Confidence Thresholds and 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 NIH Grant 1 R01 NS047493-01.

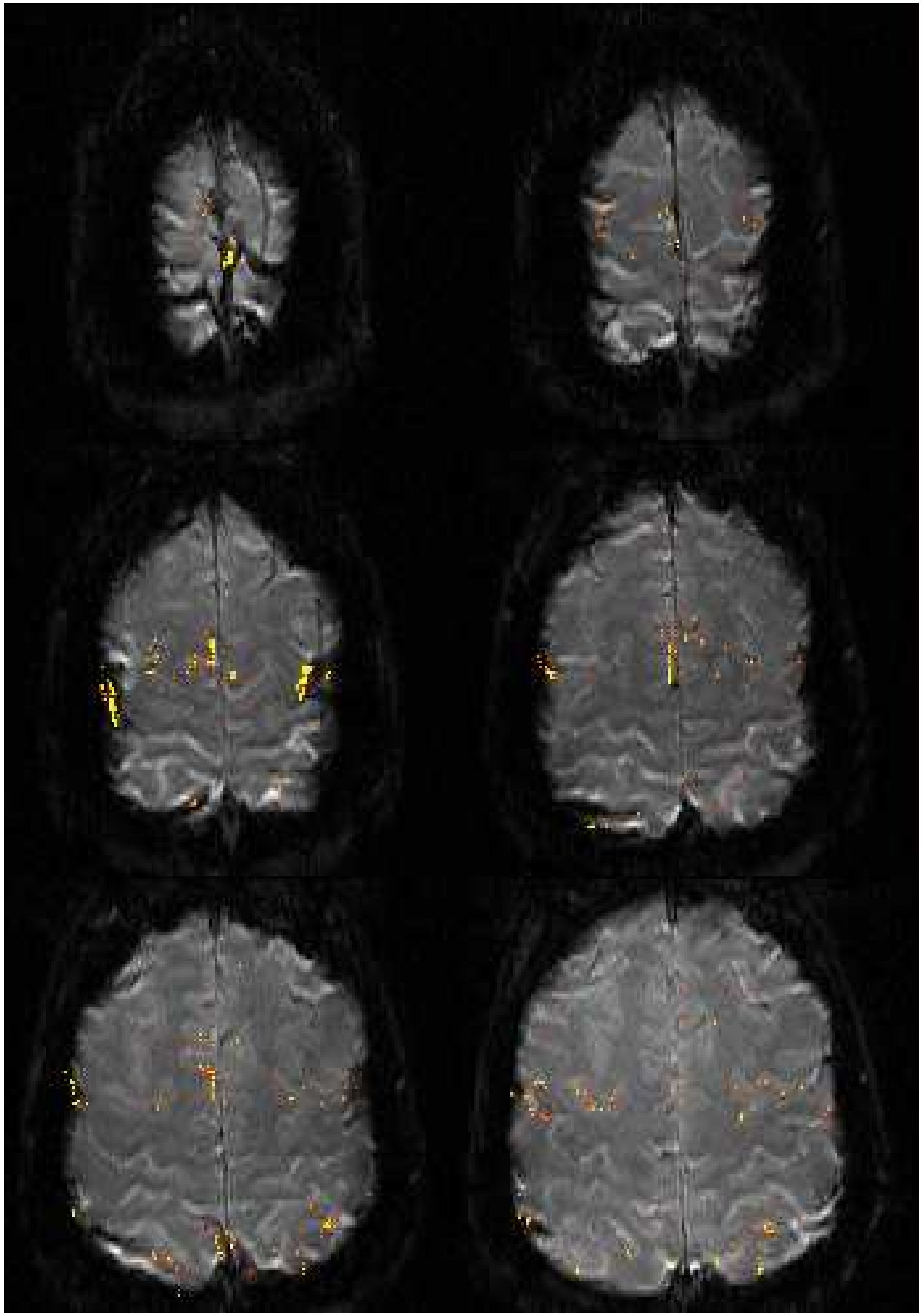
Motivating Example #1: fMRI

- 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.]

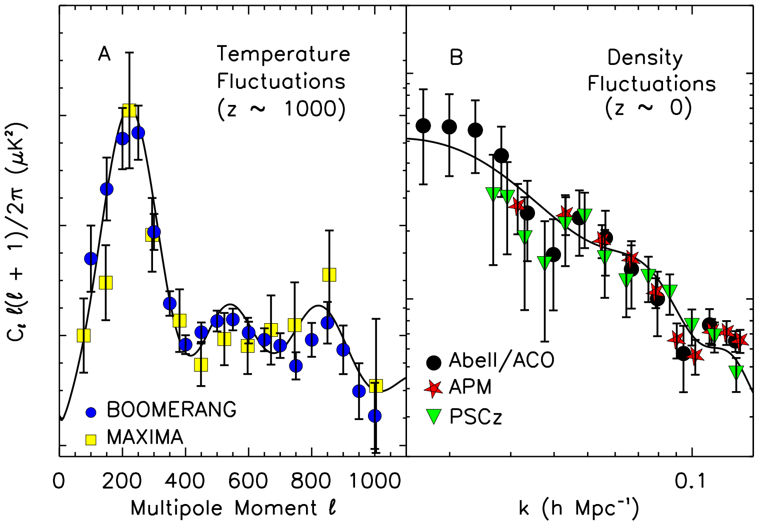
fMRI (cont'd)

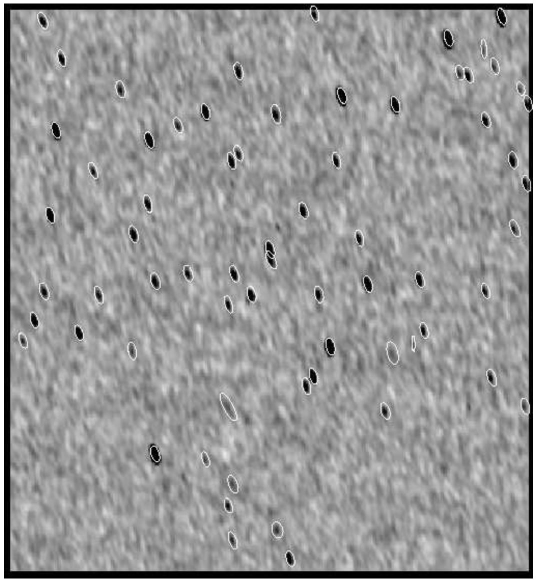
Perform hypothesis tests at many thousands of volume elements to identify loci of activation.

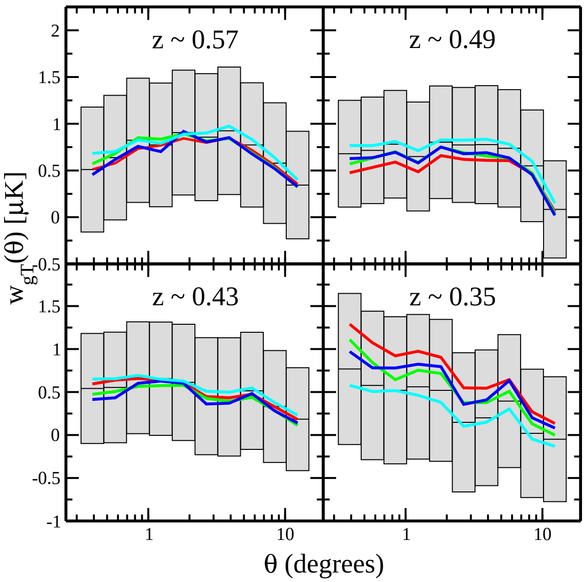


Motivating Example #2: Cosmology

- 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.

Condition 1

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

Condition 2

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

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

Objective

Develop methods for exceedance control of the False Discovery Proportion (FDP):

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

as an alternative to mean (FDR) control.

Useful in applications as the basis for a secondary inference about the pattern of false discoveries.

Also useful as an FDR diagnostic.

Plan

1. Set Up

– Testing Framework

– FDR and FDP

2. Exceedance Control for FDP

– Inversion and the $P_{(k)}$ test

– Power and Optimality

– Combining $P_{(k)}$ tests

– Augmentation

3. False Discovery Control for Random Fields

– Confidence Supersets and Thresholds

– Controlling the Proportion of False Clusters

– Scan Statistics

Plan

1. Set Up
   - Testing Framework
   - FDR and FDP
2. Exceedance Control for FDP
   - Inversion and the $P_{(k)}$ test
   - Power and Optimality
   - Combining $P_{(k)}$ tests
   - Augmentation
3. False Discovery Control for Random Fields
   - Confidence Supersets and Thresholds
   - Controlling the Proportion of False Clusters
   - Scan Statistics

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.

Testing Framework: Hypotheses

Let random vectors $X_1, \ldots, X_n$ be drawn $\textsc{iid}$ from distribution $\mathbb{P}$.

Consider m hypotheses (typically $m >> n$) of the form

$$H_{0j} : \ \mathbb{P} \in \mathcal{M}_j \quad \text{versus} \quad H_{1j} : \ \mathbb{P} \notin \mathcal{M}_j \qquad j = 1, \ldots, m,$$

for sets of probability distributions $\mathcal{M}_1, \ldots, \mathcal{M}_m$.

Common case:

$X_i = (X_{i1}, \ldots, X_{im})$ comprises m measurements on subject i.

Here, we might take $\mathcal{M}_j = \{\mathbb{P}\colon \ \mathsf{E}_{\mathbb{P}}(X_{ij}) = \mu_j\}$ for some constant μ_j.

Testing Framework: P-values

Define the following:

- Hypothesis indicators $H^m = (H_1, \ldots, H_m)$ with $H_j = 1\{\mathbb{P} \notin \mathcal{M}_j\}$.
- Set of true nulls $S_0 \equiv S_0(\mathbb{P}) = \{j \in S\colon\ H_j = 0\}$ where $S = \{1, \ldots, m\}$.
- Test statistics $Z_j = Z_j(X_1, \ldots, X_n)$ for H_{0j}, for each $j \in S$.
- P-values $P^m = (P_1, \ldots, P_m)$. Let $P_W = (P_i\colon\ i \in W \subset S)$.
- Ordered p-values $P_{(1)} < \cdots < P_{(m)}$.

Assume $P_j \mid H_j = 0 \sim \text{Unif}(0, 1)$.

Initially assume that the P_js are independent, but will weaken this later.

Testing Framework: Rejection Sets

We call a *rejection set* any $R = R(P^m) \subset S$ that indexes the rejected null hypotheses H_{0j}.

In practice, R will usually be of the form $\{j \in S \colon P_j \leq T\}$ for a random *threshold* $T = T(P^m)$.

Want to define rejection sets that control specified error criteria.

Example: say that R controls k-familywise error at level α if

$$\mathbb{P}\Big\{\#(R \cap S_0(\mathbb{P})) > k\Big\} \leq \alpha,$$

where $\#(B)$ denotes the number of points in a set B.

The False Discovery Proportion

Define the false discovery proportion (FDP) for each rejection set R by

$$\Gamma(R) \equiv \mathsf{FDP}(R) = \frac{\text{false rejections}}{\text{rejections}} = \frac{\sum_{j=1}^{m}(1 - H_j)1\left\{R \ni j\right\}}{\sum_{j=1}^{m} 1\left\{R \ni j\right\}}$$

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

The false discovery rate $\mathsf{FDR}(R)$ is defined by

$$\mathsf{FDR}(R) = \mathsf{E}(\Gamma(R)).$$

If R is defined by a threshold T, write $\Gamma(T)$ interchangeably, with $\Gamma(t)$ corresponding to a fixed threshold t.

Specifying the function $t \mapsto \overline{\Gamma}(t)$ is sufficient to determine the entire envelope for rejection sets defined by a threshold.

Plan

1. Set Up
   - Testing Framework
   - FDR and FDP

2. Exceedance Control for FDP
   - Inversion and the $P_{(k)}$ test
   - Power and Optimality
   - Combining $P_{(k)}$ tests
   - Augmentation

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

Confidence Envelopes and Exceedance Control

Goal: find a rejection set R such that

$$\mathbb{P}\Big\{\Gamma(R) > \gamma\Big\} \leq \alpha$$

for specified $0 < \alpha, \gamma < 1$. We call such an R a (γ, α) rejection set.

Our main tool for finding these are *confidence envelopes*.

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

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

Confidence Envelopes and Thresholds

It's easiest to visualize a $1 - \alpha$ confidence envelope for FDP as a random function $\overline{\text{FDP}}(t)$ on $[0, 1]$ such that

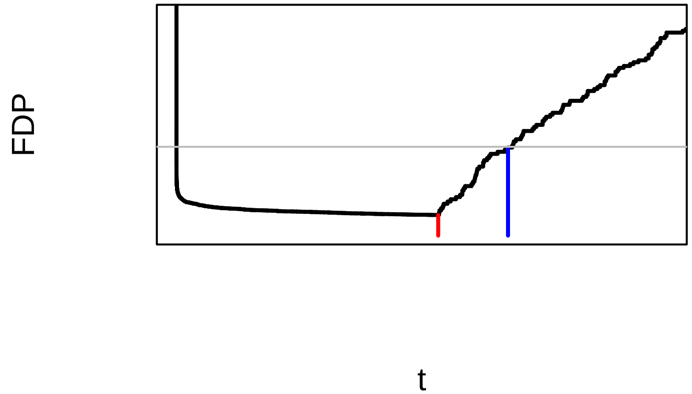
$$\mathbb{P}\Big\{\text{FDP}(t) \le \overline{\text{FDP}}(t) \text{ for all } t\Big\} \ge 1 - \alpha.$$

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

Two special cases have proven useful:

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

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



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.

Allow 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

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

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

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

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

$$\overline{\Gamma}(R) \leq \gamma.$$

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

It follows that

1. $\overline{\Gamma}$ is a $1 - \alpha$ confidence envelope for FDP, and
2. R is a (γ, α) 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 $\overline{\Gamma}$ should be close to Γ and thus result in rejection sets with high power.

  B. (Computational Tractability). The envelope $\overline{\Gamma}$ should be easy to compute.

- We 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

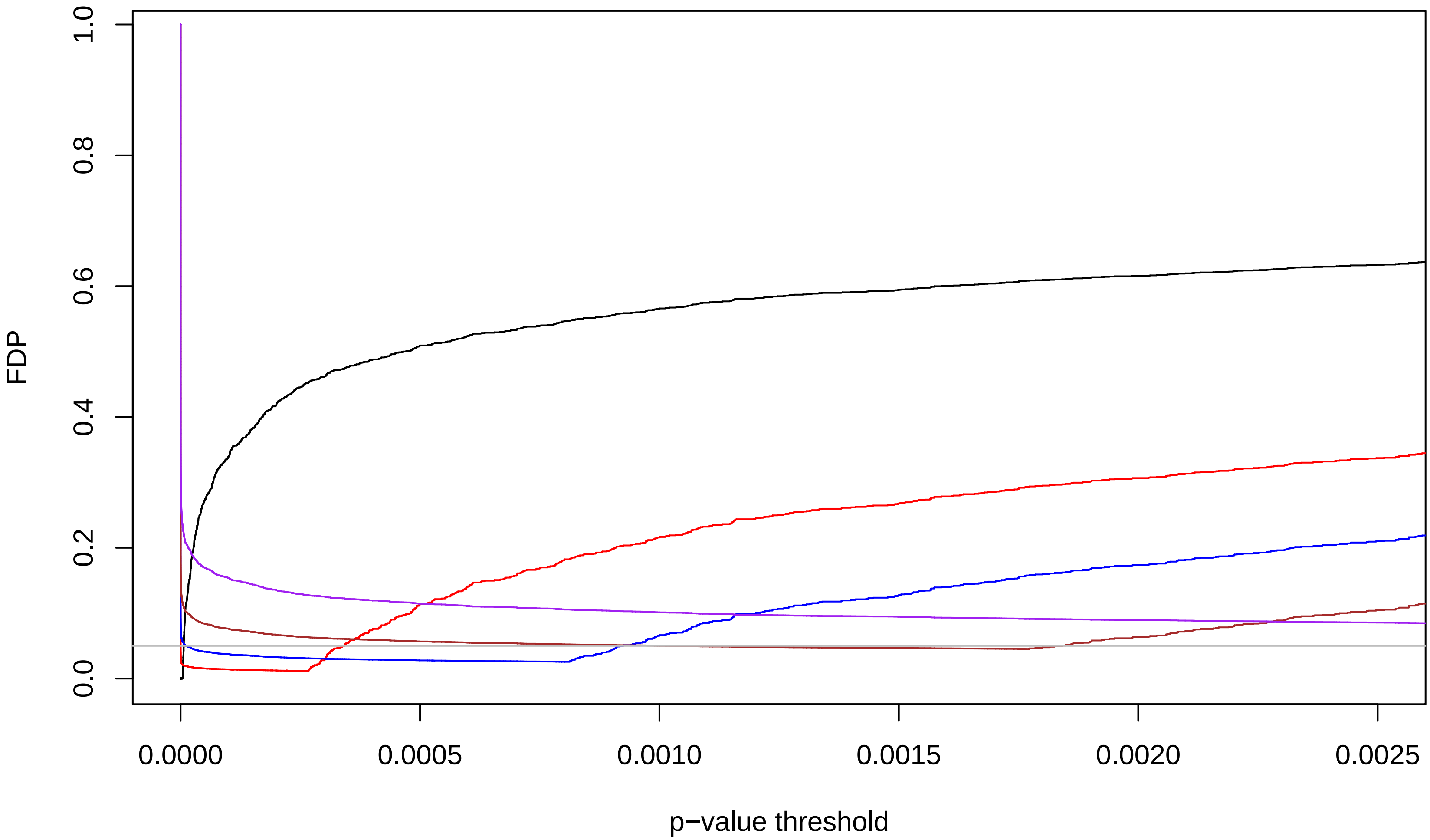
- In contrast, using the kth order statistic as a one-sided test statistic meets both desiderata.
  - For small k, these are sensitive to departures that have a large impact on FDP. Good "power."
  - Computing the confidence envelopes is linear in m.

- 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.



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 $\overline{\Gamma}$ complicates analysis.

- Combine $P_{(k)}$ tests

  Let $Q_m \subset \{1, \ldots, m\}$ with cardinality q_m. Define $\overline{\Gamma} = \min_{k \in Q_m} \overline{\Gamma}_k$, where $\overline{\Gamma}_k$ 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.

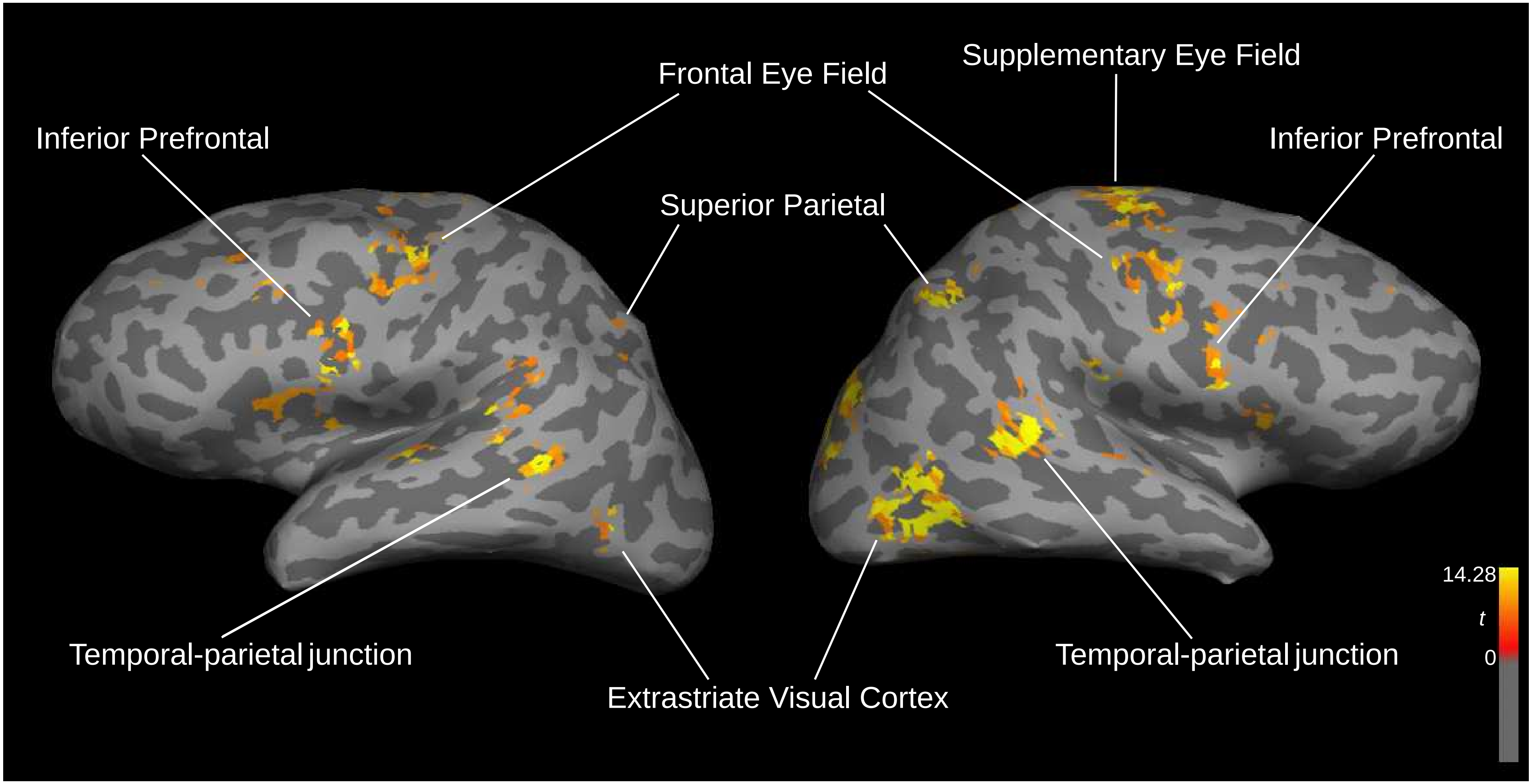
$$J_k = \min\{j : \ P_{(j)} \geq \frac{k\alpha}{m-j}\}.$$

Simulation Results

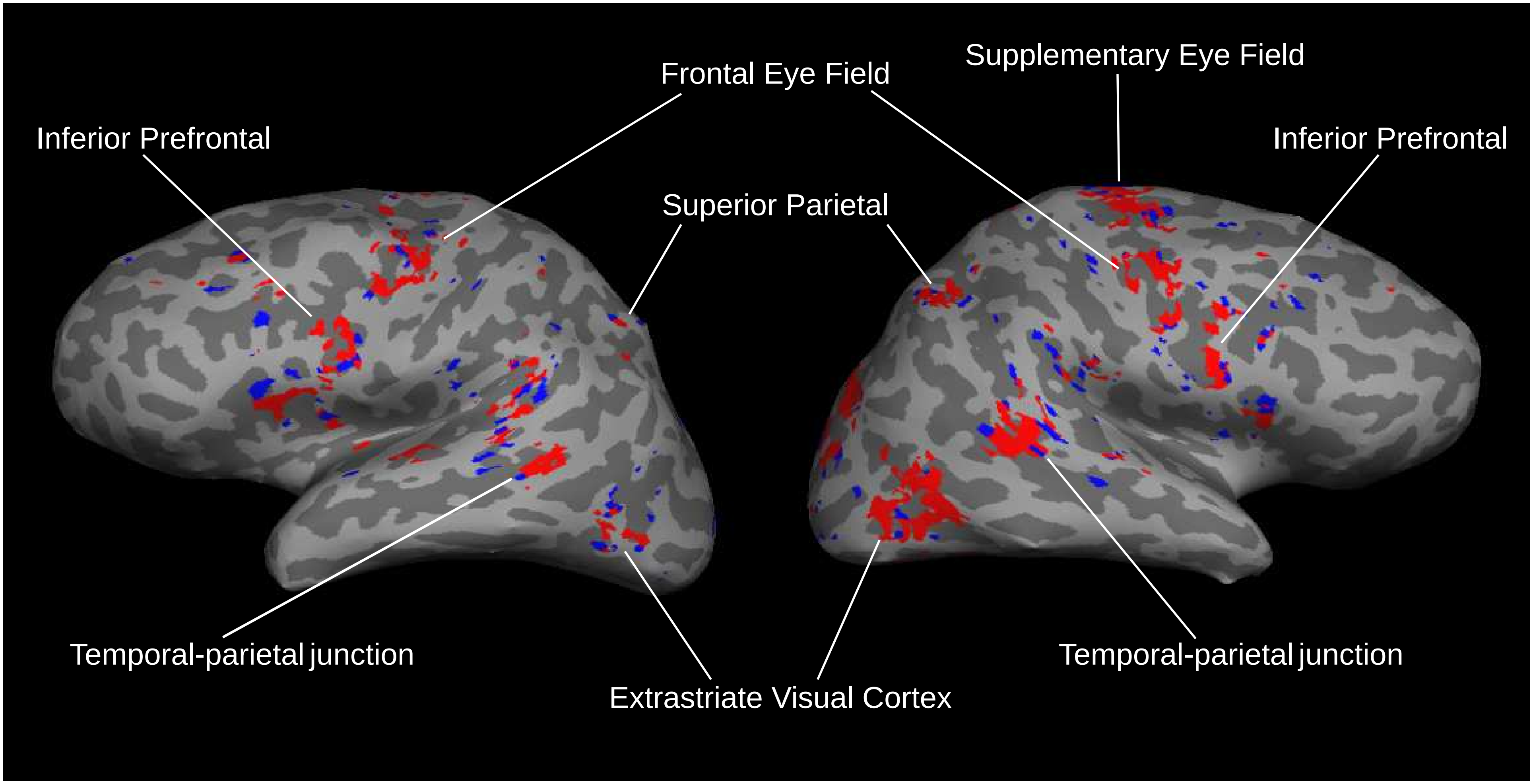
Excerpt under simple mixture model with proportion a alternatives with Normal($\theta, 1$) distribution. Here $m = 10,000$ tests, $\gamma = 0.2$, $\alpha = 0.05$.

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

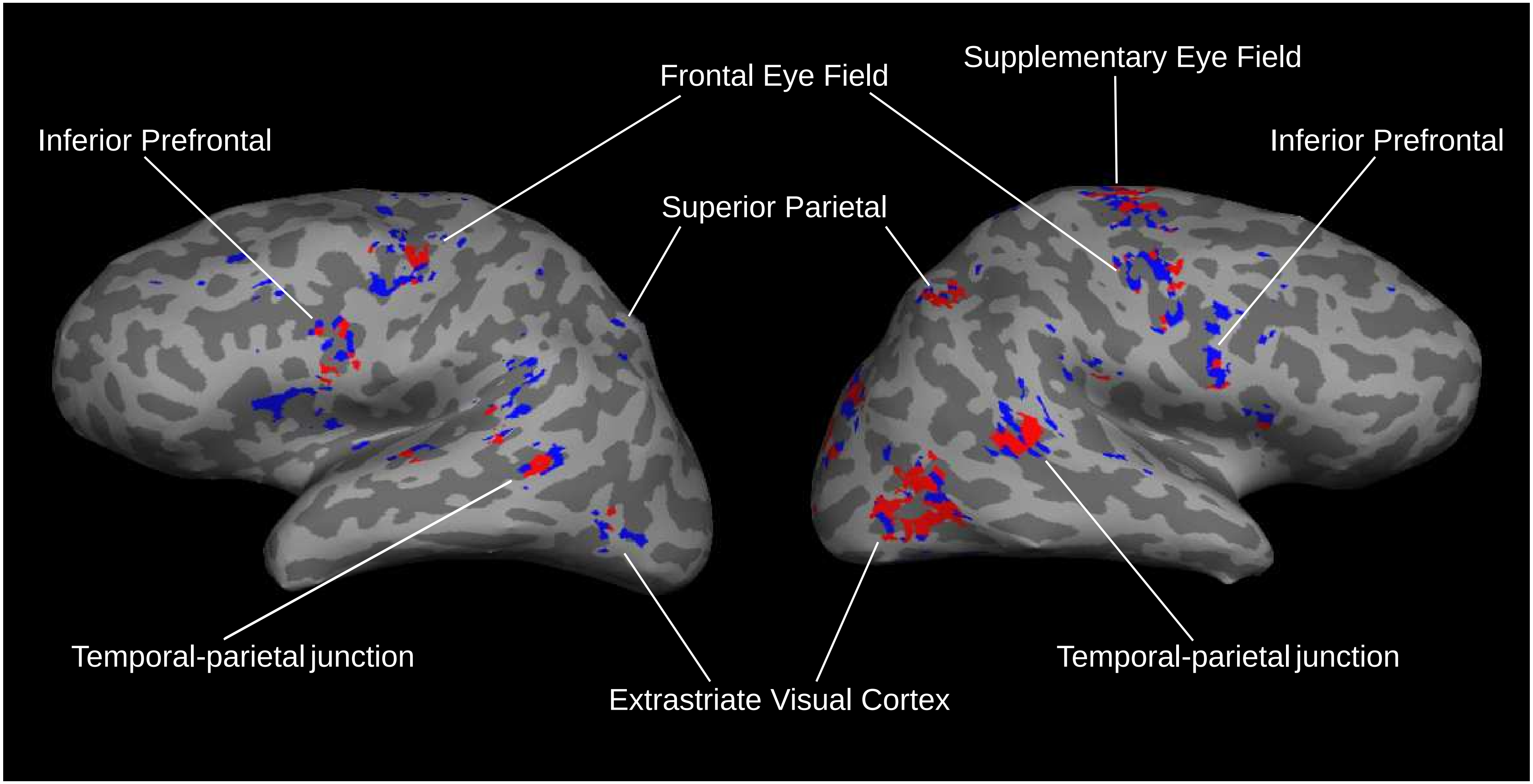
Results: (0.05,0.9) Confidence Threshold



Results: (0.05,0.9) Threshold versus BH



Results: (0.05,0.9) Threshold versus Bonferroni



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,

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

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

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

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 $R_{\mathrm{aug}} = R_{\mathrm{inv}}$.

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

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

Plan

1. Set Up
   - Testing Framework
   - FDR and FDP

2. Exceedance Control for FDP
   - Inversion and the $P_{(k)}$ test
   - Power and Optimality
   - Combining $P_{(k)}$ tests
   - Augmentation

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

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

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

  where λ is usually Lebesgue measure.

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

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

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 $\mathcal{U} = \{A \subset S\text{:}\;\; p(x(A), A) \geq \gamma\}$.

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

4. And,
$$\overline{\mathrm{FDP}}(t) = \frac{\lambda(U \cap \{s \in S : X(s) > t\})}{\lambda(\{s \in S : X(s) > t\})},$$
   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.

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:

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

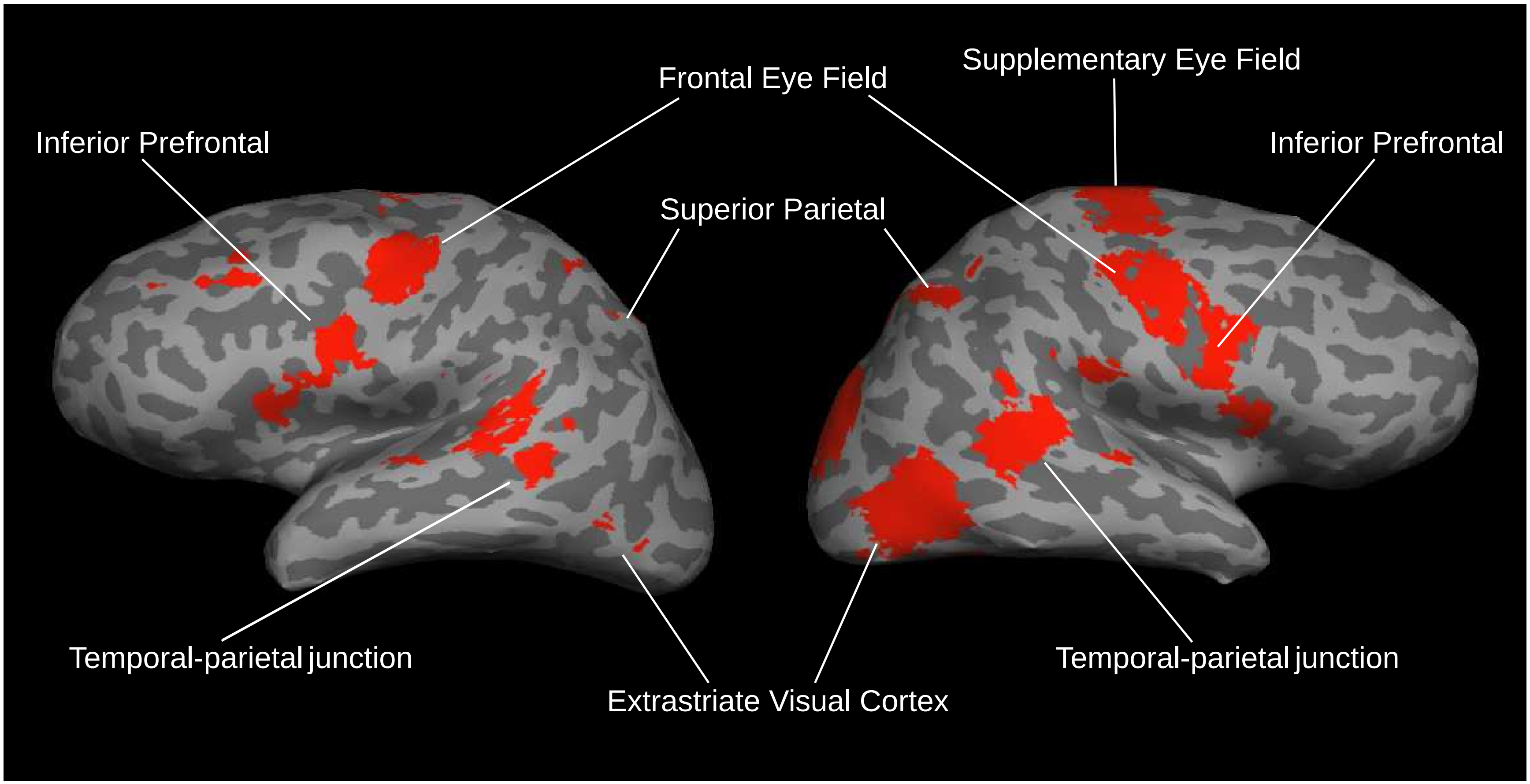
- Decompose the t-level set of X into its connected components $C_{t1}, \ldots, C_{tk_t}$.
- For each level t, let $\xi(t)$ denote the *proportion of false regions (at tolerance ϵ)* out of k_t regions.
- Then,

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

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

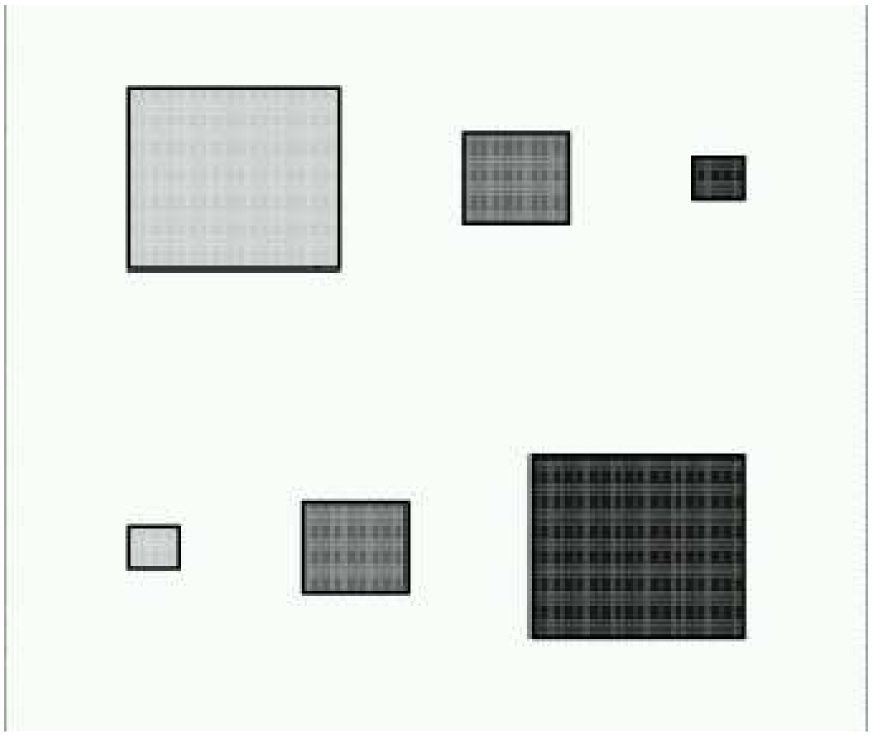
$$f(s) = \frac{\nu(s)}{\int_S \nu(u)\, du}.$$

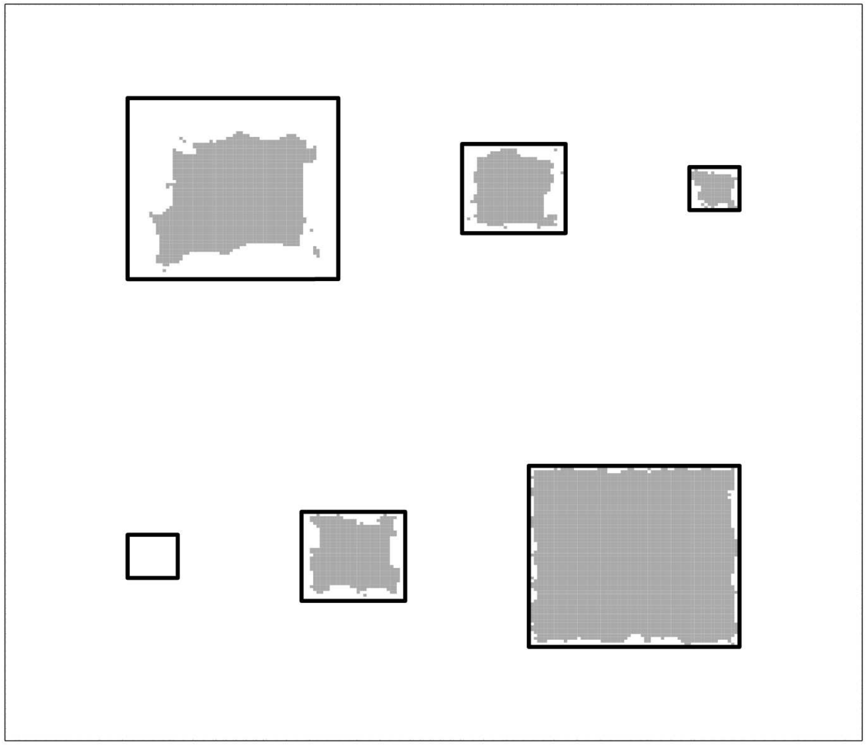
Scan statistic test for "clusters" via the statistic $T = \sup_{s \in S} N_s$.,

Our procedure:

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

Scan Statistics (cont'd)





Plan

1. Set Up
   - Testing Framework
   - FDR and FDP
2. Exceedance Control for FDP
   - Inversion and the $P_{(k)}$ test
   - Power and Optimality
   - Combining $P_{(k)}$ tests
   - Augmentation
3. False Discovery Control for Random Fields
   - Confidence Supersets and Thresholds
   - Controlling the Proportion of False Clusters
   - Scan Statistics

Take-Home Points

- Confidence thresholds have practical advantages for False Discovery Control.

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

- Works under general dependence, though there is potential gain in tailoring the procedure to the dependence structure.

- This helps with secondary inference about the structure of alternatives (e.g., controlling proportion of false regions), but a better next step is to handle that structure directly.