New Approaches to False Discovery Control

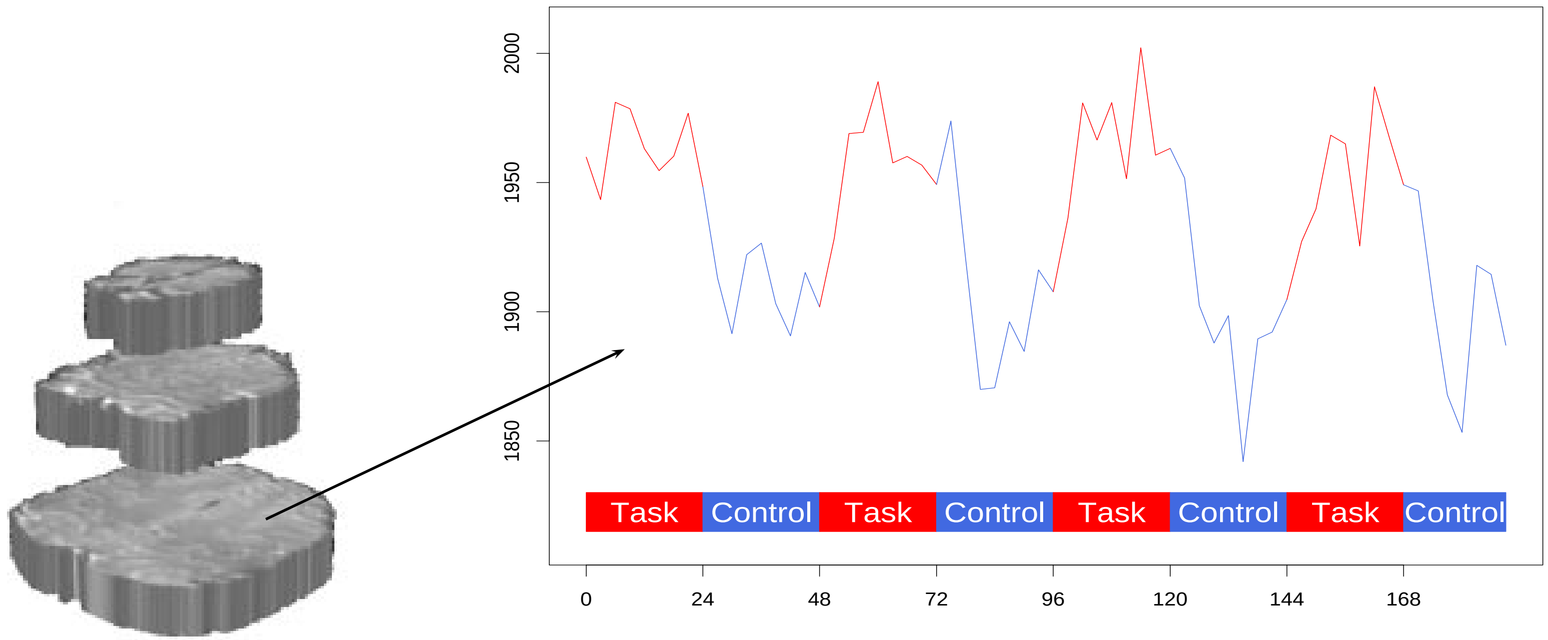
Christopher R. Genovese
Department of Statistics
Carnegie Mellon University
http://www.stat.cmu.edu/~genovese/

Larry Wasserman
Department of Statistics
Carnegie Mellon University

This work partially supported by NSF Grants SES 9866147 and ITR 0104016.

Motivating Example #1: fMRI

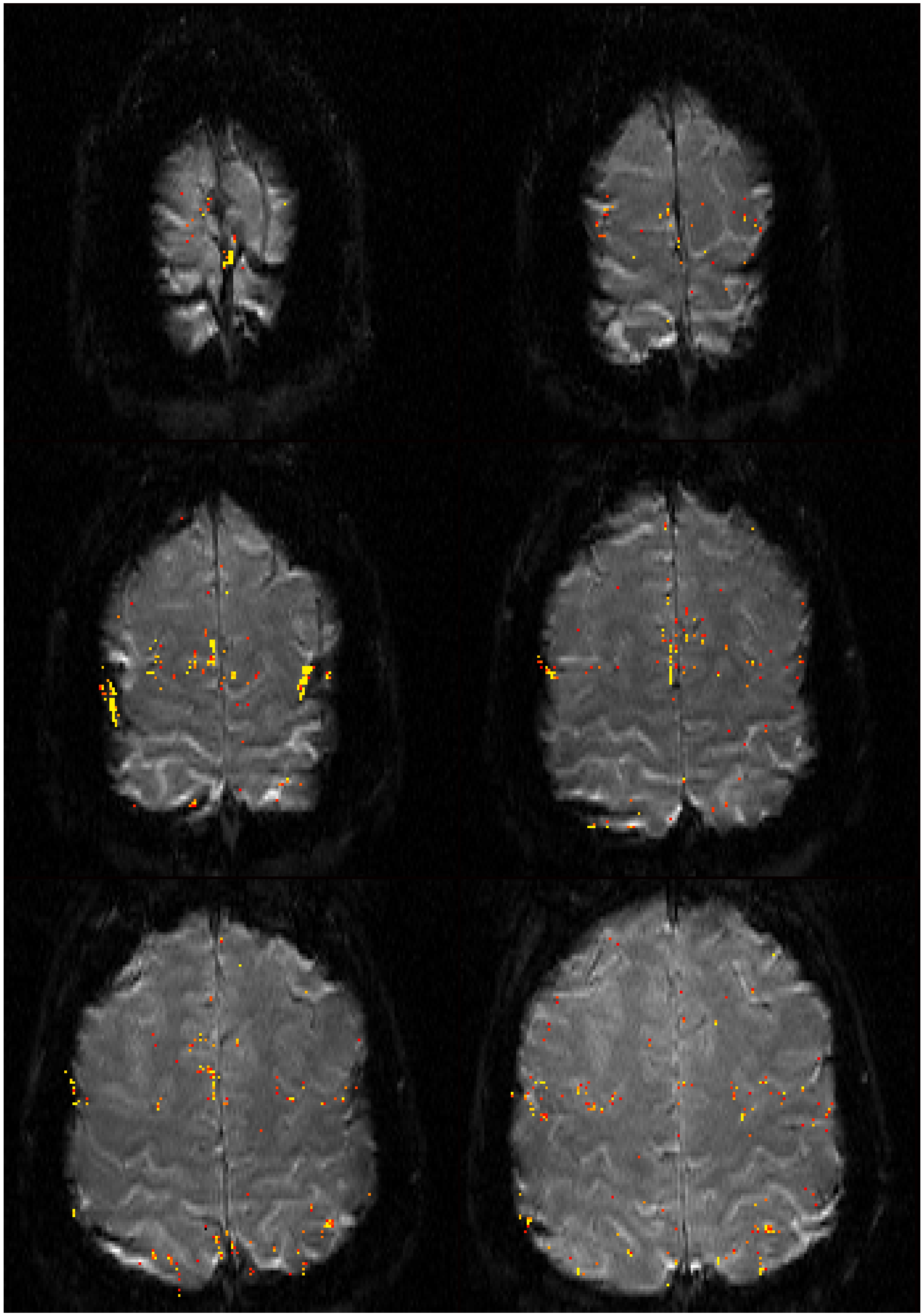
- fMRI Data: Time series of 3-d images acquired while subject performs specified tasks.



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

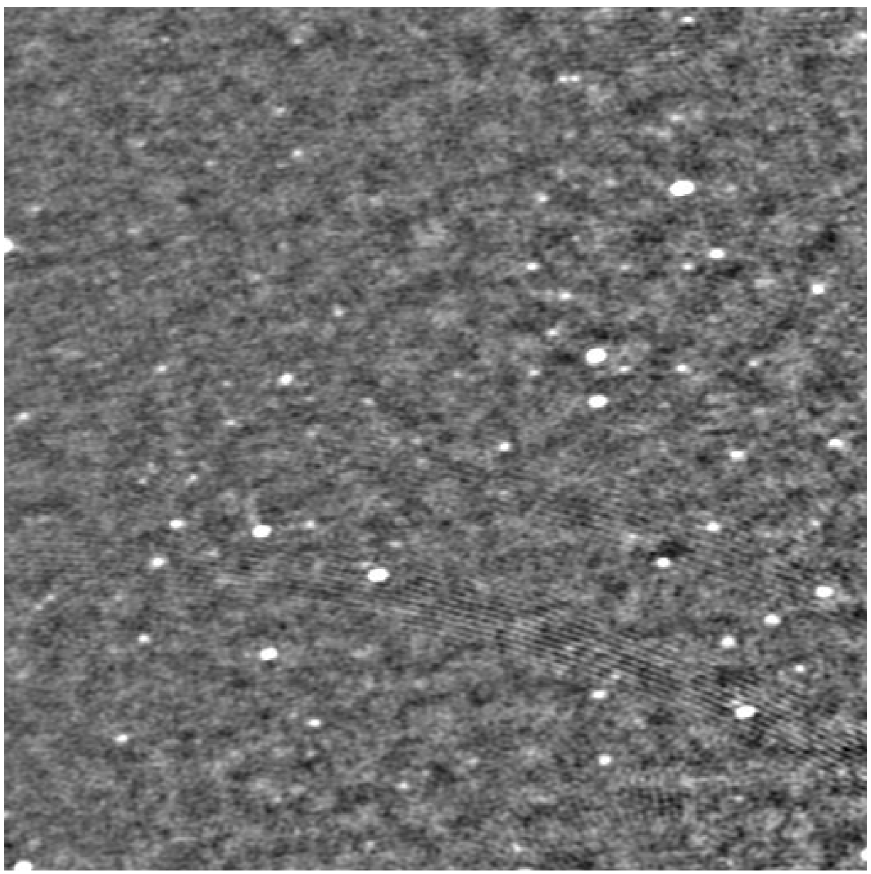
fMRI (cont'd)

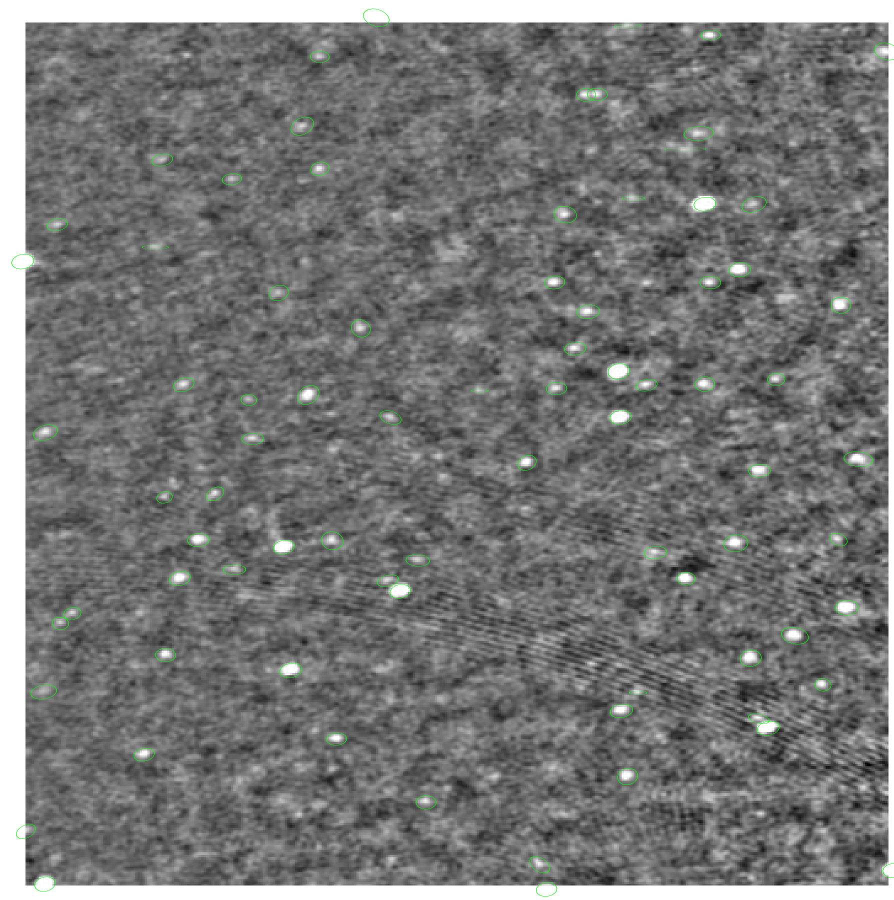
Perform hypothesis tests at many thousands of volume elements to identify loci of activation.



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 (Subject)

Gene	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 (Subject)

Gene	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.

Road Map

1. The Multiple Testing Problem
   – The Basic Problem
   – Error Criteria
2. Controlling FDR
   – Benjamini-Hochberg and Beyond
   – Outstanding Issues
   – Data Example
3. Confidence Envelopes and Thresholds
   – Exact Confidence Envelopes for the False Discovery Proportion
   – Choice of Tests
4. False Discovery Control for Random Fields
   – Confidence Supersets and Thresholds
   – Controlling the Proportion of False Clusters

Road Map

1. The Multiple Testing Problem
   - The Basic Problem
   - Error Criteria

2. Controlling FDR
   - Benjamini-Hochberg and Beyond
   - Outstanding Issues
   - Data Example

3. Confidence Envelopes and Thresholds
   - Exact Confidence Envelopes for the False Discovery Proportion
   - Choice of Tests

4. False Discovery Control for Random Fields
   - Confidence Supersets and Thresholds
   - Controlling the Proportion of False Clusters

The Multiple Testing Problem

- Perform m simultaneous hypothesis tests with a common procedure.

- For any given threshold, 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 threshold that balances the competing demands of sensitivity and specificity.

How to Choose a Threshold?

- Control Per-Comparison Type I Error
  - a.k.a. "uncorrected testing," many type I errors
  - Gives $\mathrm{P}_0\{FD_i > 0\} \le \alpha$ marginally for all $1 \le i \le m$

- Strong Control of Familywise Type I Error
  - e.g.: Bonferroni: use per-comparison significance level α/m
  - Guarantees $\mathrm{P}_0\{FD > 0\} \le \alpha$

- False Discovery Control
  - e.g.: Benjamini & Hochberg (BH, 1995, 2000): False Discovery Rate (FDR)
  - Guarantees $\mathrm{FDR} \equiv \mathrm{E}\left(\frac{FD}{D}\right) \le \alpha$

Road Map

1. The Multiple Testing Problem
   - The Basic Problem
   - Error Criteria
2. Controlling FDR
   - Benjamini-Hochberg and Beyond
   - Outstanding Issues
   - Data Example
3. Confidence Envelopes and Thresholds
   - Exact Confidence Envelopes for the False Discovery Proportion
   - Choice of Tests
4. False Discovery Control for Random Fields
   - Confidence Supersets and Thresholds
   - Controlling the Proportion of False Clusters

The Benjamini-Hochberg Procedure

- Given m p-values ordered $0 \equiv P_{(0)} < P_{(1)} < \cdots < P_{(m)}$, BH rejects any null hypothesis with $P_j \leq T_{\mathrm{BH}}$, where

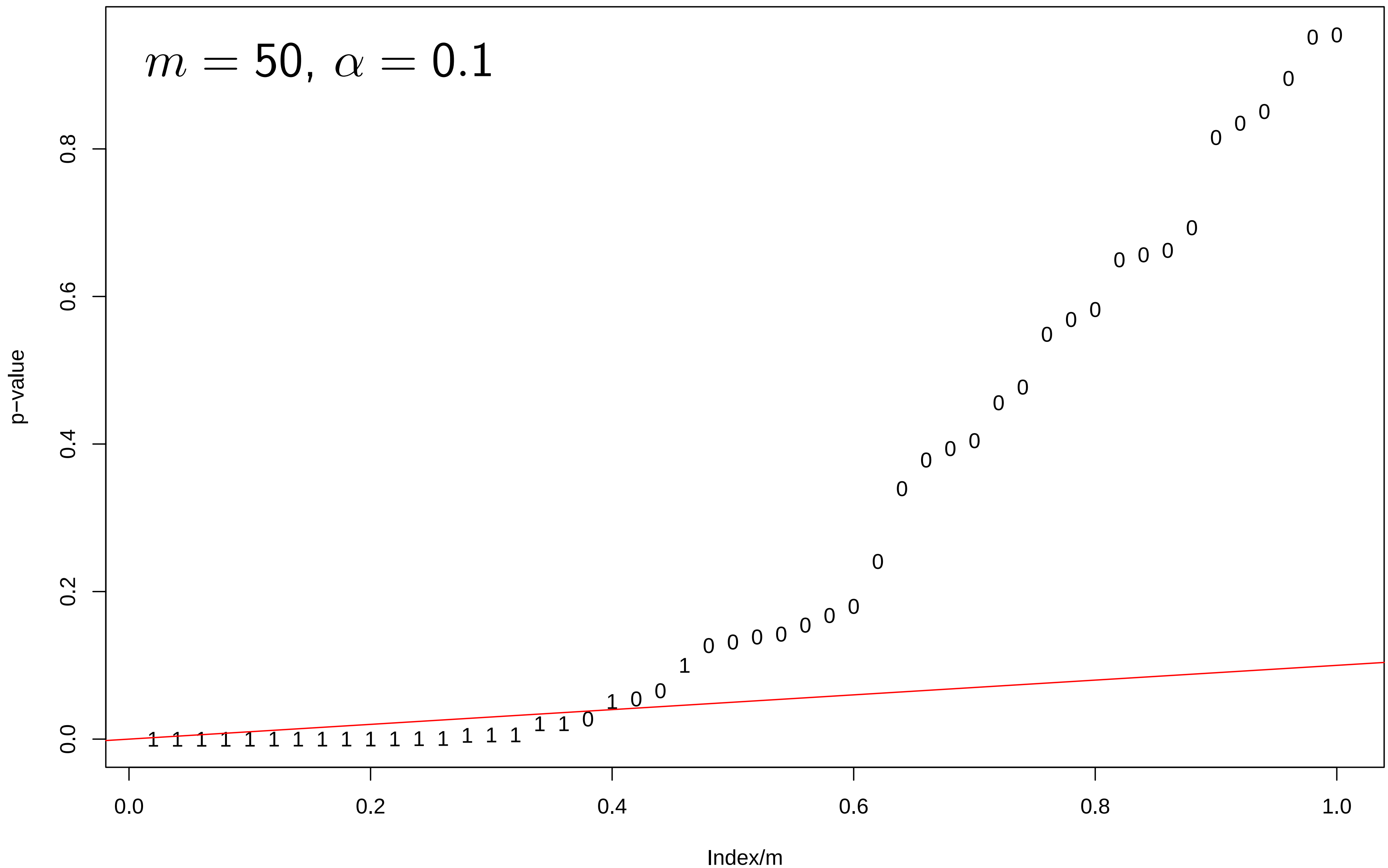
$$T_{\mathrm{BH}} = \max\left\{P_{(i)} \colon \; P_{(i)} \leq \alpha \frac{i}{m}\right\}.$$

- BH procedure guarantees that

$$\mathrm{FDR} \equiv \mathrm{E}\left(\frac{FD}{D}\right) \leq \frac{T_0}{m} \alpha.$$

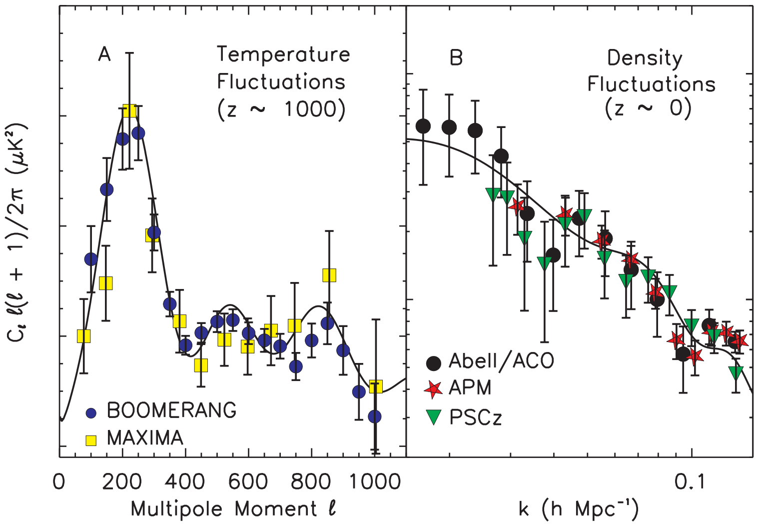
- This bound holds at least under "positive dependence".
- Gives more power than Bonferroni, fewer Type I errors than uncorrected testing.
- Replacing α by $\alpha / \sum_{i=1}^{m} 1/i$ extends FDR bound to any distribution, but this is typically *very* conservative.

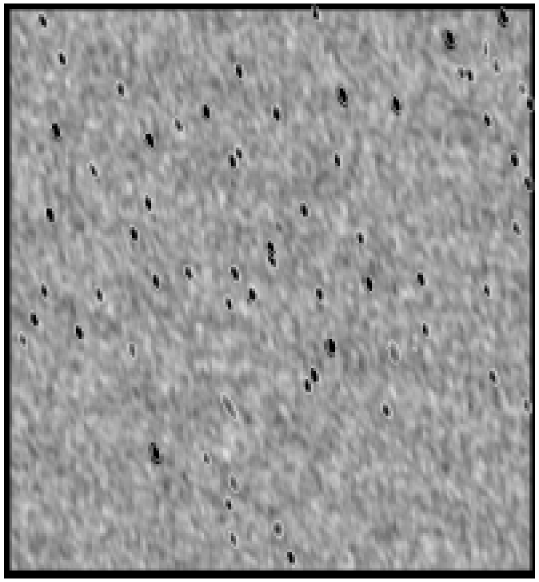
The Benjamini-Hochberg Procedure (cont'd)

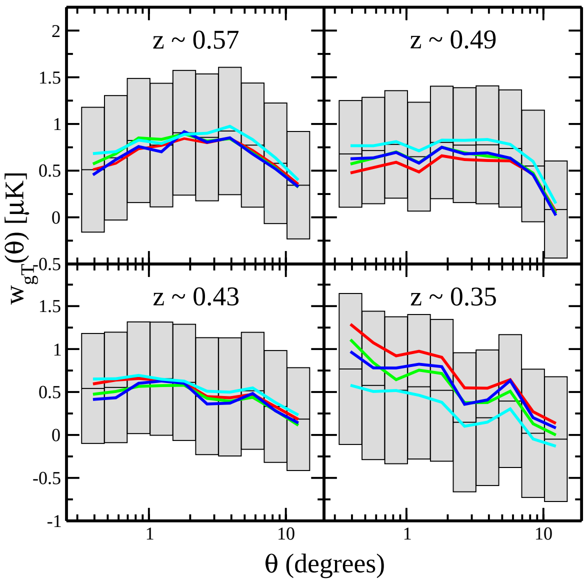


Astronomical Examples (PiCA Group)

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







Mixture Model for Multiple Testing

- Let $P^m = (P_1, \ldots, P_m)$ be the p-values for the m tests.
- Let $H^m = (H_1, \ldots, H_m)$ where $H_i = 0$ (or 1) if the i^{th} null hypothesis is true (or false).
- We assume the following model:

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

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

Mixture Model for Multiple Testing (cont'd)

- Marginally, $P_1, \ldots, P_m$ are drawn iid from

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

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

$$F = \int \xi \, d\mathcal{L}_{\mathcal{F}}(\xi).$$

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

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

- Can also work under what we call the *conditional model* where $H_1, \ldots, H_m$ are fixed, unknown.

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 $P_i \le T(P^m)$. We 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	$T_{\mathrm{BH}}(P^m) = \sup\{t\colon \widehat{G}(t) = t/\alpha\}$
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 Discovery Process

- Define two stochastic processes as a function of threshold t: the False Discovery Proportion FDP(t) and False Nondiscovery Proportion FNP(t).

$$\text{FDP}(t; P^m, H^m) = \frac{\sum_i 1\left\{P_i \le t\right\}(1 - H_i)}{\sum_i 1\left\{P_i \le t\right\} + 1\left\{\text{all } P_i > t\right\}} = \frac{\#\text{False Discoveries}}{\#\text{Discoveries}}$$

$$\text{FNP}(t; P^m, H^m) = \frac{\sum_i 1\left\{P_i > t\right\} H_i}{\sum_i 1\left\{P_i > t\right\} + 1\left\{\text{all } P_i \le t\right\}} = \frac{\#\text{False Nondiscoveries}}{\#\text{Nondiscoveries}}$$

- These converge to Gaussian processes away from $t = 0$.

The False Discovery Rate

- For a given procedure T, let FDP and FNP denote the value of these processes at $T(P^m)$.

- Then, the False Discovery Rate (FDR) and the False Nondiscovery Rate (FNR) are given by

$$\mathsf{FDR} = \mathsf{E}(\mathsf{FDP}) \qquad \mathsf{FNR} = \mathsf{E}(\mathsf{FNP}).$$

- The BH guarantee becomes

$$\mathsf{FDR} \leq (1 - a)\alpha \leq \alpha,$$

where the first inequality is an equality in the continuous case.

The BH Procedure Revisited

- If $\widehat{G}$ is the empirical cdf of the m p-values, $\widehat{G}(P_{(i)}) = i/m$, so

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

  Note that $\mathrm{FDR}(t) \approx \frac{(1-a)t}{G(t)}$, so BH bounds $\widehat{\mathrm{FDR}}$ taking $a = 0$.

- BH performs best in very sparse cases ($T_0 \approx m$); power can be improved in non-sparse cases by more complicated procedures.

- One can think of BH as a plug-in procedure for estimating

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

- Genovese and Wasserman (2002) showed that T_{BH} converges to a fixed-threshold at u^*.

Optimal Thresholds

- In the continuous case, Benjamini and Hochberg's argument shows that

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

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

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

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

- Note that $t^* \geq u^*$. If we knew a, this would correspond to using the BH procedure with $\alpha/(1-a)$ in place of α.

Optimal Thresholds (cont'd)

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

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

- Ignoring $O()$ terms and choosing t to minimize E(FNP(t)) subject to E(FDP(t)) $\leq \alpha$, yields $t^*(a, G)$ as the optimal threshold.

- T_{PI} considered in some form by Benjamini & Hochberg (2000), Storey (2003), and Genovese and Wasserman (2003).

Selected Recent Work on FDR

Abromovich, Benjamini, Donoho, & Johnstone (2000)

Benjamini & Hochberg (1995, 2000)

Benjamini & Yekutieli (2001)

Efron, Tibshirani, & Storey, J. (2001)

Finner & Roters (2001, 2002)

Hochberg & Benjamini (1999)

Genovese & Wasserman (2001,2002,2003)

Pacifico, Genovese, Verdinelli, & Wasserman (2003)

Sarkar (2002)

Seigmund, Taylor, & Storey (2003)

Storey (2001,2002)

Storey & Tibshirani (2001)

Tusher, Tibshirani, Chu (2001)

Yekutieli & Benjamini (2001)

Outstanding Issues

- Interpretation
  - How to choose α?
  - How to interpret the FDR bound?
- Dependence
  - Is positive regression dependence enough? How do we test for it?
  - BH method appears to be very hard to "break;" plug-in more sensitive to dependence.
  - Extensions of new methods to handle dependence structure.
- Spatial Structure
  - Standard multiple-testing methods ignore location information.
  - Focal regions are easier to identify than arbitrarily placed voxels.
  - Regions rather than voxels are the units of interest.
  - This is the key to much improved inference in applications like fMRI.

Data Example

- Monkeys exhibit *visual remapping* in parietal cortex

  When the eyes move so that the receptive field of a neuron lands on a previously stimulated location, the neuron fires even though no stimulus is present.

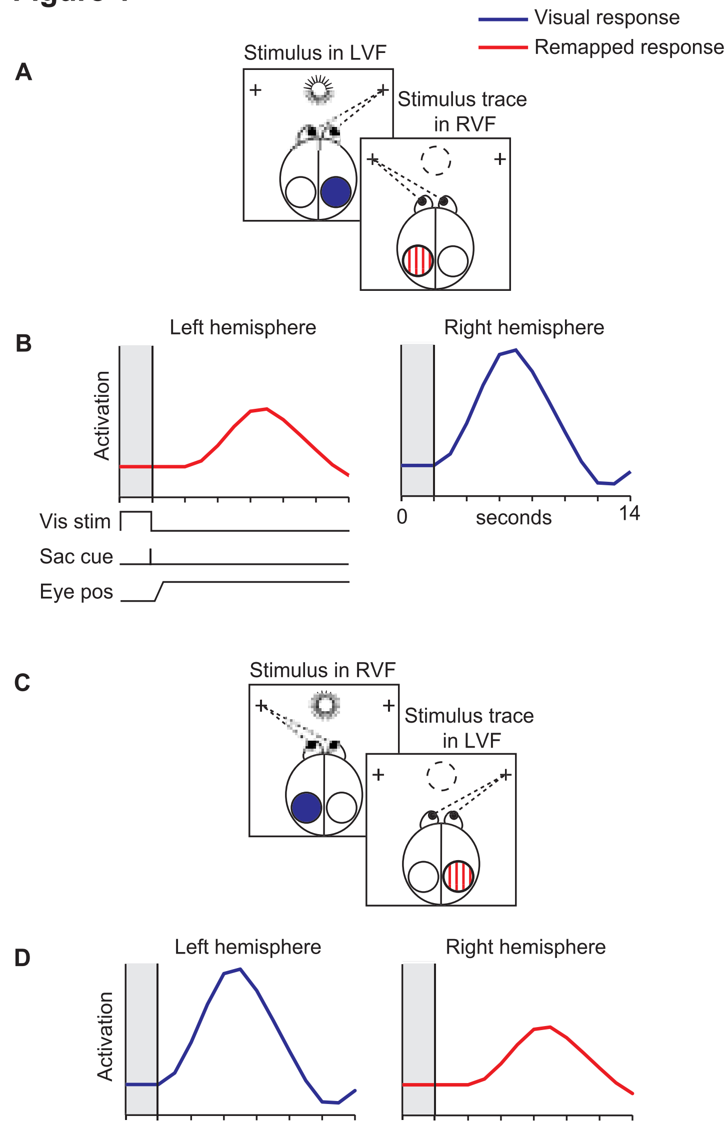
  Implies transformation in neural representation with eye movements. (Duhamel et al. 1992)

- Seek evidence for remapping in human cortex.

- See Merriam, Genovese, and Colby (2003). *Neuron*, 39, 361–373 for more details.

- EPI-RT acquisition, TR 2s, TE 30ms, 20 oblique slices, 3.125mm $\times$ 3.125mm $\times$ 3mm voxels.

Figure 1



Road Map

1. The Multiple Testing Problem
   - The Basic Problem
   - Error Criteria
2. Controlling FDR
   - Benjamini-Hochberg and Beyond
   - Outstanding Issues
   - Data Example
3. Confidence Envelopes and Thresholds
   - Exact Confidence Envelopes for the False Discovery Proportion
   - Choice of Tests
4. False Discovery Control for Random Fields
   - Confidence Supersets and Thresholds
   - Controlling the Proportion of False Clusters

Confidence Envelopes and Thresholds

- In practice, it would be useful to be able to control quantiles of the FDP process.

- We want a procedure T that for specified A and γ guarantees

$$\mathsf{P}_G\Big\{ \mathsf{FDP}(T) > A \Big\} \leq \gamma$$

  We call this an $(A, 1-\gamma)$ *confidence-threshold procedure*.

- Three methods: (i) asymptotic closed-form threshold, (ii) asymptotic confidence envelope, and (iii) exact small-sample confidence envelope. (See Genovese & Wasserman 2003, to appear *Annals of Statistics*.)

  I'll focus here on (iii).

Confidence Envelopes and Thresholds (cont'd)

- A $1 - \gamma$ confidence envelope for FDP is a random function $\overline{\text{FDP}}(t)$ on $[0, 1]$ such that

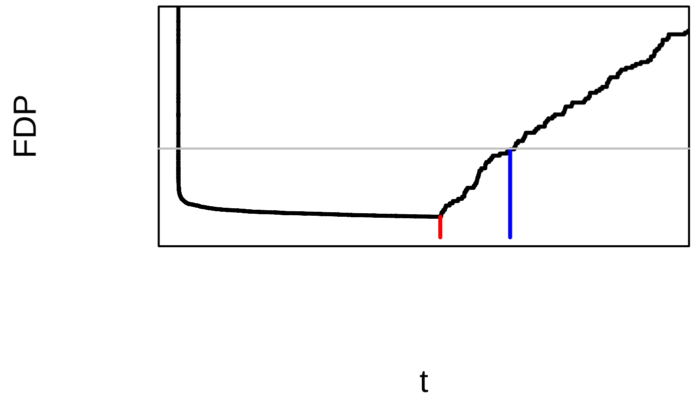
$$\text{P}\Big\{\text{FDP}(t) \le \overline{\text{FDP}}(t) \text{ for all } t\Big\} \ge 1 - \gamma.$$

- Given such an envelope, we can construct confidence thresholds.

  Two special cases have proven useful.

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

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



Exact Confidence Envelopes

- Given $V_1, \ldots, V_j$, let $\varphi_j(v_1, \ldots, v_j)$ be a level γ test of the null hypothesis that $V_1, \ldots, V_j$ are IID Uniform$(0, 1)$.

- Define
$$p_0^m(h^m) = (p_i\colon h_i = 0,\ 1 \le i \le m)$$
$$m_0(h^m) = \sum_{i=1}^{m} (1 - h_i)$$
and
$$\mathcal{U}_\gamma(p^m) = \left\{ h^m \in \{0, 1\}^m\colon \varphi_{m_0(h^m)}\left(p_0^m(h^m)\right) = 0 \right\}.$$
Note that as defined, $\mathcal{U}_\gamma$ always contains the vector $(1, 1, \ldots, 1)$.

- Let
$$\mathcal{G}_\gamma(p^m) = \left\{ \mathsf{FDP}(\cdot; h^m, p^m)\colon\ h^m \in \mathcal{U}_\gamma(p^m) \right\}$$
$$\mathcal{M}_\gamma(p^m) = \left\{ m_0(h^m)\colon\ h^m \in \mathcal{U}_\gamma(p^m) \right\}.$$

Exact Confidence Envelopes (cont'd)

- THEOREM. For all $0 < a < 1$, F, and positive integers m,

$$\mathsf{P}\Big\{H^m \in \mathcal{U}_\gamma(P^m)\Big\} \geq 1 - \gamma$$
$$\mathsf{P}\Big\{M_0 \in \mathcal{M}_\gamma(P^m)\Big\} \geq 1 - \gamma$$
$$\mathsf{P}\Big\{\mathsf{FDP}(\cdot; H^m, P^m) \in \mathcal{G}_\gamma\Big\} \geq 1 - \gamma.$$

- Define $\overline{\mathsf{FDP}}$ to be the pointwise supremum over $\mathcal{G}_\gamma$. This is a $1 - \gamma$ confidence envelope for FDP.

- Confidence thresholds follow directly. For example,

$$T_\alpha = \sup\Big\{t : \overline{\mathsf{FDP}}(t) \leq \alpha\Big\}$$

  is an $(\alpha, 1 - \gamma)$ confidence threshold.

Choice of Tests

- The confidence envelopes depend strongly on choice of tests.
- Two desiderata for selecting uniformity tests:
  - "Power", such that $\overline{\text{FDP}}$ is close to FDP, and
  - Computability, given that there are 2^m subsets to test.
- Want an automatic way to choose a good test
- Traditional uniformity tests, such as the (one-sided) Kolmogorov-Smirnov (KS) test, do not usually meet both conditions.

  For example, the KS test is sensitive to deviations from uniformity equally though all the p-values.

The $P_{(k)}$ Tests

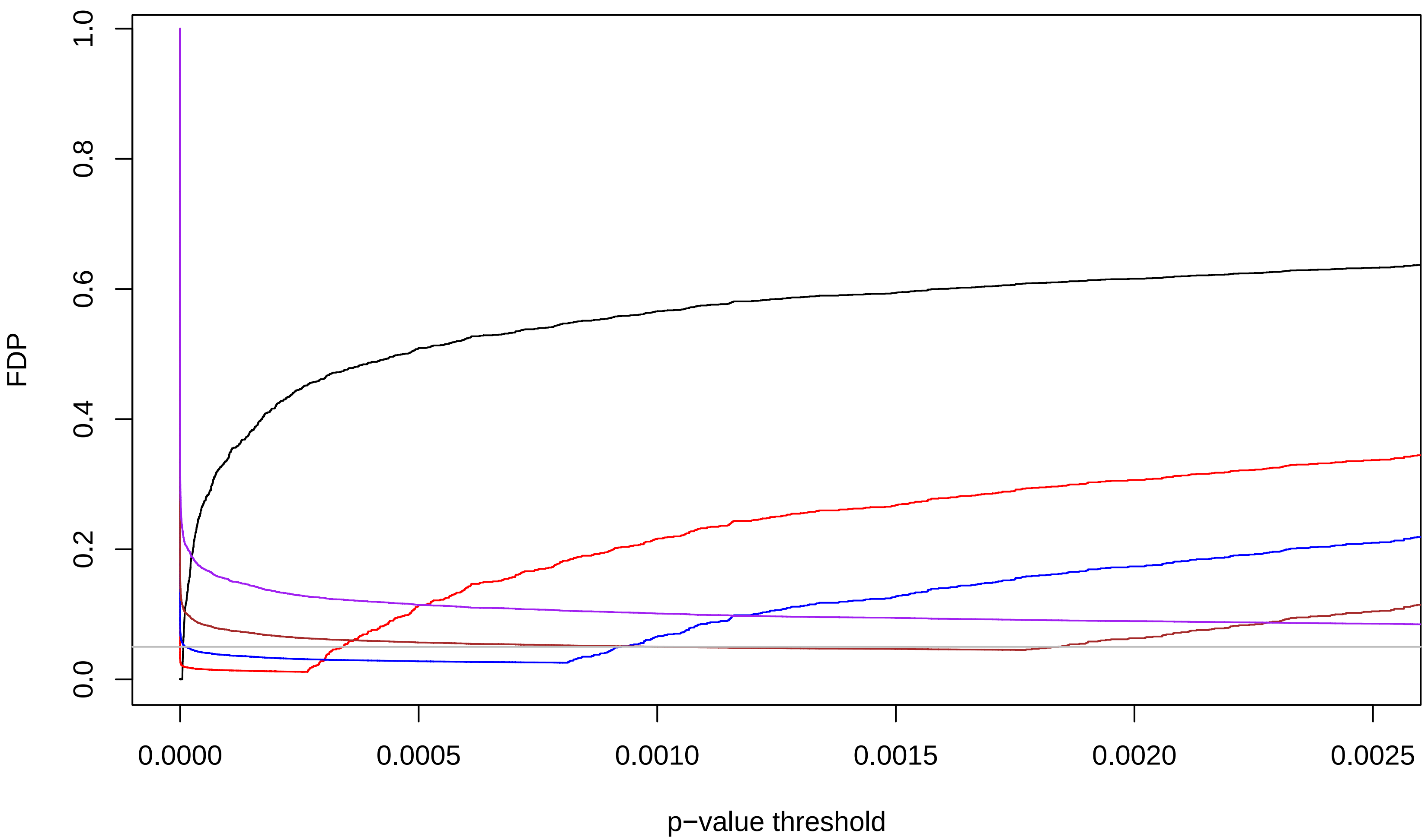
- In contrast, using the kth order statistic as a one-sided test statistic meets both desiderata.
  - For small k, these are sensitive to departures that have a large impact on FDP. Good "power."
  - Computing the confidence envelopes is linear in m.

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

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

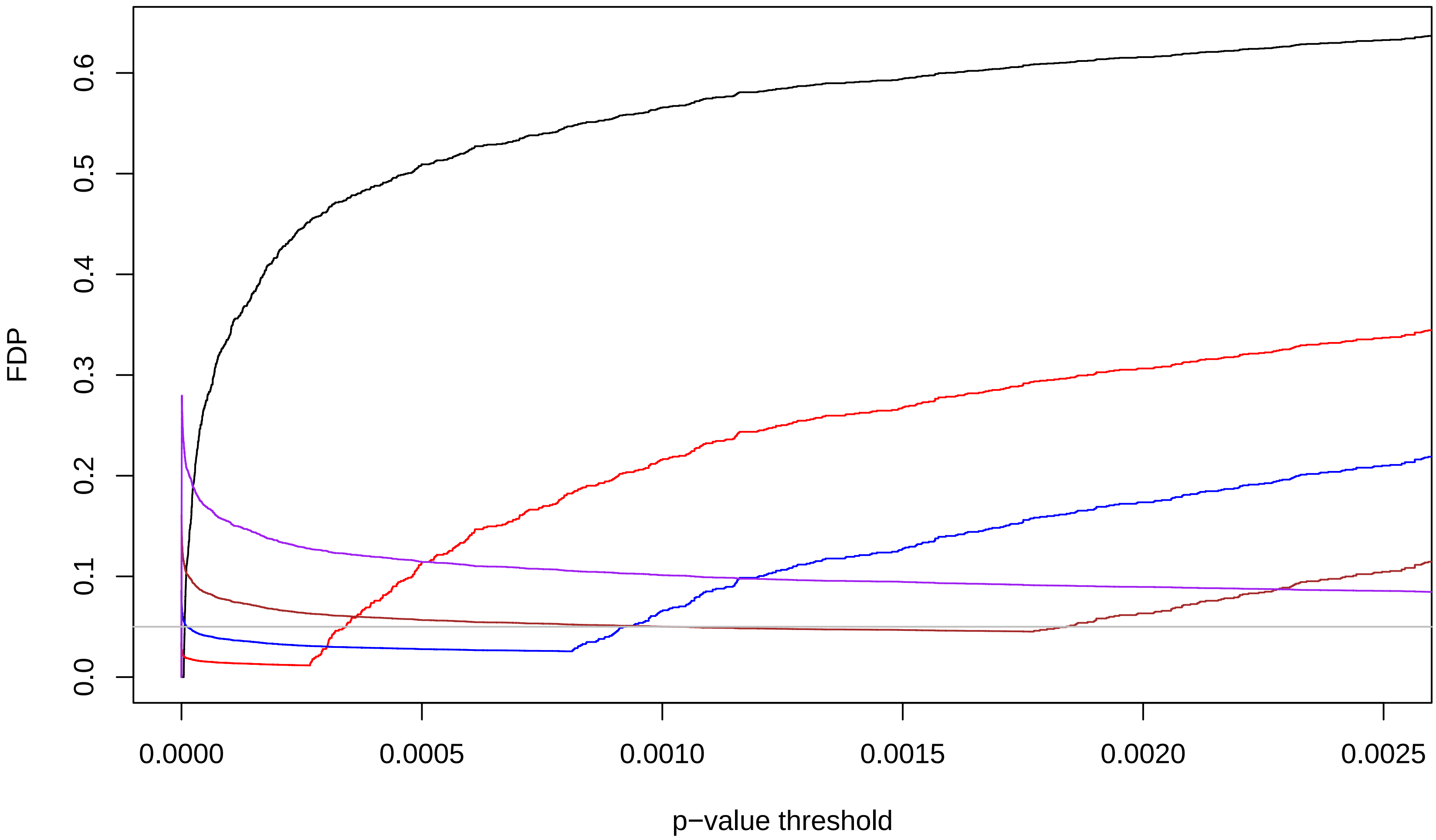
Results: $P_{(k)}$ 90% Confidence Envelopes

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

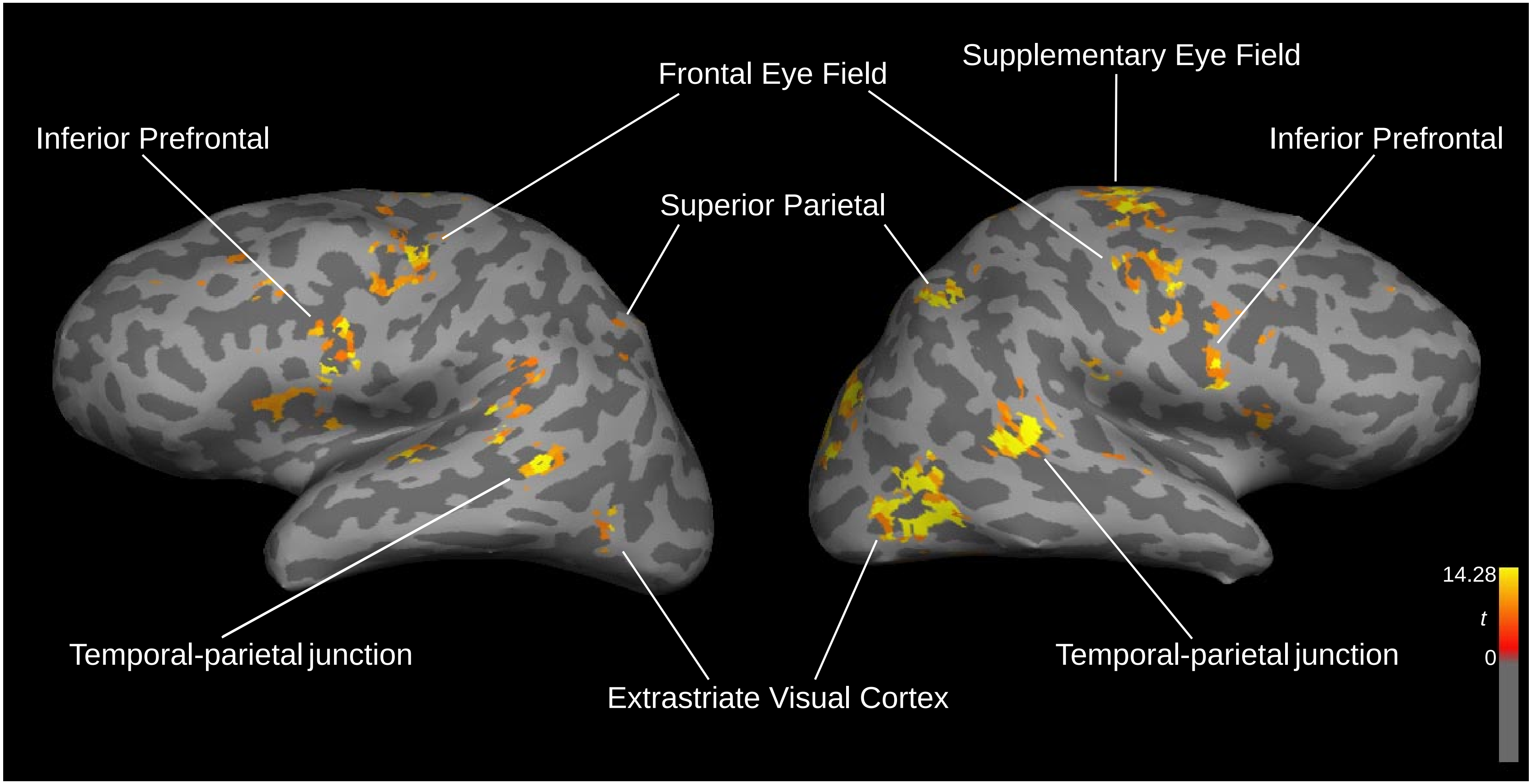


Results: $P_{(k)}$ 90% Modified Envelopes

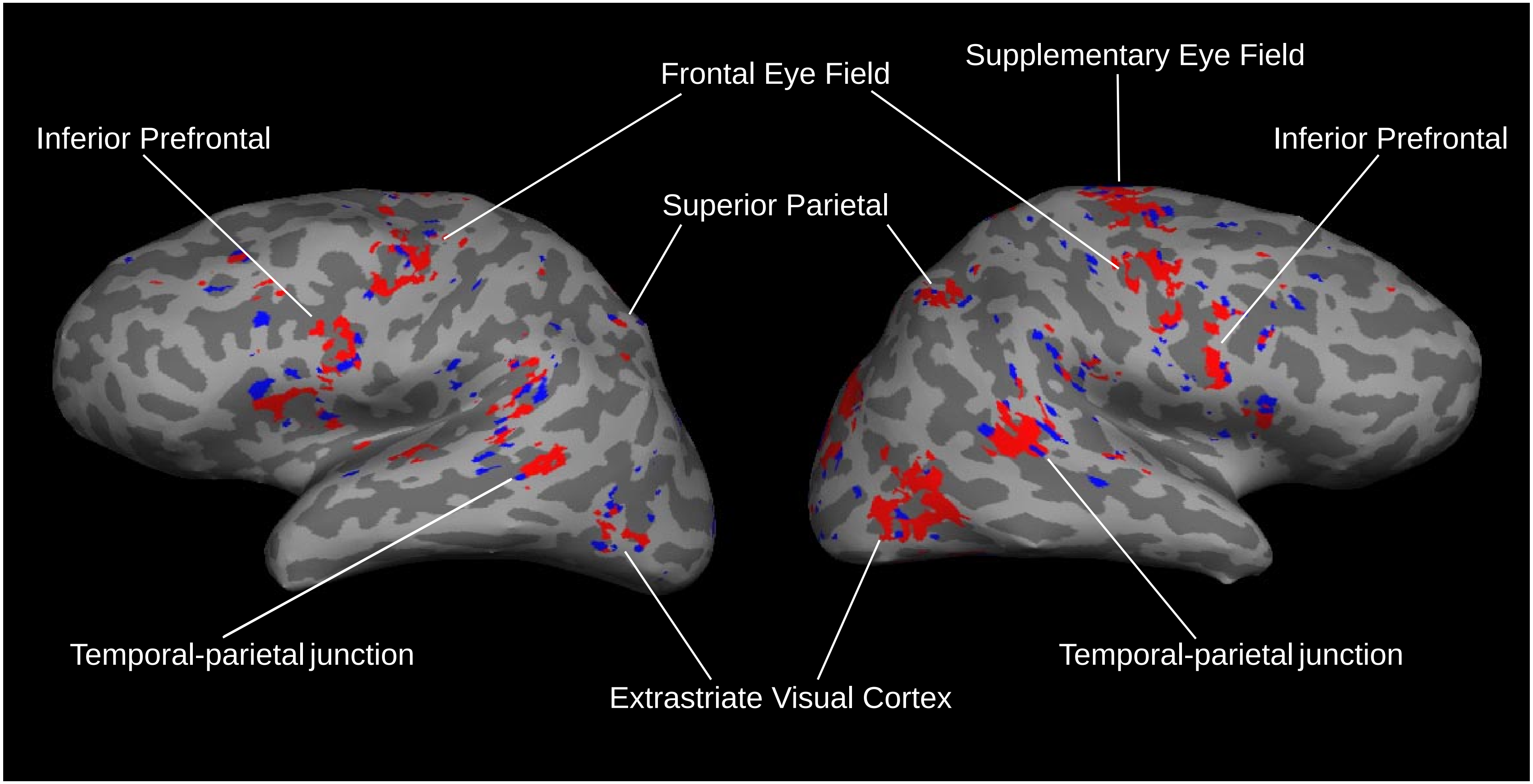
For $k = 1, 10, 25, 50, 100$, with 0.05 FDP level marked.



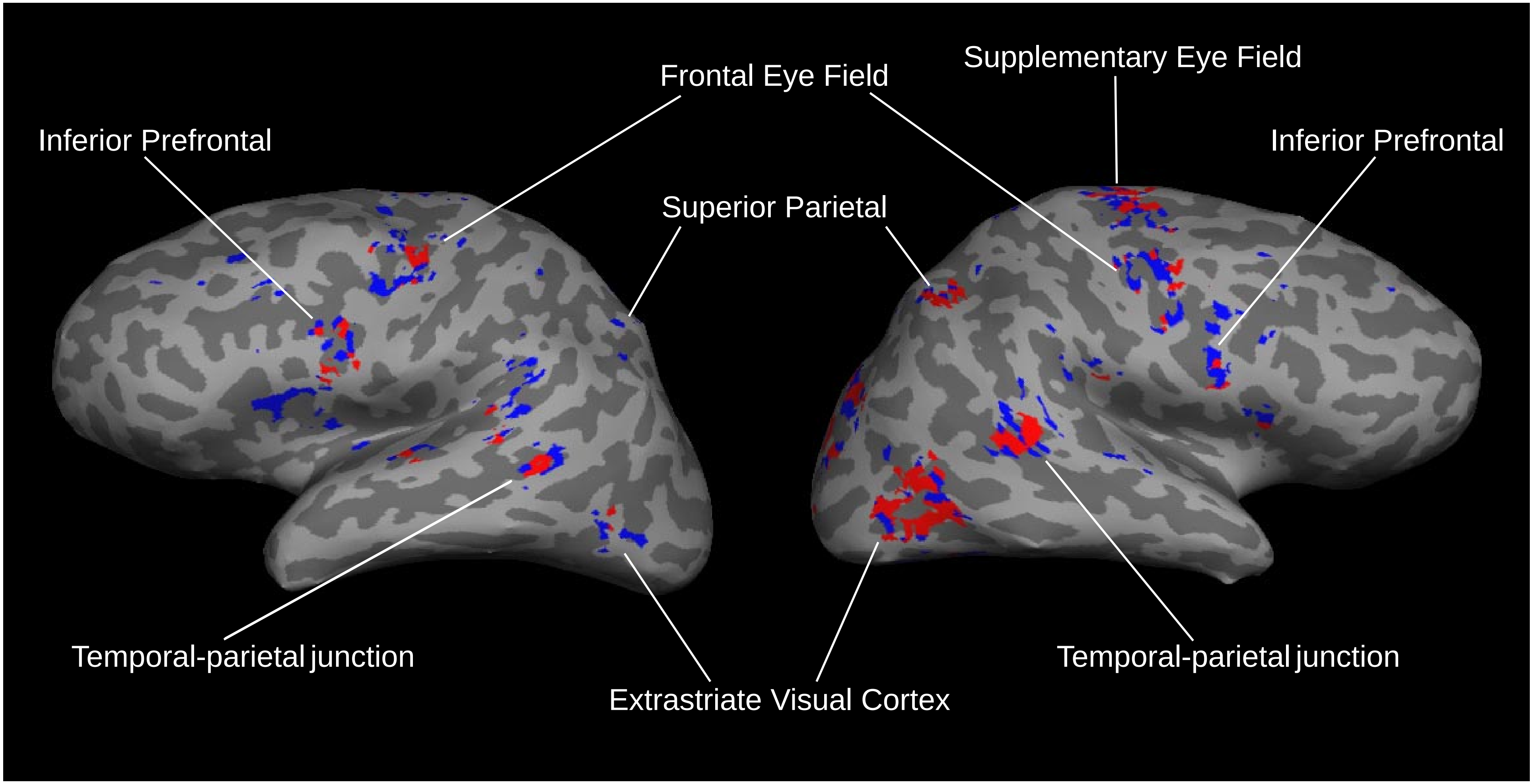
Results: (0.05,0.9) Confidence Threshold



Results: (0.05,0.9) Threshold versus BH



Results: (0.05,0.9) Threshold versus Bonferroni



Choosing k

- Direct Approach

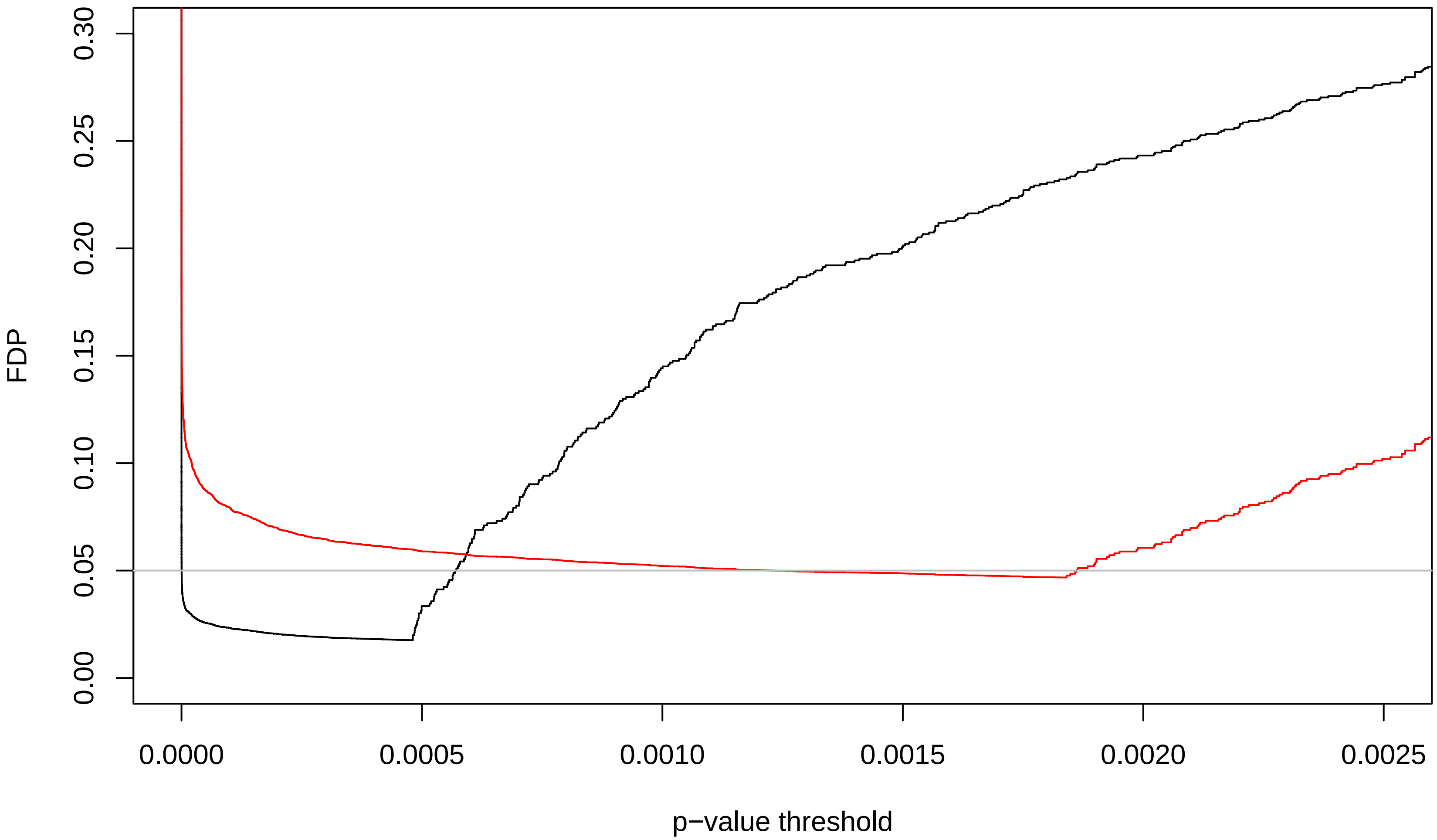
  Simulate from prior family, such as Normal($\theta, 1$), Noncentral $t(\theta)$, or mixtures of these.

  Compute the optimal k, $k^*(\theta, m)$.

- Data-dependent approaches
  - Estimate a and F, and simulate from corresponding mixture.
  - Parametric estimate $k^*(\widehat{\theta}, m)$.
  - Solve for optimal k distribution using smoothed estimate of G.

  The data-dependence only has a small effect on coverage.

Results: Direct versus Fitting Approach



Road Map

1. The Multiple Testing Problem
   - The Basic Problem
   - Error Criteria
2. Controlling FDR
   - Benjamini-Hochberg and Beyond
   - Outstanding Issues
   - Data Example
3. Confidence Envelopes and Thresholds
   - Exact Confidence Envelopes for the False Discovery Proportion
   - Choice of Tests
4. False Discovery Control for Random Fields
   - Confidence Supersets and Thresholds
   - Controlling the Proportion of False Clusters

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 by

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

  where λ is usually Lebesgue measure.

- As in the cases discussed earlier, we can control FDR or quantiles of FDP.

- Our approach is again based on constructing a confidence envelope for FDP by finding a confidence superset U of S_0.

Confidence Supersets and Envelopes

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 $p(z, A) = \mathsf{P}\Big\{X(A) \geq z\Big\}$.

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

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

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.

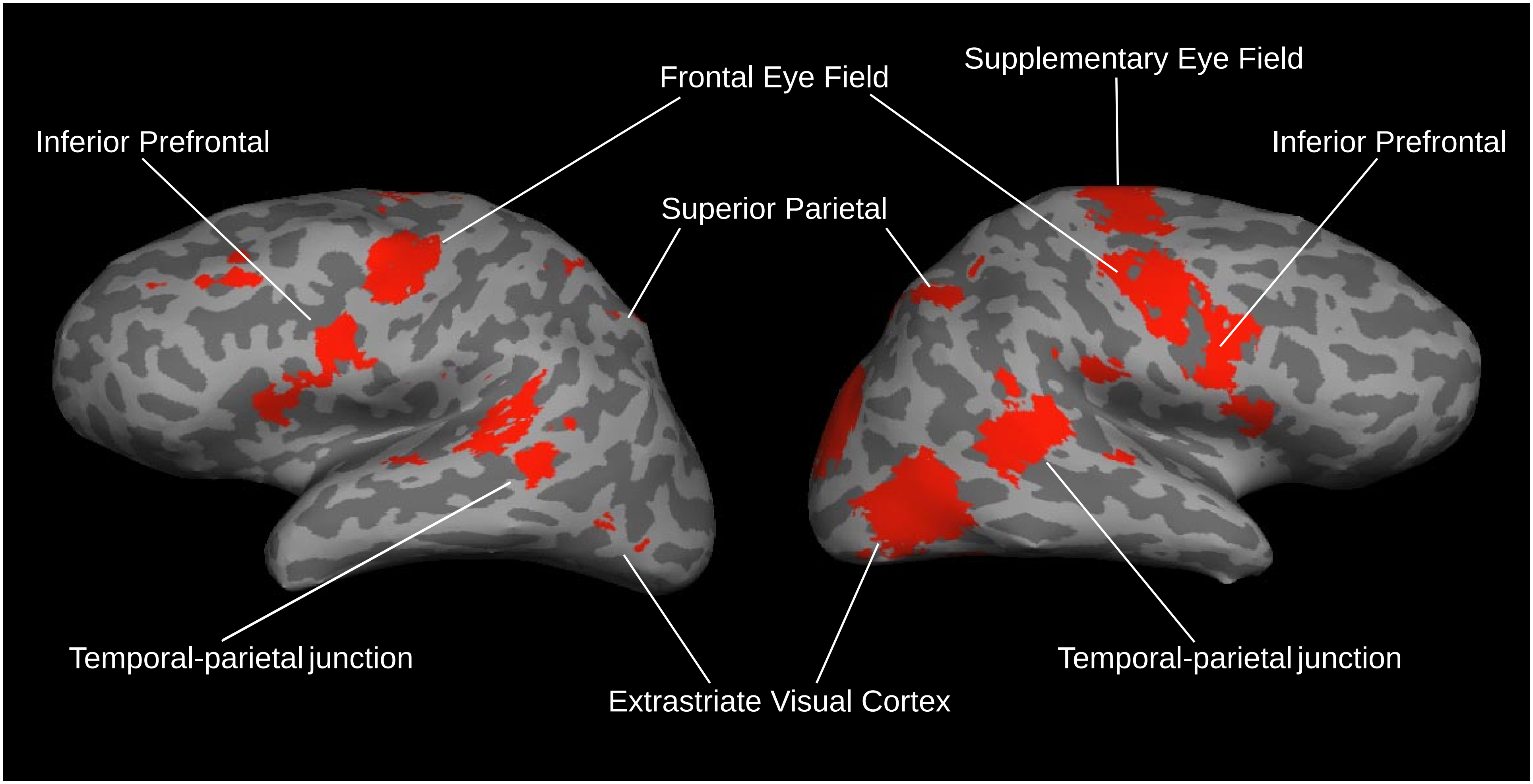
Note: We need not carry out the tests for all subsets.

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.

Results: (0.05,0.9) Confidence Threshold

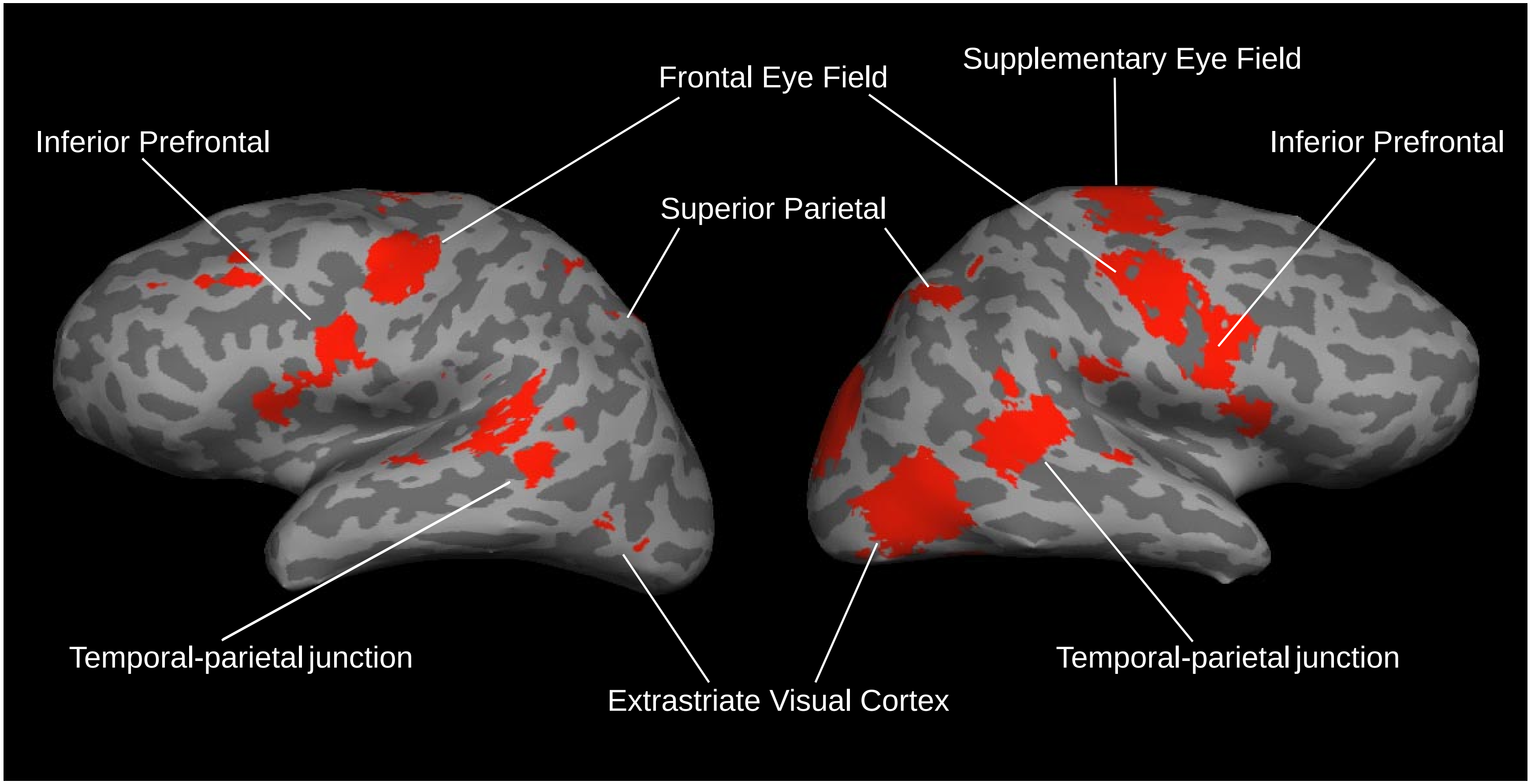


Controlling the Proportion of False Regions

- 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

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



Take-Home Points

- Confidence thresholds have practical advantages for False Discovery Control.

  In particular, we gain a stronger inferential guarantee with little effective loss of power.

- Dependence complicates the analysis greatly, but confidence envelopes appear to be valid under positive dependence.

- For spatial applications, adjacency relations can be highly informative but are typically ignored by multiple-testing methods.

  Controlling proportion of false regions is a first step.

  Region-based false discovery control (work in progress) is the next step.