^{m} \frac{1}{i}$. Typically very conservative.

- In practice, simulation studies suggest BH is quite hard to "break".

Dependence (cont'd)

- The challenge of dependence for adaptive procedures is finding an estimator of $1 - a$ (or T_0) that performs well under various dependence structures.

  This turns out to be far from easy.

- BKY (2004) show that their two stage procedure continues to control FDR under positive dependence.

  They argue, convingingly, that this is the best option when the degree of dependence is unknown.

- There are also advantages to be explored in using the estimated dependence structure itself to improve performance.

pFDR and Bayesian Connections

- Storey (2001) considers the "positive FDR," defined by

$$\text{pFDR}(t) = \mathbb{E}\left(\text{FDP}(t) \,\middle|\, D(t) > 0\right).$$

  Note that $\text{FDR}(t) = \text{pFDR}(t) \cdot \mathbb{P}\Big\{D(t) > 0\Big\} \leq \text{pFDR}(t)$.

- Storey (2001) makes a nice Bayesian connection.

  Taking a under the mixture model to be the prior probability that a null hypothesis is false, it follows taht

$$\text{pFDR}(t) = \frac{(1-a)t}{G(t)} = \frac{(1-a)\mathbb{P}\Big\{P \leq t \mid H = 0\Big\}}{\mathbb{P}\Big\{P \leq t\Big\}} = \mathbb{P}\Big\{H = 0 \mid P \leq t\Big\}.$$

- Storey (2003) also introduces the *q-value* as the minimum pFDR for which the given statistic is rejected.

  This has a Bayesian interpretation as a "posterior Bayesian p-value".

Empirical Bayes Testing

- Efron et al (2001) construct an empirical Bayes measure of "local FDR". They note that

$$\mathbb{P}\Big\{H_i = 0 \mid P^m\Big\} = \frac{(1-a)}{g(P_i)} \equiv q(P_i),$$

  where $g = G'$.

- This suggests a rejection rule $q(p) \leq \alpha$, but we need to estimate q, e.g., $\widehat{q}(p) = \frac{1-\widehat{a}}{\widehat{g}(p)}$.
- Under weak conditions, can show that

$$q(t) \leq \alpha \;\text{ implies }\; \frac{(1-a)t}{G(t)} \equiv \mathbb{E}(\text{FDR}) \leq \alpha$$

  So EBT is conservative and performance depends on behavior of $\widehat{q}$.

Exceedance Control

- Genovese and Wasserman (2002, 2004) introduce the idea of "exceedance control" where we bound $\mathbb{P}\{\mathsf{FDP} > \gamma\}$ rather than $\mathsf{FDR} \equiv \mathbb{E}\,\mathsf{FDP}$.

- van der Laan, Dudoit, and Pollard (2004) introduce FDR and exceedance controlling procedures based on "augmenting" a FWER-controlling test.

- Motivates the term "False Discovery Control" since we're no longer just controlling FDR.

Plan

1. The Multiple Testing Problem
   - Error Criteria and Power
   - False Discovery Control and the BH Method
2. Why BH Works
   - A Useful Model
   - A Stochastic Process Perspective
   - Performance Characteristics
3. Toward False Discovery Control: Variations on BH
   - Improving Power
   - Dependence
   - Alternative Formulations
4. Exceedance Control and Random Fields

Objective

- Want a procedure to control the exceedance False Discovery Proportion (FDP):

$$\mathbb{P}\left\{\frac{\text{False Discoveries}}{\text{Discoveries}} > \gamma\right\} \leq \alpha \qquad \text{for } 0 < \alpha, \gamma < 1,$$

- This allows tuning of the inference to account for variability in the FDP distribution.
- Also useful as a basis for secondary inference about patterns of false discovery. Examples:
  - Controlling proportion of false regions (rather than pixels) in spatial/image problems.
  - Scan statistics for finding "hot spots" in random fields.
- Also useful as an FDR diagnostic.

Rejection Sets and the FDP

- Let $S = \{1, \ldots, m\}$ and let $S_0 = \{j \in S\colon\ H_j = 0\}$.
- Call a rejection set any $R \equiv R(P^m) \subset S$ that indexes the set of *rejected* null hypotheses.

  The prototypical rejection set is defined by a threshold: $R_T = \{j \in S\colon\ P_j \leq T\}$.
- A rejection set is another way to represent a multiple testing procedure. That is, if $\mathbb{P}\Big\{\#(R \cap S_0) > 0\Big\} \leq \alpha$, then R controls FWER at level α.
- Similarly, we can defined the FDP for any such procedure:

$$\text{FDP}(R) = \frac{\text{false rejections}}{\text{rejections}} = \frac{\sum_{j=1}^{m}(1 - H_j)1\Big\{R \ni j\Big\}}{\sum_{j=1}^{m} 1\Big\{R \ni j\Big\}},$$

  where the ratio is defined to be zero if the denominator is zero.

Confidence Envelopes

- Our main tools for exceedance control are *confidence envelopes*.

  A $1-\alpha$ confidence envelope for FDP is a random function $\overline{\mathsf{FDP}}(C) \equiv \overline{\mathsf{FDP}}(C; P_1, \ldots, P_m)$ such that

$$\mathbb{P}\Big\{\overline{\mathsf{FDP}}(C) \geq \mathsf{FDP}(C), \text{ for all } C\Big\} \geq 1-\alpha.$$

- If we take the largest rejection set R such that $\overline{\mathsf{FDP}}(R) \leq \gamma$, then,

$$\mathbb{P}\Big\{\mathsf{FDP}(R) \leq \gamma\Big\} \geq \mathbb{P}\Big\{\mathsf{FDP} \leq \overline{\mathsf{FDP}}\Big\} \geq 1-\alpha,$$

  so we have controlled the FDP exceedance at the target levels.

Confidence Envelopes (cont'd)

In terms of thresholds, a $1-\alpha$ confidence envelope for FDP satisfies

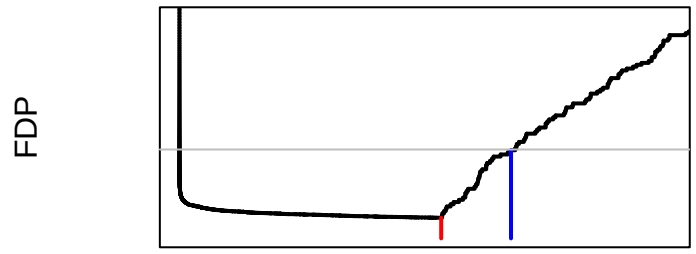
$$\mathbb{P}\Big\{\mathsf{FDP}(t) \le \overline{\mathsf{FDP}}(t) \text{ for all } t\Big\} \ge 1-\alpha.$$

With this, we can construct thresholds that give
$\mathbb{P}\Big\{\mathsf{FDP}(T) \le \gamma\Big\} \ge 1-\alpha$.

Two special cases have proven useful:

– *Fixed-ceiling:* $T = \sup\{t\colon\ \overline{\mathsf{FDP}}(t) \le \alpha\}$.

– *Minimum-envelope:* $T = \sup\{t\colon\ \overline{\mathsf{FDP}}(t) = \min_t \overline{\mathsf{FDP}}(t)\}$.



Inversion Construction: Main Idea

- Construct confidence envelope by inverting a set of uniformity tests.
- Specifically, consider all subsets of the p-values that cannot be distinguished from a sample of Uniforms by a suitable level α test.
- Consider each of these subsets as one configuration of true nulls.
- Maximize FDP pointwise over these configurations.

Inversion Construction: Step 1

For every $W \subset S$, test at level α the hypothesis that

$$P_W = (P_i\colon\ i \in W)$$

is a sample from a Uniform$(0, 1)$ distribution:

$$H_0 : W \subset S_0 \quad \text{versus} \quad H_1 : W \not\subset S_0.$$

Formally, let $\Psi = \{\psi_W\colon\ W \subset S\}$ be a set of non-randomized tests such that

$$\mathbb{P}\Big\{\psi_W(U_1, \ldots, U_{\#(W)}) = 1\Big\} \leq \alpha$$

whenever $U_1, \ldots, U_{\#(W)} \leftarrow$ Uniform$(0, 1)$.

Inversion Construction: Step 2

Let $\mathcal{U}$ denote the collection of all subsets W not rejected in the previous step:

$$\mathcal{U} = \{W\colon\ \psi_W(P_W) = 0\}.$$

Now define

$$\overline{\mathsf{FDP}}(C) = \begin{cases} \max\limits_{B \in \mathcal{U}} \dfrac{\#(B \cap C)}{\#(C)} & \text{if } C \neq \emptyset, \\ 0 & \text{otherwise.} \end{cases}$$

If $\mathcal{U}$ is closed under unions, then

$$\overline{\mathsf{FDP}}(C) = \frac{\#(U \cap C)}{\#(C)}$$

where $U = \cup\{V\colon\ V \in \mathcal{U}\}$. This is a *confidence superset* for S_0:

$$\mathbb{P}\Big\{S_0 \subset U\Big\} \geq 1 - \alpha.$$

Inversion Construction: Step 3

Choose $R = R(P_1, \ldots, P_m)$ as large as possible such that

$$\overline{\mathsf{FDP}}(R) \leq \gamma.$$

(Typically, take R of the form $R = \left\{j\colon\ P_j \leq T\right\}$ where the *confidence threshold* $T = \sup\{t :\ \overline{\mathsf{FDP}}(t) \leq c\}$.)

It follows that

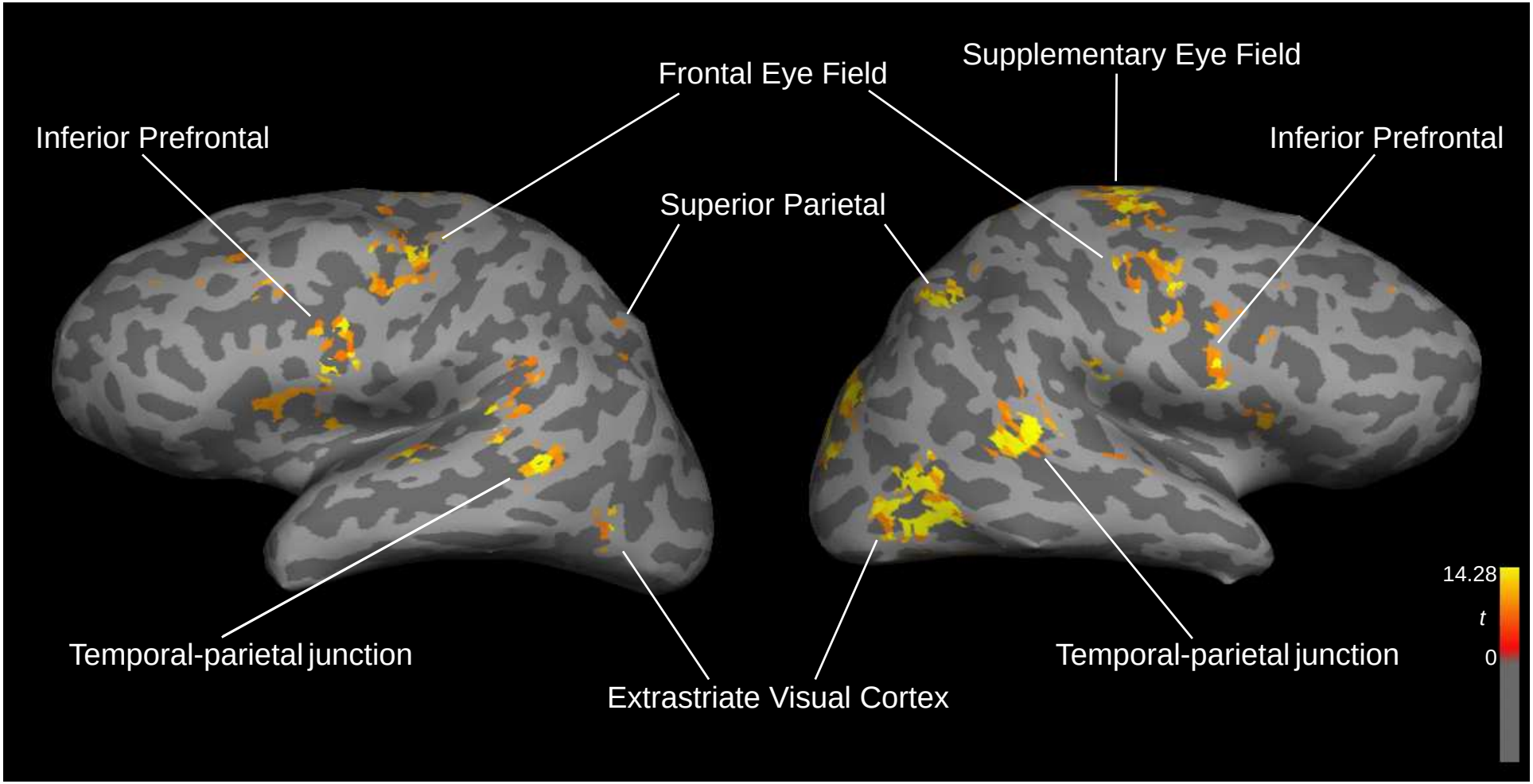
1. $\overline{\mathsf{FDP}}$ is a $1 - \alpha$ confidence envelope for FDP, and
2. R is a (γ, α) exceedance-controlling rejection set.

Note: Can also calibrate this procedure to control FDR.

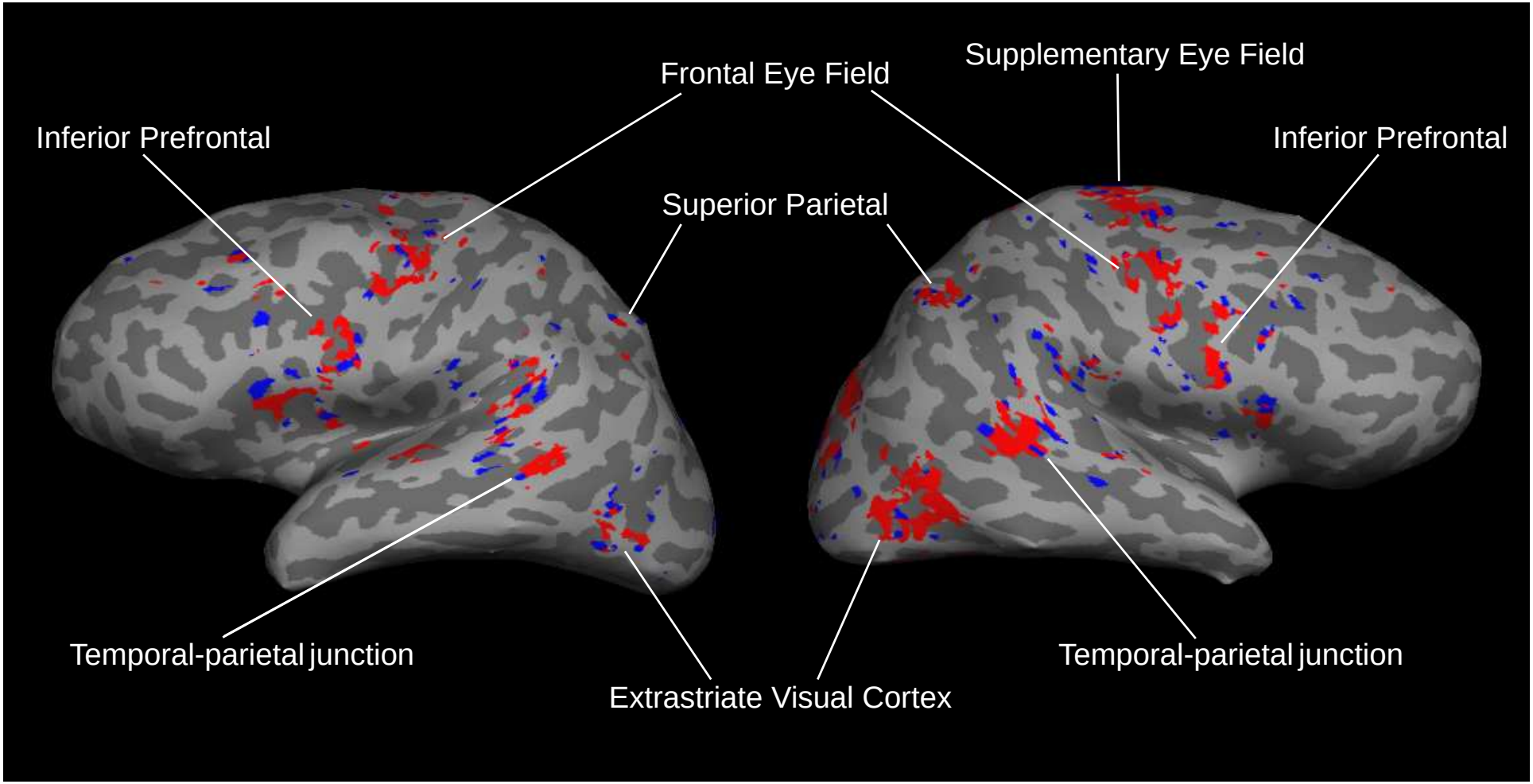
Choice of Tests

- The confidence envelopes depend strongly on choice of tests.
- Two desiderata for selecting uniformity tests:
  A. (Power). The envelope $\overline{\text{FDP}}$ should be close to FDP and thus result in rejection sets with high power.
  B. (Computational Tractability). The envelope $\overline{\text{FDP}}$ should be easy to compute.
- Traditional uniformity tests, such as the (one-sided) Kolmogorov-Smirnov (KS) test, do not usually meet both conditions.
- One good approach is to combine uniformity tests based on the kth order statistic.

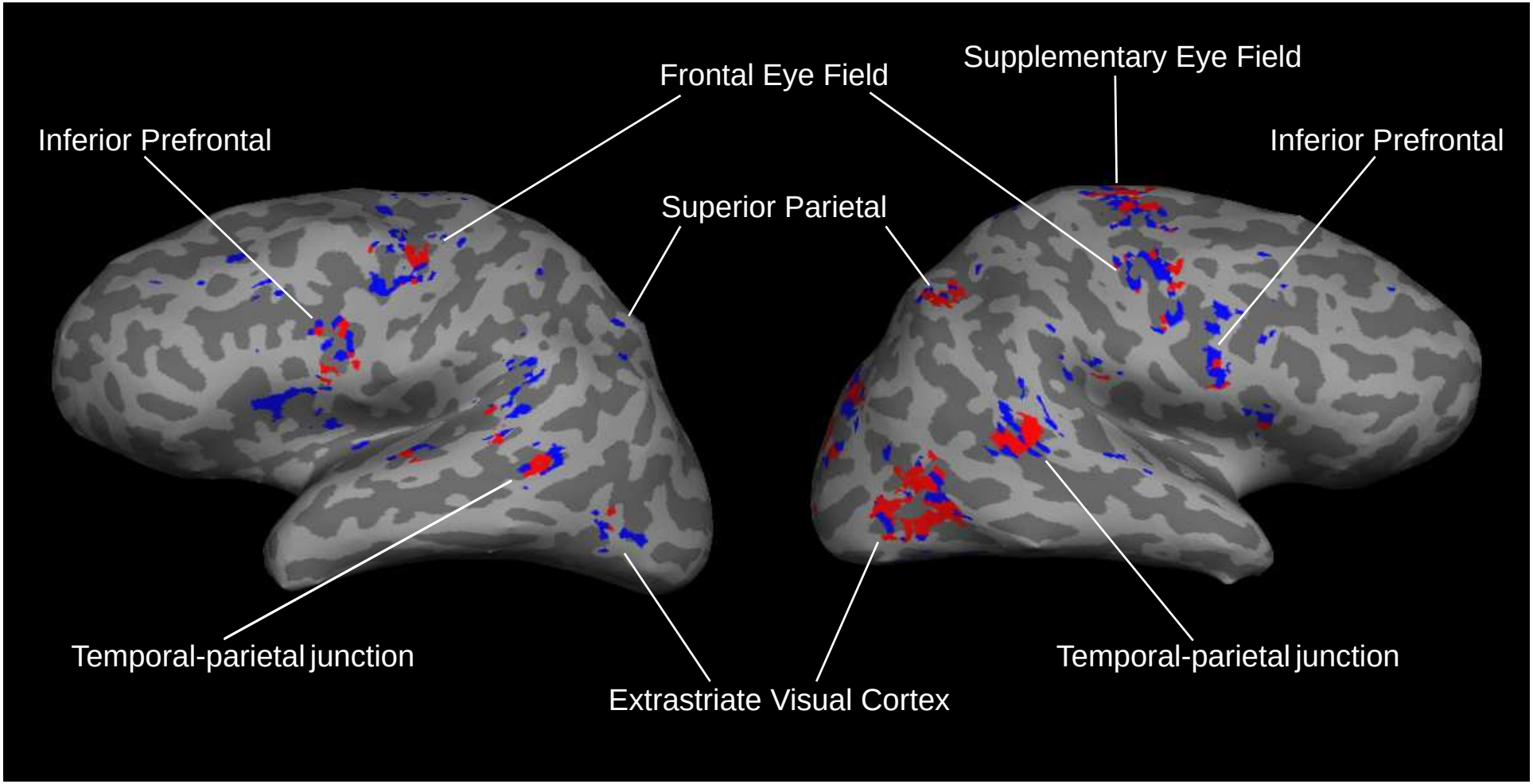
Results: (0.05,0.9) Confidence Threshold



Results: (0.05,0.9) Threshold versus BH



Results: (0.05,0.9) Threshold versus Bonferroni



False Discovery Control for Random Fields

- Multiple testing methods based on the excursions of random fields are widely used, especially in functional neuroimaging (e.g., Cao and Worsley, 1999) and scan clustering (Glaz, Naus, and Wallenstein, 2001).

- False Discovery Control extends to this setting as well.

- For a set S and a random field $X = \{X(s)\colon\ s \in S\}$ with mean function $\mu(s)$, use the realized value of X to test the collection of one-sided hypotheses

$$H_{0,s} : \mu(s) = 0 \text{ versus } H_{1,s} : \mu(s) > 0.$$

Let $S_0 = \{s \in S :\ \mu(s) = 0\}$.

False Discovery Control for Random Fields

- Define a spatial version of FDP for threshold T by

$$\mathsf{FDP}(T) = \frac{\lambda(S_0 \cap \{s \in S : X(s) \geq t\})}{\lambda(\{s \in S : X(s) \geq t\})},$$

  where λ is usually Lebesgue measure.

- As before, we can control FDR or FDP exceedance.

- Our approach is again based on the inversion method for constructing a confidence envelope for FDP.

Controlling the Proportion of False Regions

- Say a region R is false at tolerance ϵ if more than an ϵ proportion of its area is in S_0:

$$\frac{\lambda(R \cap S_0)}{\lambda(R)} \geq \epsilon.$$

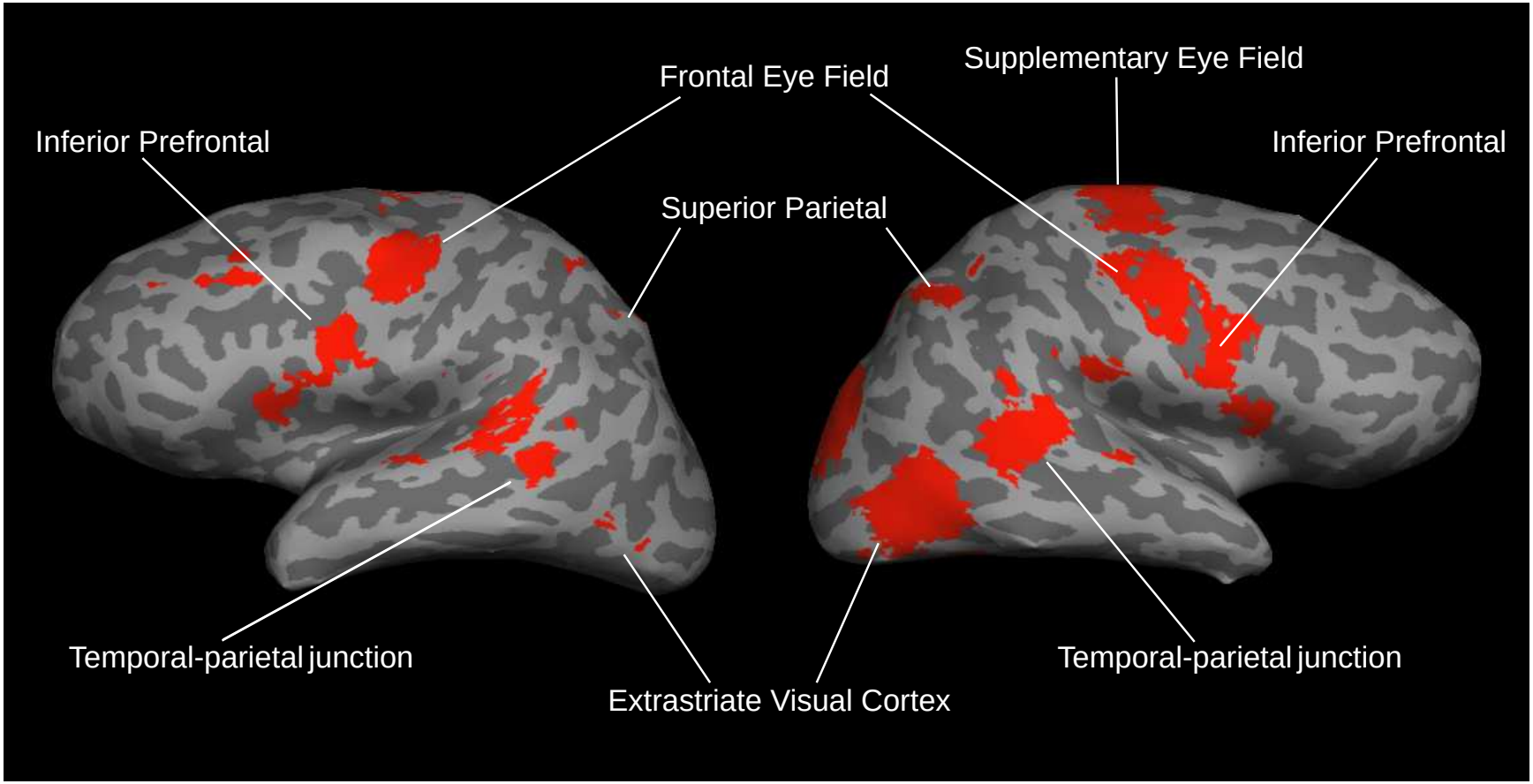
- Decompose the t-level set of X into its connected components $C_{t1}, \ldots, C_{tk_t}$.
- For each level t, let $\xi(t)$ denote the *proportion of false regions (at tolerance ϵ)* out of k_t regions.
- Then,

$$\overline{\xi}(t) = \frac{\#\left\{1 \leq i \leq k_t : \ \frac{\lambda(C_{ti} \cap U)}{\lambda(C_{ti})} \geq \epsilon\right\}}{k_t}$$

gives a $1 - \gamma$ confidence envelope for ξ.

Results: False Region Control Threshold

$\mathbb{P}\{\text{prop'n false regions} \leq 0.1\} \geq 0.95$ where false means null overlap $\geq 10\%$



Scan Statistics

Let $X = (X_1, \ldots, X_N)$ be a realization of a point process with intensity function $\nu(s)$ defined on a compact set $S \subset \Re^d$. Assume that $\nu(s) = \nu_0$ on $S_0 \subset S$ and $\nu(s) > \nu_0$ otherwise.

Assume that conditional on $N = n$, X is an IID sample from the density

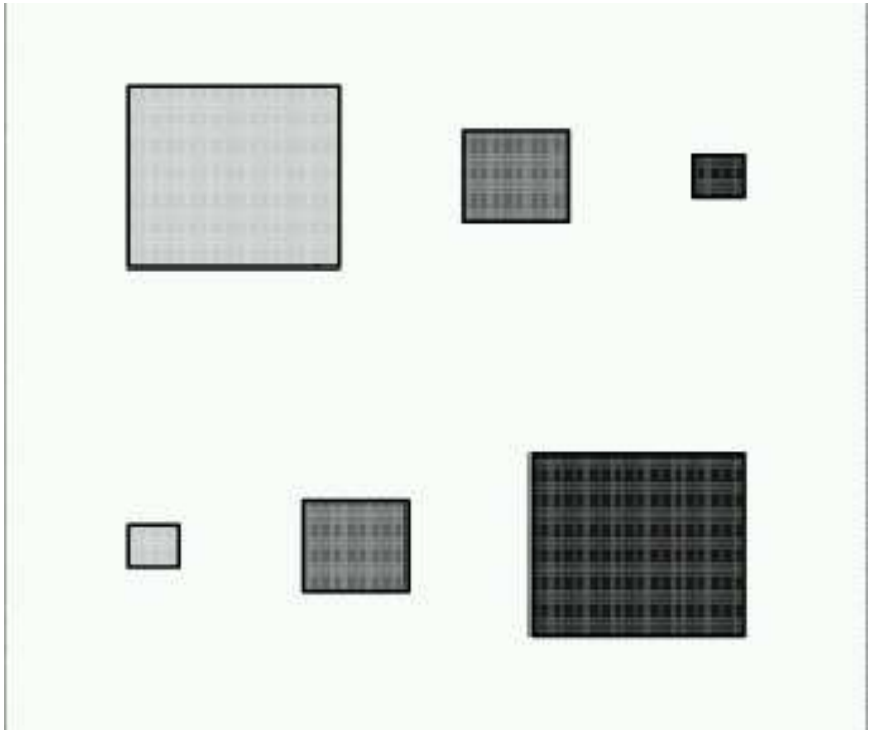
$$f(s) = \frac{\nu(s)}{\int_S \nu(u)\, du}.$$

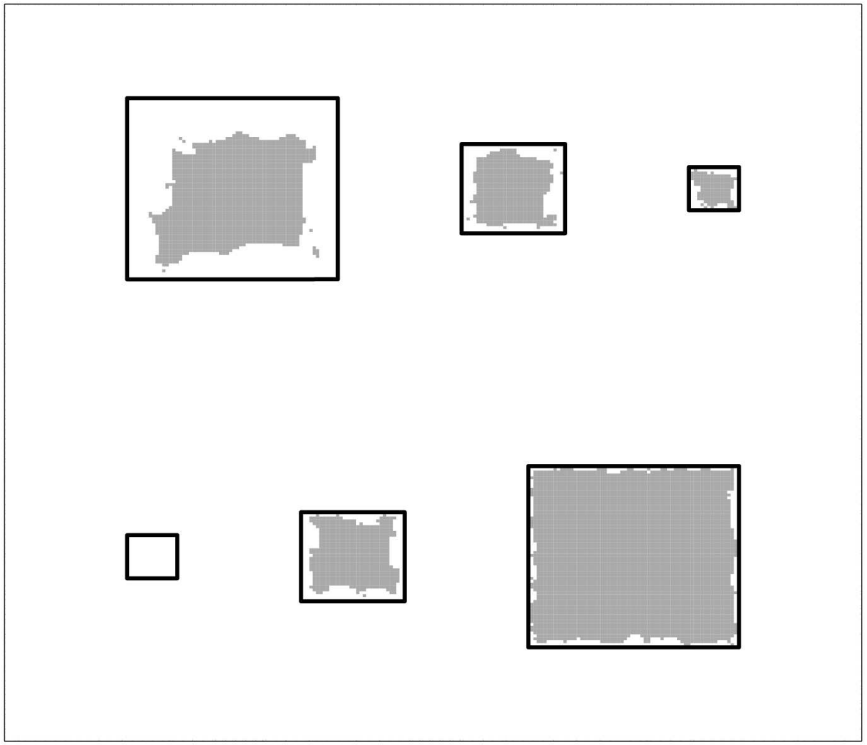
Scan statistic test for "clusters" via the statistic $T = \sup_{s \in S} N_s$.,

Our procedure:

1. Kernel estimators $\widehat{f}_H$ with a set of bandwidths $\mathcal{H}$.
2. Bias adjustment
3. False Discovery Control

Scan Statistics (cont'd)





Plan

1. The Multiple Testing Problem
   - Error Criteria and Power
   - False Discovery Control and the BH Method
2. Why BH Works
   - A Useful Model
   - A Stochastic Process Perspective
   - Performance Characteristics
3. Toward False Discovery Control: Variations on BH
   - Improving Power
   - Dependence
   - Alternative Formulations
4. Exceedance Control and Random Fields

Take-Home Points

- False Discovery Control provides a useful alternative to traditional multiple testing methods.
- The BH method is fast and robust, but it overcontrols FDR. Good adaptive methods exist that can increase power (e.g., BKY 2004).
- Exceedance control has practical advantages.

  In particular, gives a tunable inferential guarantee without too much loss of power.

  Works under general dependence and can be used to make inferences about the pattern of false discoveries.
- Important open problems include explicitly accounting for dependence and taking advantage of spatial structure in the alternatives.

Selected References

Abramovich, F., Benjamini, Y., Donoho, D. and Johnstone, I. (2000). Adapting to unknown sparsity by controlling the false discovery rate. Technical report 2000-19. Department of Statistics. Stanford University.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57, 289-300.

Benjamini, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational Behavior, Statistics*, **25**, 60–83.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29, 1165-1188.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96, 1151-1160.

Finner, H. and Roters, M. (2002). Multiple hypotheses testing and expected number of type I errors. *Annals of Statistics*, 30, 220–238.

Selected References (cont'd)

Genovese, C. R. and Wasserman, L. (2001). False discovery rates. Technical Report, Department of Statistics, Carnegie Mellon.

Genovese, C. R. and Wasserman, L. (2002). Operating characteristics and extensions of the FDR procedure. *Journal of the Royal Statistical Society B*, **64**, 499–518.

Genovese, C. R. and Wasserman, L. (2003). A stochastic process approach to false discovery control. *Annals of Statistics*, in press.

Harvönek & Chytil (1983). Mechanizing hypotheses formation – a way for computerized exploratory data analysis? *Bulletin of the International Statistical Institution*, **50**, 104–121.

Helperin, M., Lan, G.K.K., and Hamdy, M.I. (1988). Some implications of an alternative definition of the multiple comparison problem. *Biometrika*, **75**, 773–778.

Hochberg, Y. and Benjamini, Y. (1990). More powerful procedures for multiple significance testing. *Statistics in Medicine*, **9**, 811–818.

Hommel, G. and Hoffman, T. (1987) Controlled uncertainty. In P. Bauer, G. Hommel, and E. Sonnemann, (Eds.), *Multiple hypothesis testing* (pp. 154–161). Heidelberg: Springer.

Selected References (cont'd)

Sarkar, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *Annals of Statistics*, 30, 239–257.

Schweder, T. and Spjøtvoll, E. (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika*, **69**, 493–502.

Simes, J. R. 1986. An improved Bonferroni procedure for multiple tests of significance. Biometrika 73: 75–754.

Shafer, J. (1995). Multiple hypothesis testing. *Annual Reviews in Psychology*, **46**:561–584.

Storey, J. D. (2001). The positive False Discovery Rate: A Bayesian interpretation and the q-value. *Annals of Statistics*, in press.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society B*, **64**, 479–498.

Storey, J.D., Taylor, J. E., and Siegmund, D. (2002). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society B*, in press.

Yekutieli, D. and Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, **82**, 171–196.

Appendix

Improving Power (cont'd)

- Benjamini and Hochberg (2000) introduced the idea of using the BH procedure to estimate T_0.
  - Use BH at level α. If no rejections, stop.
  - Otherwise, define $\widehat{T}_{0,k} = \dfrac{m+1-k}{1-P_{(k)}}$, for $k = 1, \ldots, m$.
  - Find first $k^* \geq 2$ such that $\widehat{T}_{0,k} > \widehat{T}_{0,k-1}$.
  - Estimate $\widehat{T}_0 = \min(m, \lceil \widehat{T}_{0,k^*} \rceil)$.
  - Use BH at level $\alpha' = \alpha m / \widehat{T}_0$.
- Here, the intermediate estimators $\widehat{T}_{0,k}$ are derived from the number of rejections at fixed threshold $P_{(k)}$, adjusted for the expected $T_0 \cdot P_{(k)}$ false rejections.
- This procedure controls FDR and has good power under independence.

Improving Power (cont'd)

- Storey (2002) gave an alternative adaptive procedure that uses

$$\widehat{1-a} = \frac{1-\widehat{G}(\lambda)}{1-\lambda},$$

for some fixed λ, often $\lambda = 1/2$. The rationale for this estimator is that most of the p-values near 1 should be null, implying $1 - G(\lambda) \approx (1-a)(1-\lambda)$.

- Storey et al. (2003) modified this estimator for theoretical reasons to

$$\widehat{1-a} = \frac{1+\frac{1}{m}-\widehat{G}(\lambda)}{1-\lambda},$$

with the proviso that only nulls with $P_{(i)} \leq \lambda$ can be rejected.

- With this modification, this procedure tends to have higher power than BH2000 *under independence*.

Improving Power (cont'd)

- Genovese and Wasserman (2002) show that this procedure controls FDR asymptotically.

  Storey et al. (2003) show by a nice martingale argument that it controls FDR for a finite number of independent tests.

  They also extended it to a particular form of dependence among the tests.

- Efron et al. (2001) considered a variant with λ set to the median p-value.

  This was motivated primarily toward computing their empirical Bayes local FDR.

Inversion for Random Fields: Details

1. For every $A \subset S$, test $H_0 : A \subset S_0$ versus $H_1 : A \not\subset S_0$ at level γ using the test statistic $X(A) = \sup_{s \in A} X(s)$.

   The tail area for this statistic is $p(z, A) = \mathbb{P}\Big\{X(A) \geq z\Big\}$.

2. Let $\mathcal{U} = \{A \subset S\colon \ p(x(A), A) \geq \gamma\}$.

3. Then, $U = \bigcup_{A \in \mathcal{U}} A$ satisfies $\mathbb{P}\Big\{U \supset S_0\Big\} \geq 1 - \gamma$.

4. And,
$$\overline{\mathsf{FDP}}(t) = \frac{\lambda(U \cap \{s \in S : X(s) > t\})}{\lambda(\{s \in S : X(s) > t\})},$$
   is a confidence envelope for FDP.

Note: We need not carry out the tests for all subsets.

Gaussian Fields

- With Gaussian Fields, our procedure works under similar smoothness assumptions as familywise random-field methods.
- For our purposes, approximation based on the expected Euler characteristic of the field's level sets will not work because the Euler characteristic is non-monotone for non-convex sets.

  (Note also that for non-convex sets, not all terms in the Euler approximation are accurate.)
- Instead we use a result of Piterbarg (1996) to approximate the p-values $p(z, A)$.
- Simulations over a wide variety of S_0s and covariance structures show that coverage of U rapidly converges to the target level.

The $P_{(k)}$ Tests

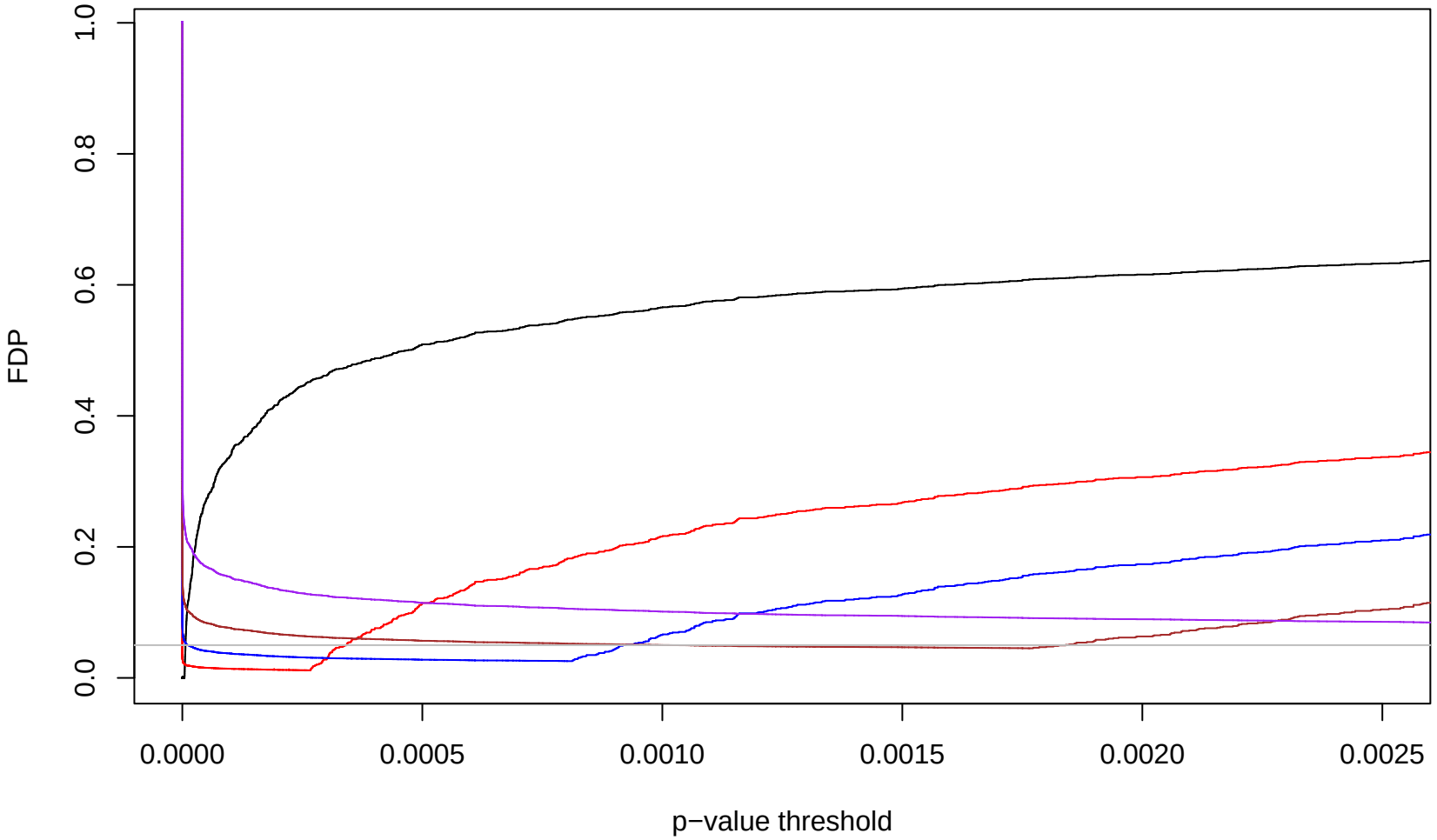
- In contrast, using the kth order statistic as a one-sided test statistic meets both desiderata.
  - For small k, these are sensitive to departures that have a large impact on FDP. Good "power."
  - Computing the confidence envelopes is linear in m.

- We call these the $P_{(k)}$ tests.

  They form a sub-family of weighted, one-sided KS tests.

Results: $P_{(k)}$ 90% Confidence Envelopes

For $k = 1, 10, 25, 50, 100$, with 0.05 FDP level marked.



Power and Optimality

The $P_{(1)}$ test corresponds to using the maximum test statistic on each subset.

Heuristic suggests this is sub-optimal: Andy-Warhol-ize.

Consider simple mixture distribution for the p-values:

$$G = (1 - a)U + aF,$$

where F is a Uniform$(0, 1/\beta)$ distribution.

Then we can construct the optimal threshold T_* (and corresponding rejection set R_*).

For any fixed k, the $P_{(k)}$ threshold satisfies

$$T_k = o_P(1)$$
$$\frac{T_*}{T_k} \xrightarrow{P} \infty.$$

Combining $P_{(k)}$ tests

- Fixed k.

  Can be effective if based on information about the alternatives, but can yield poor power.

- Estimate optimal k

  Often performs well, but two concerns: (i) if $\widehat{k} > k_{\text{opt}}$, rejection set can be empty; (ii) dependence between $\widehat{k}$ and $\overline{\text{FDP}}$ complicates analysis.

- Combine $P_{(k)}$ tests

  Let $Q_m \subset \{1, \ldots, m\}$ with cardinality q_m. Define $\overline{\text{FDP}} = \min_{k \in Q_m} \overline{\text{FDP}}_k$, where $\overline{\text{FDP}}_k$ is a $P_{(k)}$ envelope with level α/q_m.

  Generally performs well and appears to be robust.

Dependence

Extending the inversion method to handle dependence is straightforward.

Still assume each P_j is marginally Uniform$(0, 1)$ under null, but allow arbitrary joint distribution.

One formula changes: replace beta quantiles in uniformity tests with a simpler threshold.

$$J_k = \min\{j : \ P_{(j)} \geq \frac{k\alpha}{m - j}\}.$$

Simulation Results

Excerpt under simple mixture model with proportion a alternatives with Normal$(\theta, 1)$ distribution. Here $m = 10,000$ tests, $\gamma = 0.2$, $\alpha = 0.05$.

a	θ	FDP Combined	Power Combined	FDP $P_{(1)}$	Power $P_{(1)}$	FDP $P_{(10)}$	Power $P_{(10)}$
0.01	5	0.102	0.980	0.000	0.889	0.118	0.980
0.05	5	0.179	0.994	0.004	0.917	0.172	0.994
0.10	5	0.178	0.998	0.001	0.905	0.162	0.997
0.01	4	0.080	0.741	0.022	0.407	0.109	0.759
0.05	4	0.125	0.950	0.000	0.424	0.045	0.887
0.10	4	0.164	0.974	0.002	0.436	0.044	0.915
0.01	3	0.000	0.265	0.000	0.098	0.000	0.000
0.05	3	0.127	0.623	0.000	0.106	0.005	0.463
0.10	3	0.137	0.790	0.000	0.087	0.018	0.472
0.01	2	0.000	0.000	0.000	0.010	0.000	0.000

Augmentation

van der Laan, Dudoit and Pollard (2004) introduce an alternative method of exceeedance control, called augmentation

Suppose that R_0 is a rejection region that controls familywise error at level α. If $R_0 = \emptyset$ take $R = \emptyset$. Otherwise, let A be a set with $A \cap R = \emptyset$ and set $R = R_0 \cup A$. Then,

$$\mathbb{P}\Big\{\mathsf{FDP}(R) > \gamma\Big\} \leq \alpha \qquad \text{where} \qquad \gamma = \frac{\#(A)}{\#(A) + \#(R_0)}.$$

The same logic extends to k-familywise error and also gives $1 - \alpha$ confidence envelopes.

For instance, if R_0 is defined by a threshold, then

$$\overline{\mathsf{FDP}}(C) = \begin{cases} \dfrac{\#(C - R_0)}{\#(C)} & \text{if } C \neq \emptyset, \\ 0 & \text{otherwise.} \end{cases}$$

Augmentation and Inversion

Augmentation and Inversion lead to the same rejection sets.

That is, for any R_{aug}, we can find an inversion procedure with $R_{\text{aug}} = R_{\text{inv}}$.

Conversely under suitable conditions on the tests, for any R_{inv}, we can find an augmentation procedure with $R_{\text{inv}} = R_{\text{aug}}$.

When $\mathcal{U}$ is not closed under unions, inversion produces rejection sets that are not augmentations of a familywise test.