Doing Cosmology with Balls and Envelopes

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 Grant SES 9866147 and ITR 0104016.

Overview: Two Talks in One

- “Envelopes”
  - Situation: Performing many simultaneous hypothesis tests
  - Problem: Attain needed power while still controlling false discoveries in some principled way.
  - Approach: Bound the proportion of false discoveries among rejected nulls with high probability.
- “Balls”
  - Situation: Have noisy samples of an unknown function.
  - Problem: Make inferences about various features of the function.
  - Approach: Construct uniformly valid confidence sets for the unknown function.

Notation

$$\mathsf{E}X \equiv \langle X \rangle$$

$$\widehat{\theta} \equiv \text{estimate of } \theta$$

$$\sup \equiv \max$$

$$\inf \equiv \min$$

Also, focus on the blue stuff.

Road Map: "Envelopes"

1. The Multiple Testing Problem
   - Idea and Examples
   - Error Criteria
2. Controlling FDR
   - The Benjamini-Hochberg Procedure
   - Increasing Power
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: “Envelopes”

1. The Multiple Testing Problem
   - Idea and Examples
   - Error Criteria
2. Controlling FDR
   - The Benjamini-Hochberg Procedure
   - Increasing Power
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

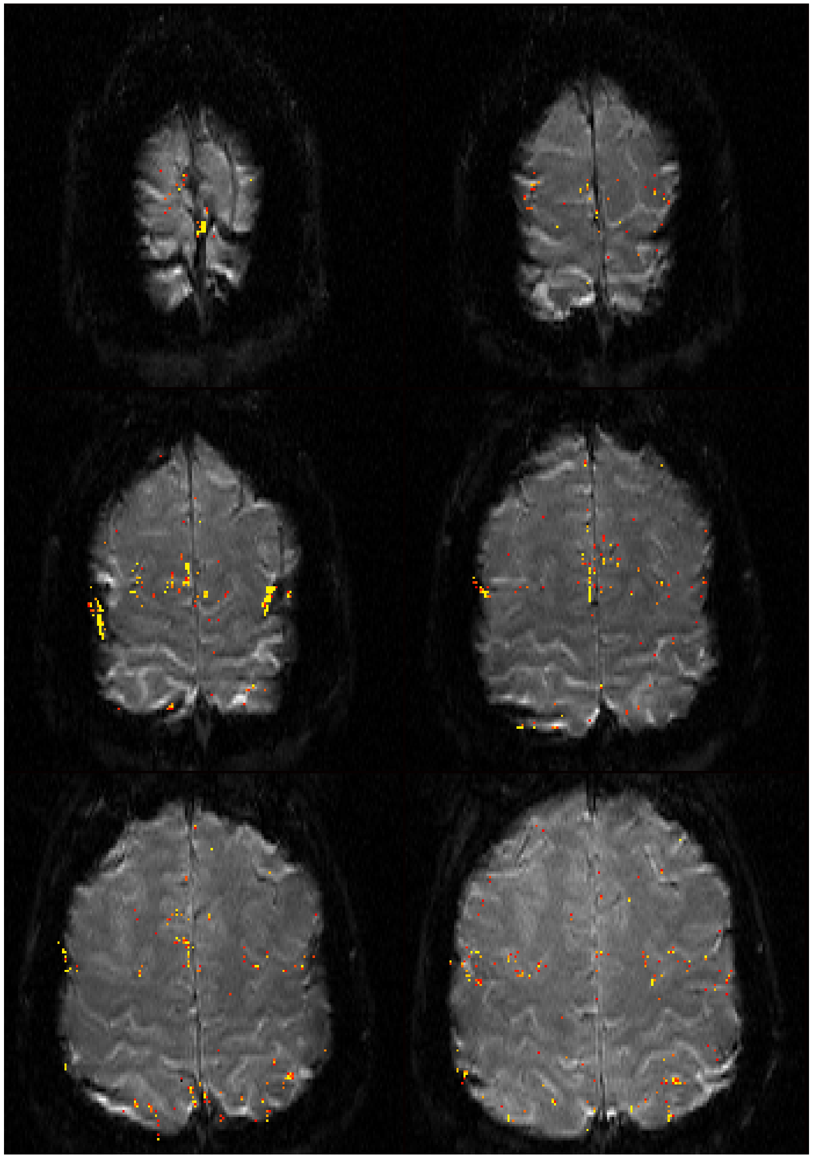
  All quantities except m, D, and N are unobserved.

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

Motivating Examples

- fMRI Data
- Astronomical Source Detection
- DNA Microarrays
- Scan Statistics

These all involve many thousands of tests and interesting spatial structure.



How to Choose a Threshold?

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

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

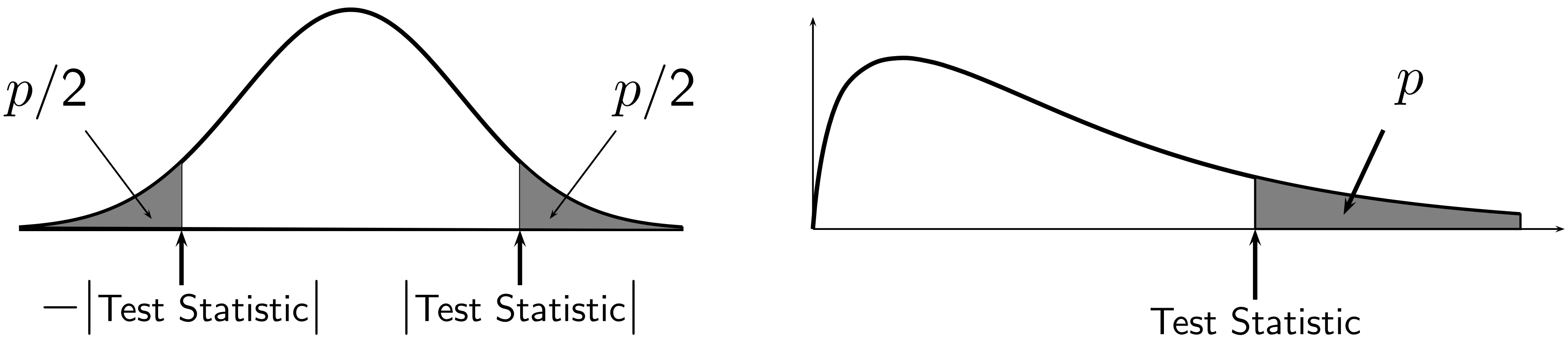
- False Discovery Control
  - e.g.: Benjamini & Hochberg (BH, 1995, 2000): False Discovery Rate (FDR)
  - Guarantees $\mathrm{FDR} \equiv \mathrm{E}\left(\frac{FD}{D}\right) \leq \alpha$

Road Map: “Envelopes”

1. The Multiple Testing Problem
   - Idea and Examples
   - Error Criteria
2. Controlling FDR
   - The Benjamini-Hochberg Procedure
   - Increasing Power
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

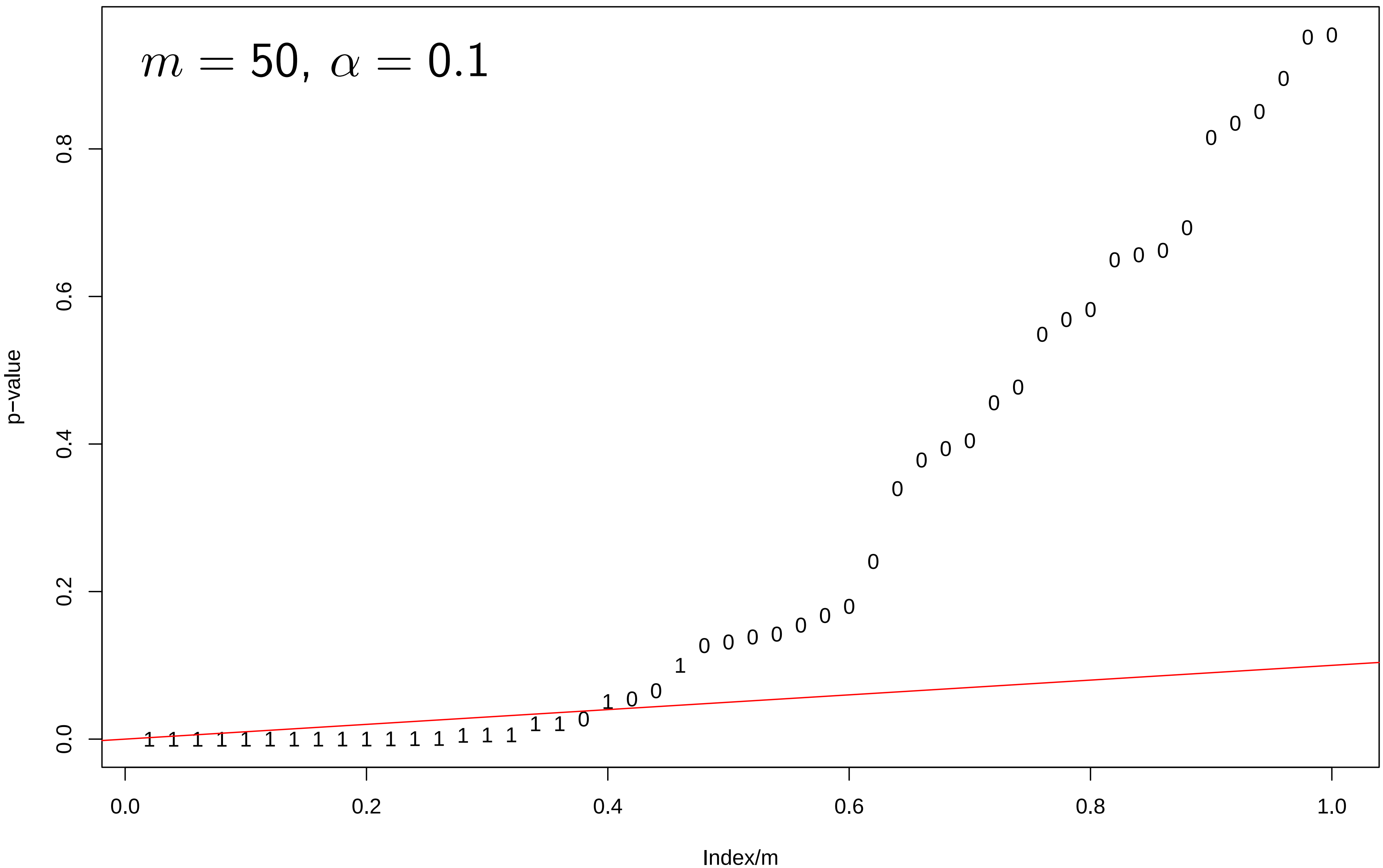
- Convenient to work with p-values



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

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

The Benjamini-Hochberg Procedure (cont'd)



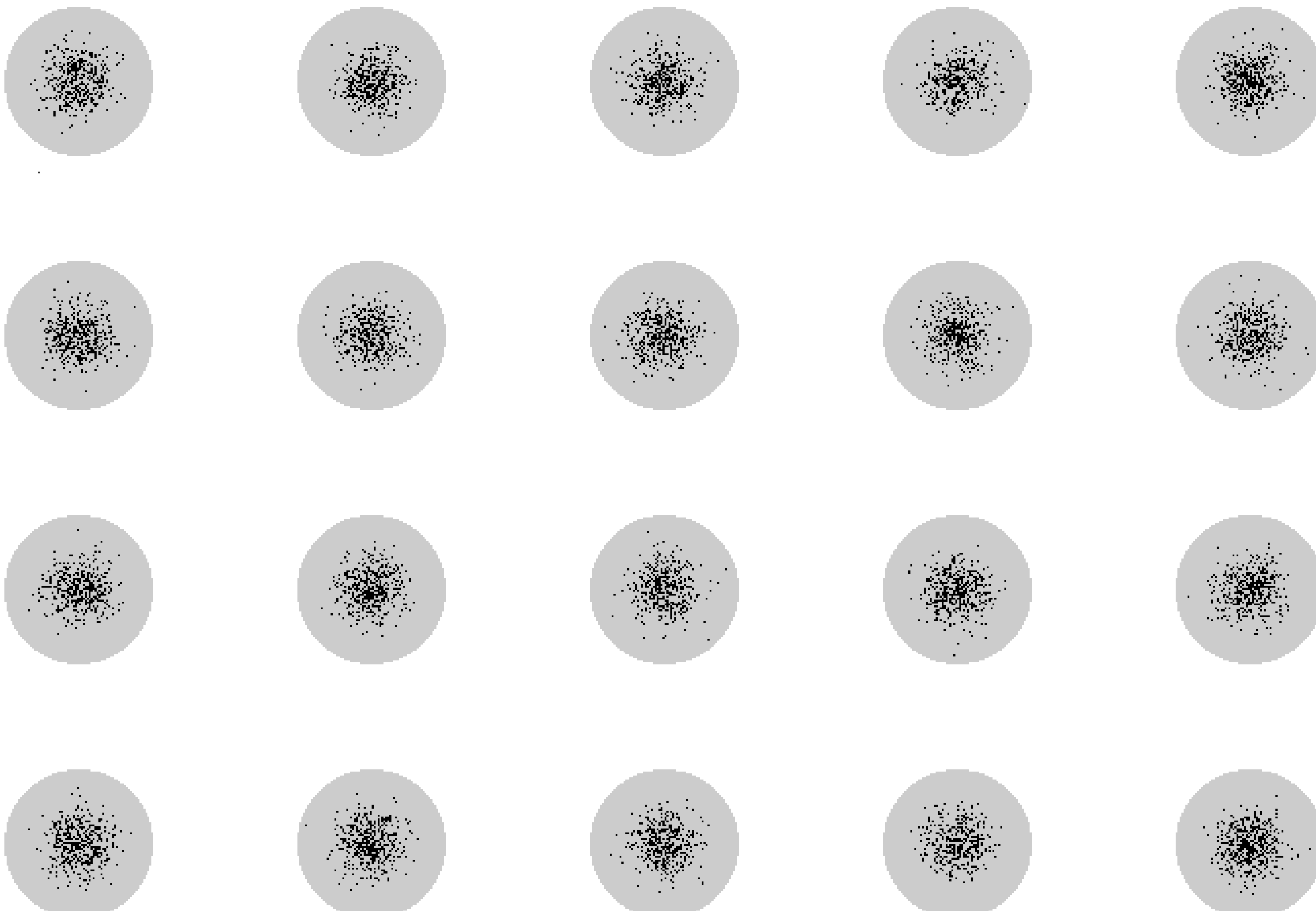
The Benjamini-Hochberg Procedure (cont'd)

- BH guarantees that FDR $\equiv \mathsf{E}\left(\frac{FD}{D}\right) \leq \frac{T_0}{m}\alpha$.
- Gives more power than Bonferroni, fewer Type I errors than uncorrected testing.
- If $\widehat{G}$ is the empirical cdf of the m p-values, $\widehat{G}(P_{(i)}) = i/m$, so

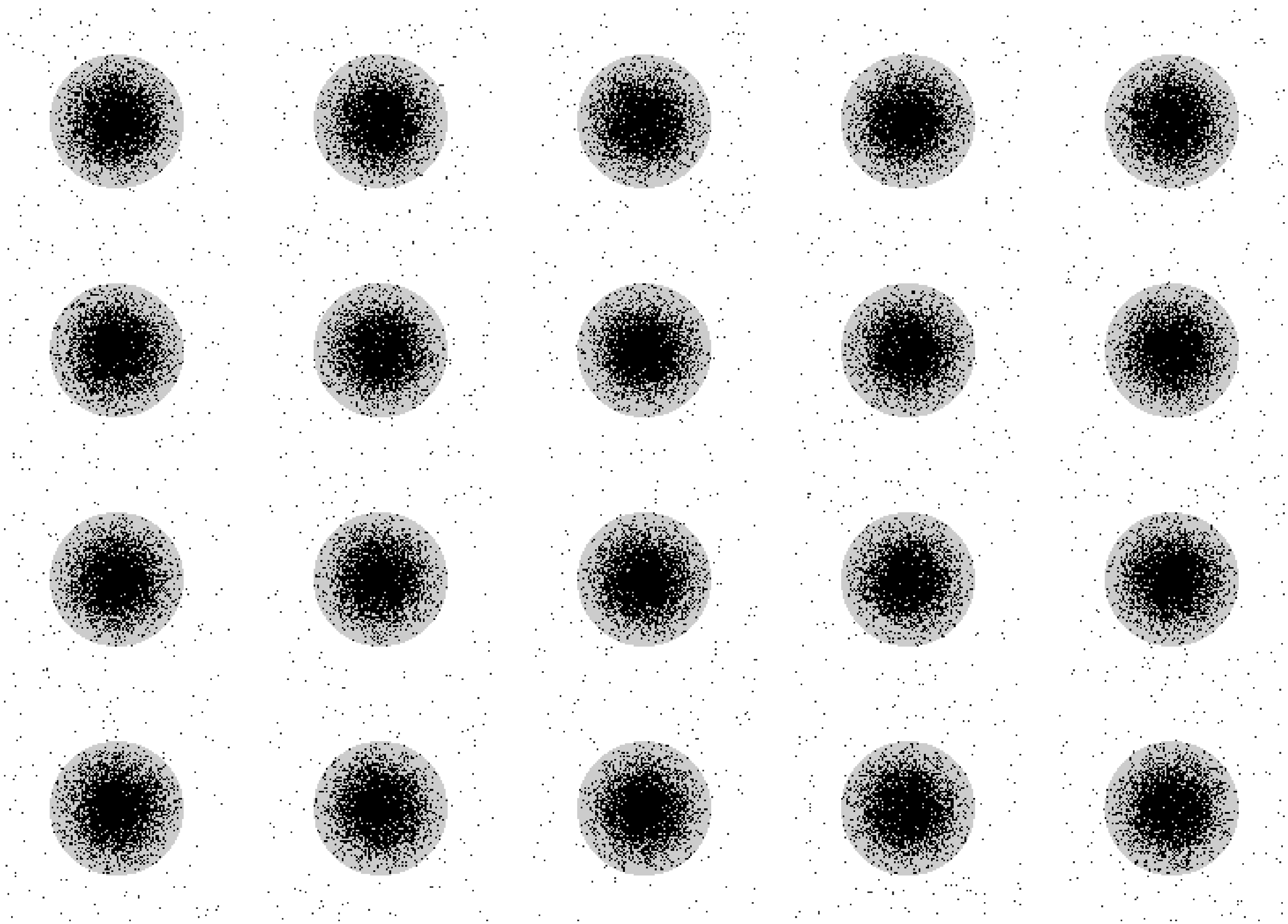
$$T_{\mathrm{BH}} = \max\left\{t\colon\ \widehat{G}(t) = \frac{t}{\alpha}\right\} = \max\left\{t\colon\ \frac{t}{\widehat{G}(t)} \leq \alpha\right\}.$$

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

Simulated Example: Bonferroni

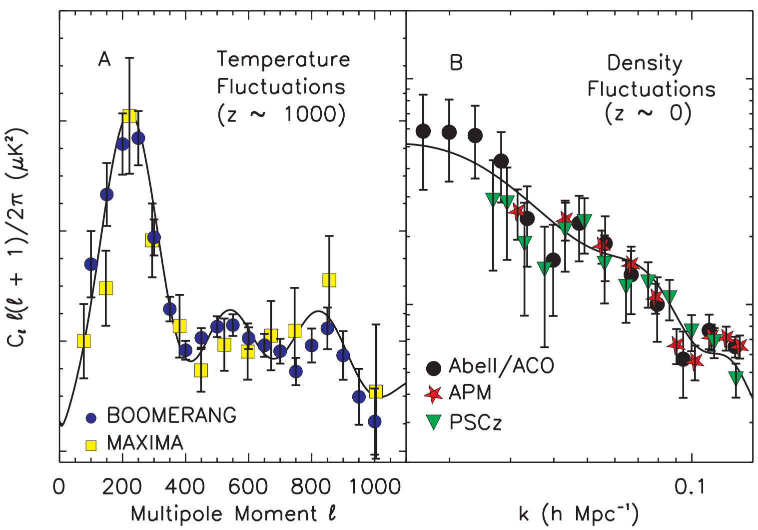


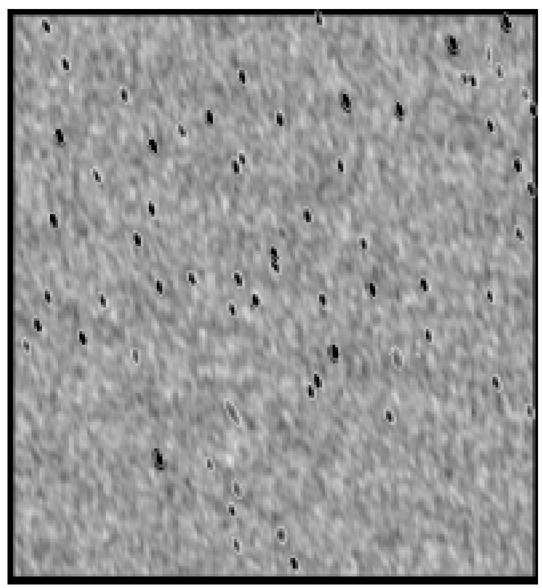
Simulated Example: BH

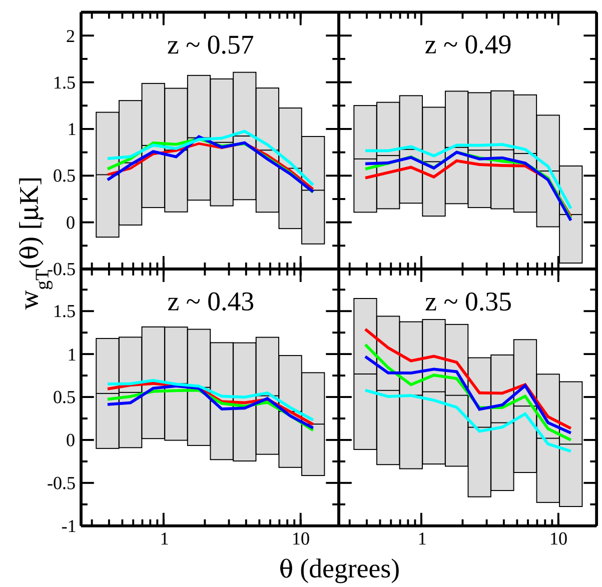


Astronomical Examples (PiCA Group)

- 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, drawn independently from

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

where:

1. $0 \le a \le 1$ 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].

- Let $H^m = (H_1, \ldots, H_m)$ where $H_i = 0$ (or 1) if the i^{th} null hypothesis is true (or false).

  Assume the H_is are independent Bernoulli$\langle a \rangle$, but everything works with the H_i's fixed as well.

Mixture Model for Multiple Testing (cont'd)

- We assume the following model (Efron et al., 2001; Efron, 2003):

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

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

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

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

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 $P_i \leq T(P^m)$. 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 $\mathsf{FDP}(t)$ and False Nondiscovery Proportion $\mathsf{FNP}(t)$.

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

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

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 $\mathsf{FDR} \leq (1 - a)\alpha \leq \alpha$.
- This bound holds at least under "positive dependence".
- Replacing α by $\alpha / \sum_{i=1}^{m} 1/i$ extends FDR bound to any distribution, but this is typically *very* conservative.

Selected Recent Work on FDR

Abromovich, Benjamini, Donoho, & Johnstone (2000)

Benjamini & Hochberg (1995, 2000)

Benjamini & Yekutieli (2001)

Efron, Tibshirani, & Storey (2001)

Efron, Tibshirani, Storey, & Tusher (2002)

Finner & Roters (2001, 2002)

Hochberg & Benjamini (1999)

Genovese & Wasserman (2001,2002,2003)

Pacifico, Genovese, Verdinelli, & Wasserman (2003)

Sarkar (2002)

Seigmund, Taylor, & Storey (2003)

Storey (2002,2003)

Storey & Tibshirani (2001)

Tusher, Tibshirani, Chu (2001)

Yekutieli & Benjamini (2001)

Road Map: "Envelopes"

1. The Multiple Testing Problem
   - Idea and Examples
   - Error Criteria
2. Controlling FDR
   - The Benjamini-Hochberg Procedure
   - Increasing Power
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

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

  In practice, it would be useful to control quantiles of FDP.

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

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

  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 $\overline{\mathrm{FDP}}(t)$ on $[0,1]$ such that

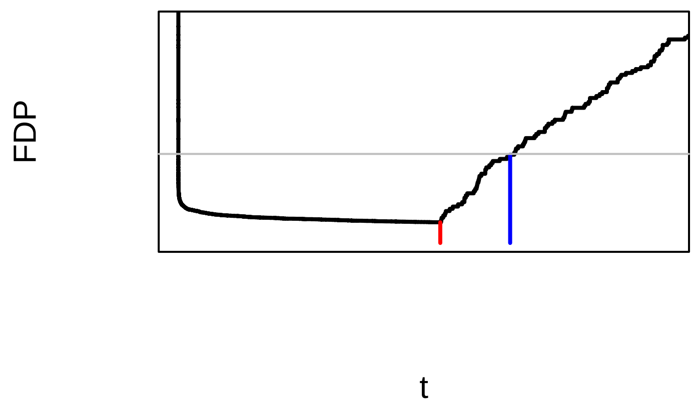
$$\mathrm{P}\Big\{\mathrm{FDP}(t) \le \overline{\mathrm{FDP}}(t) \text{ for all } t\Big\} \ge 1-\gamma.$$

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

  Two special cases have proven useful.

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

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



Exact Confidence Envelopes

- Short version: take max FDP over all subsets that look Uniform.
- 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
$$\mathcal{G}_\gamma(p^m) = \left\{ \mathrm{FDP}(\cdot; h^m, p^m)\colon\ h^m \in \mathcal{U}_\gamma(p^m) \right\}$$
$$\mathcal{M}_\gamma(p^m) = \left\{ m_0(h^m)\colon\ h^m \in \mathcal{U}_\gamma(p^m) \right\}.$$

Exact Confidence Envelopes (cont'd)

- Short version: it works.
- 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 sup 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.
- Want an automatic way to choose a good test
- 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.
- Traditional uniformity tests, such as the (one-sided) Kolmogorov-Smirnov (KS) test, *do not meet both conditions*.

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

The $P_{(k)}$ Tests

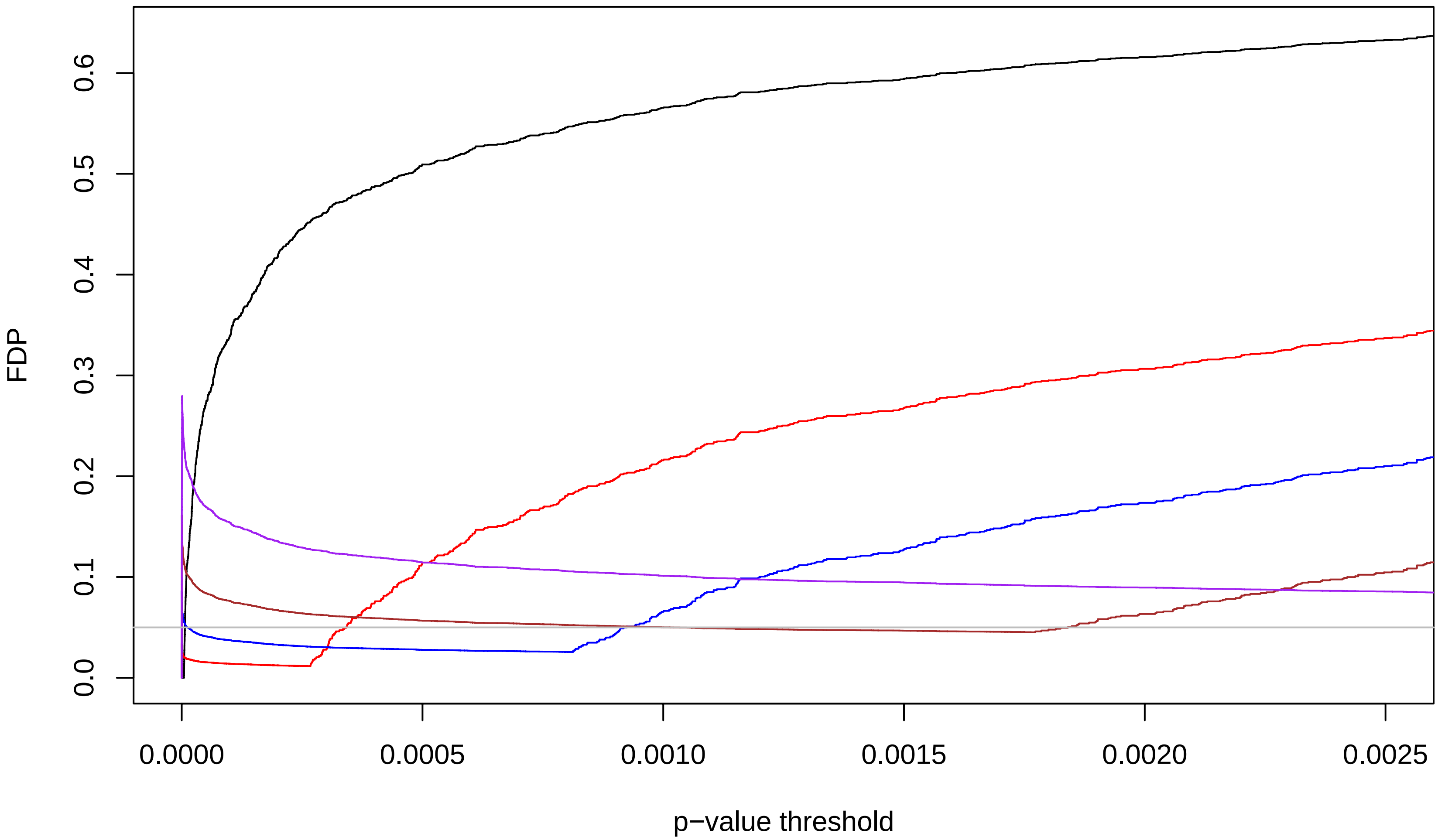
- In contrast, using the kth order statistic as a one-sided test statistic meets both desiderata.
  - For small k, these are sensitive to departures that have a large impact on FDP. Good "power."
  - Computing the confidence envelopes is linear in m.

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

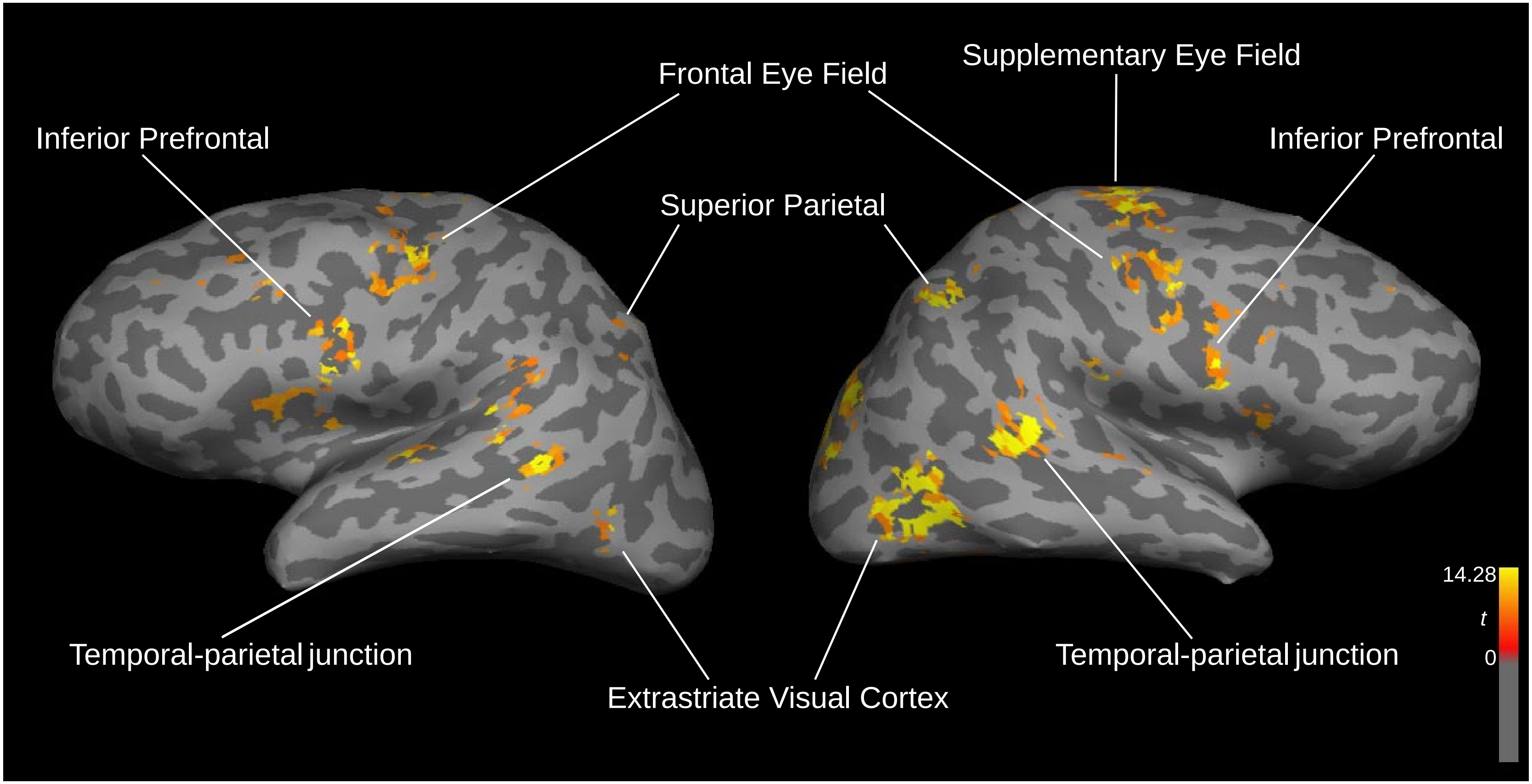
  They form a sub-family of weighted, one-sided KS tests.

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

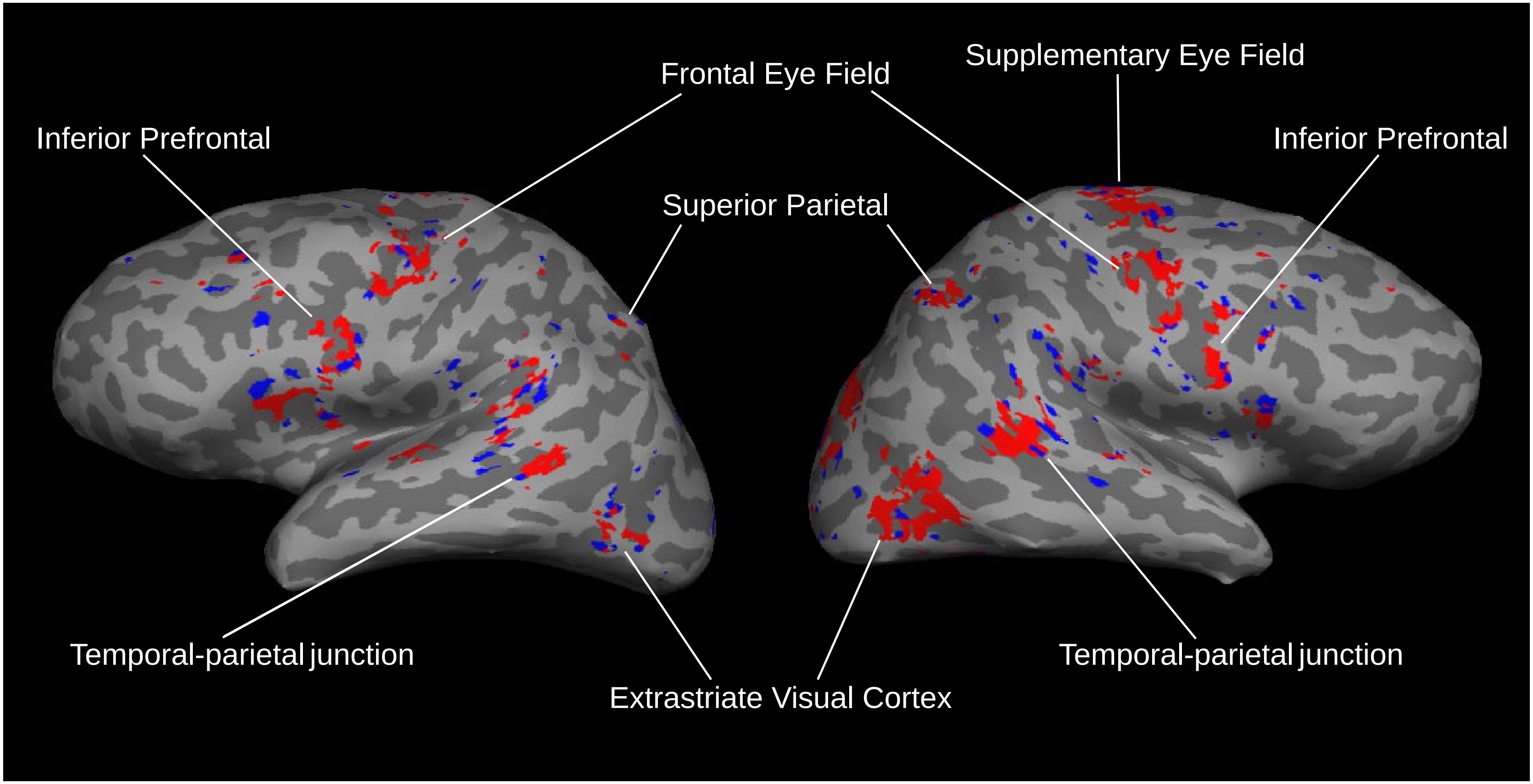
For $k = 1, 10, 25, 50, 100$, with 0.05 FDP level marked.



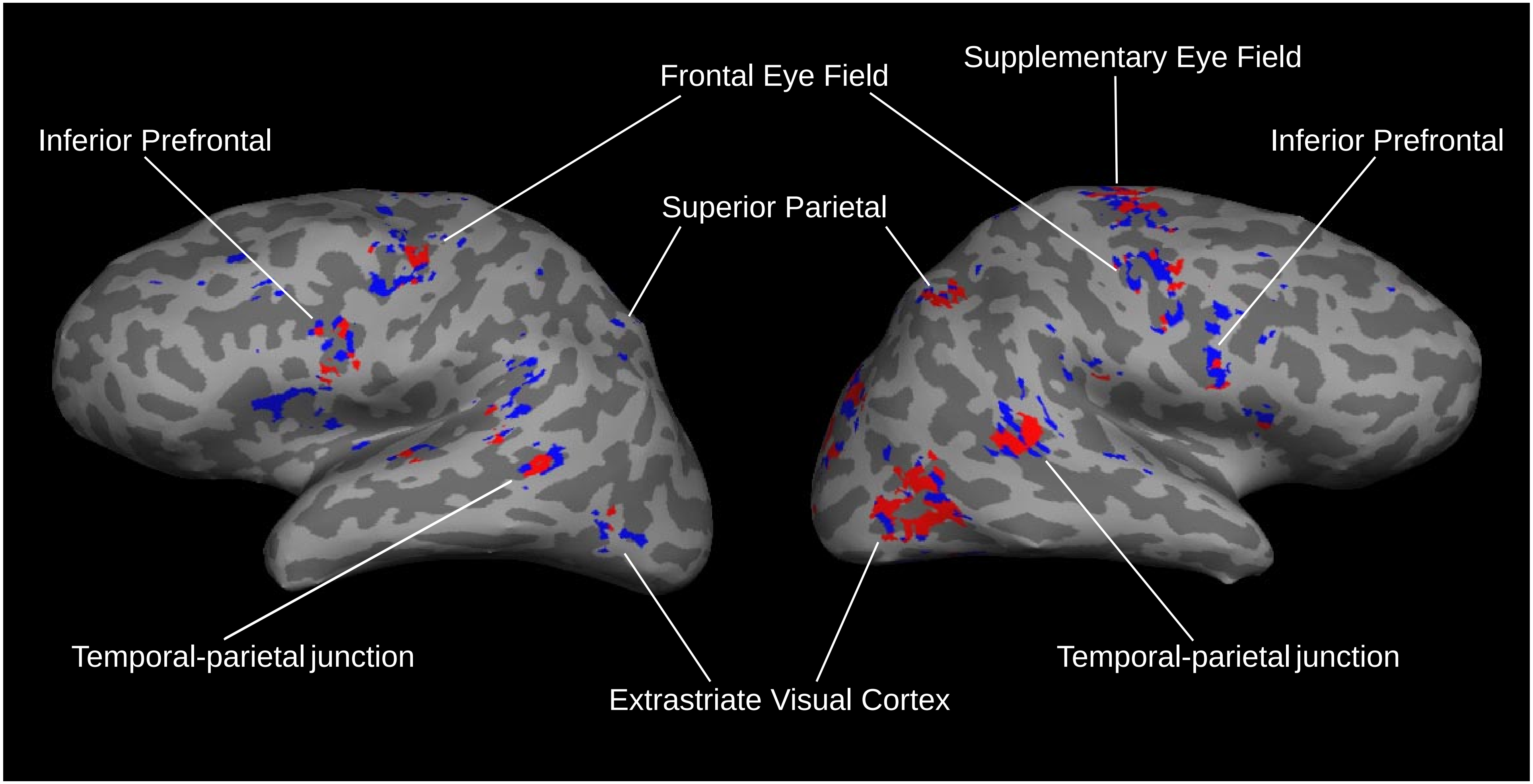
Results: (0.05,0.9) Confidence Threshold



Results: (0.05,0.9) Threshold versus BH



Results: (0.05,0.9) Threshold versus Bonferroni



Choosing k

- Direct (Simulation) Approach

  Simulate from pre-specified parametric family or mixtures of these.

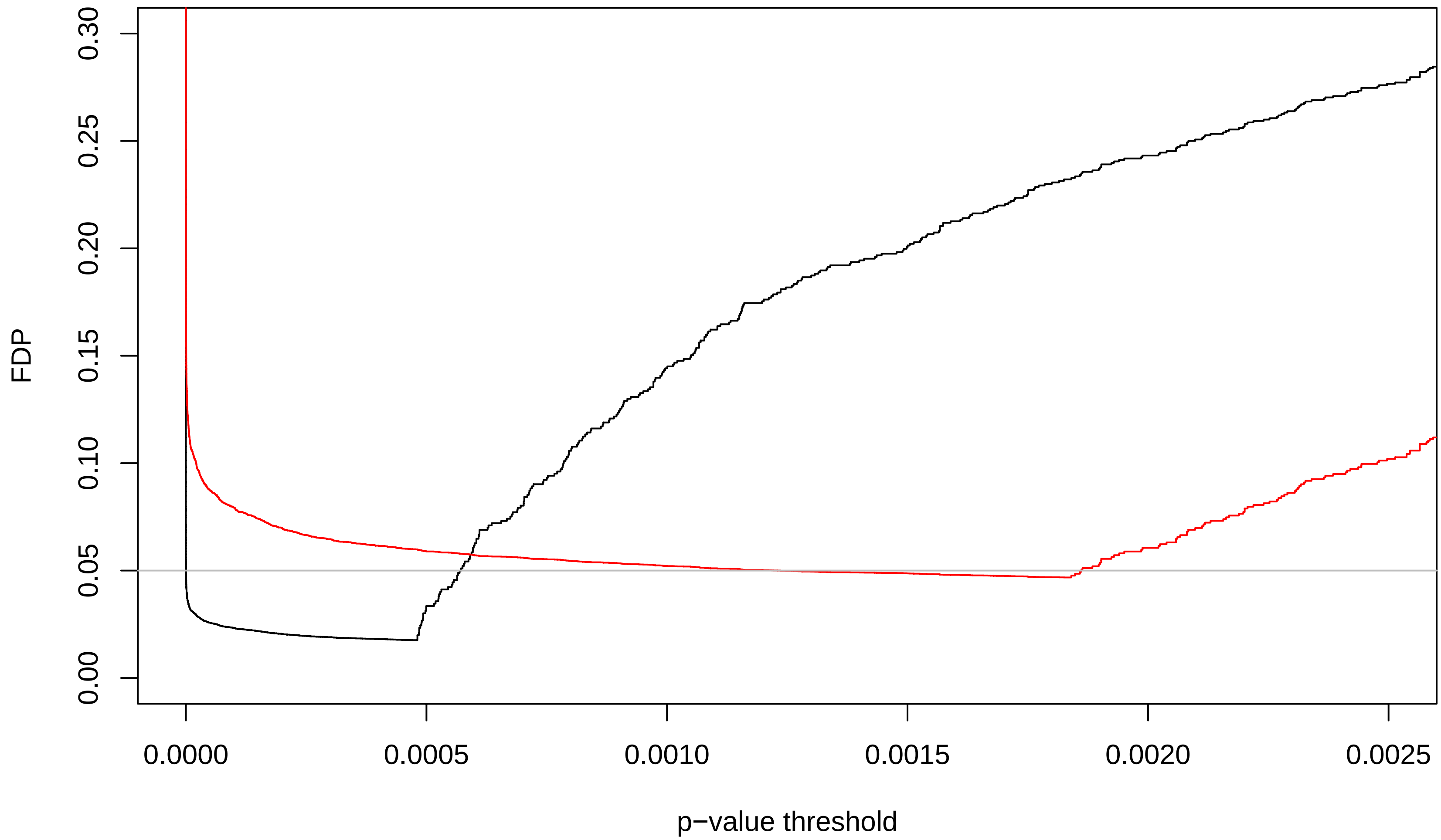
  Compute the optimal k, $k^*(\theta, m)$.

- 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

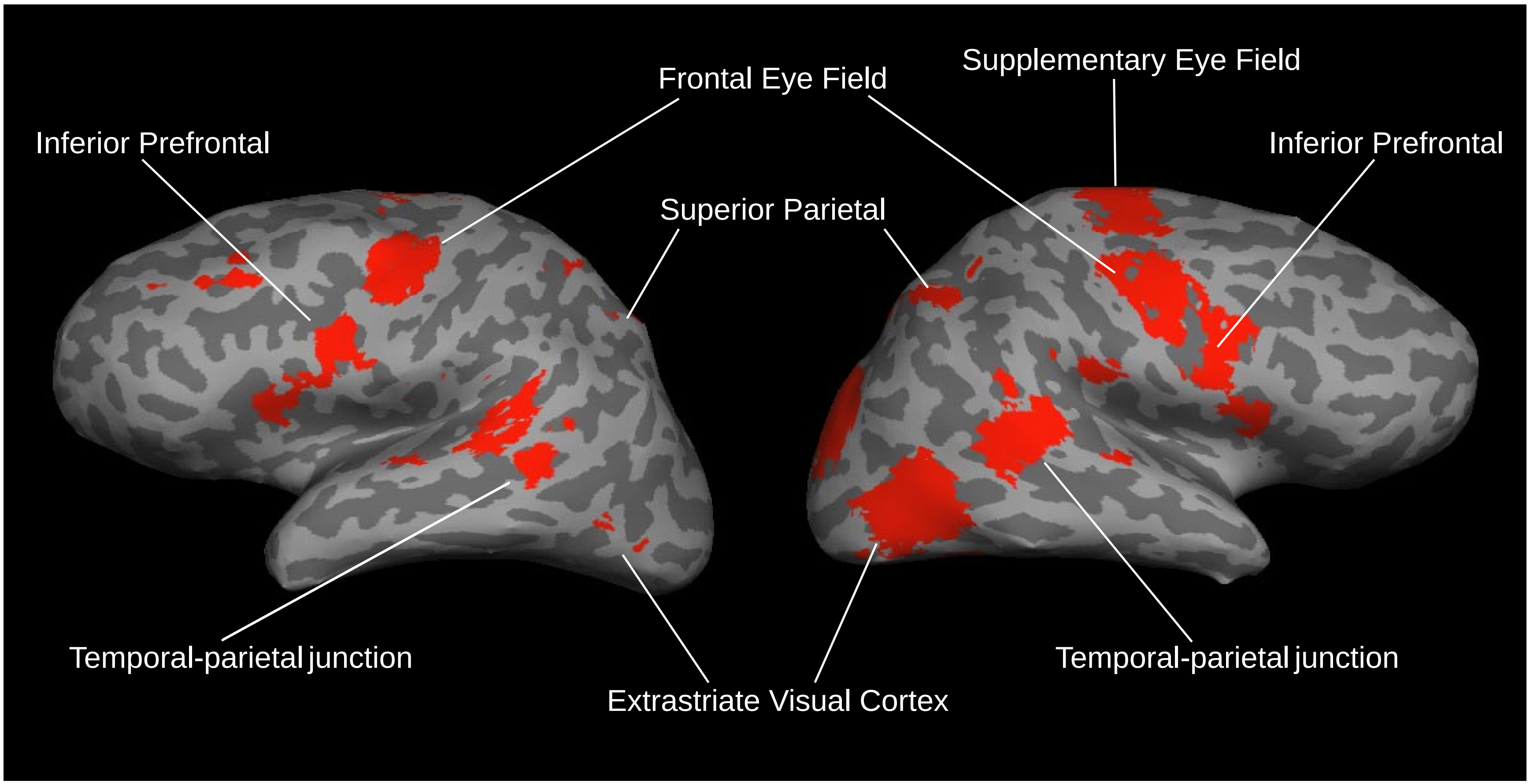


Road Map: “Envelopes”

1. The Multiple Testing Problem
   – Idea and Examples
   – Error Criteria
2. Controlling FDR
   – The Benjamini-Hochberg Procedure
   – Increasing Power
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

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: “Envelopes”

- Confidence thresholds have 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, we care about clusters/regions/sources not “pixels”. Current methods ignore spatial information.

  Controlling proportion of false regions is a first step.

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

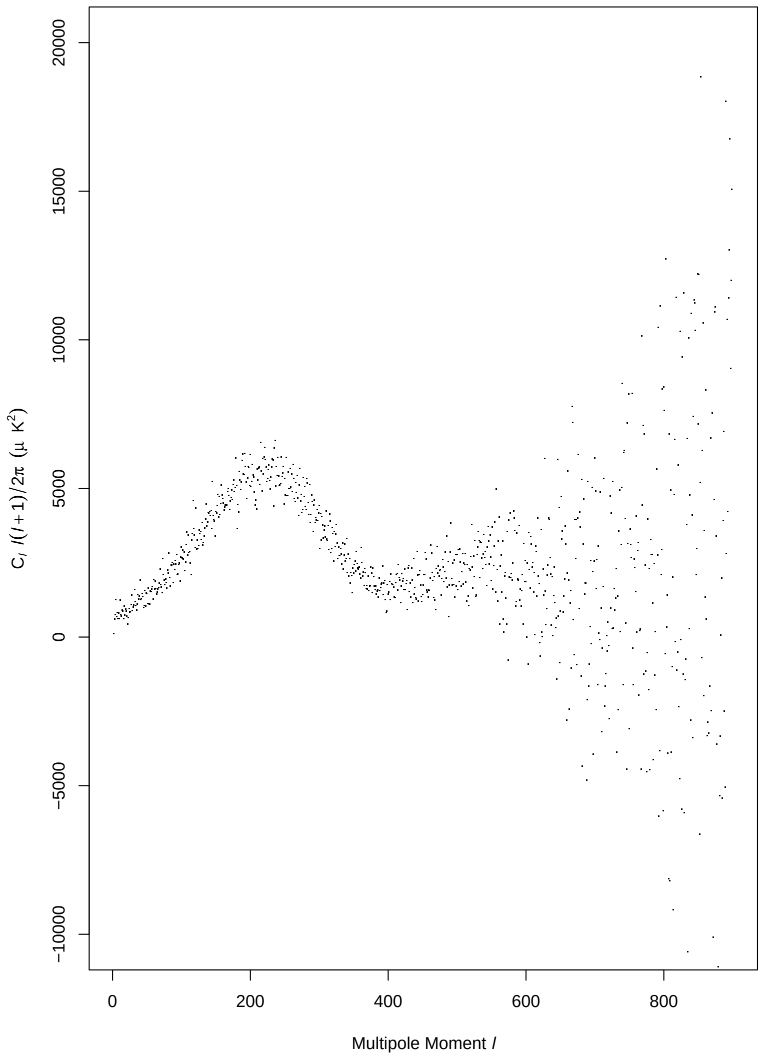
Road Map: “Balls”

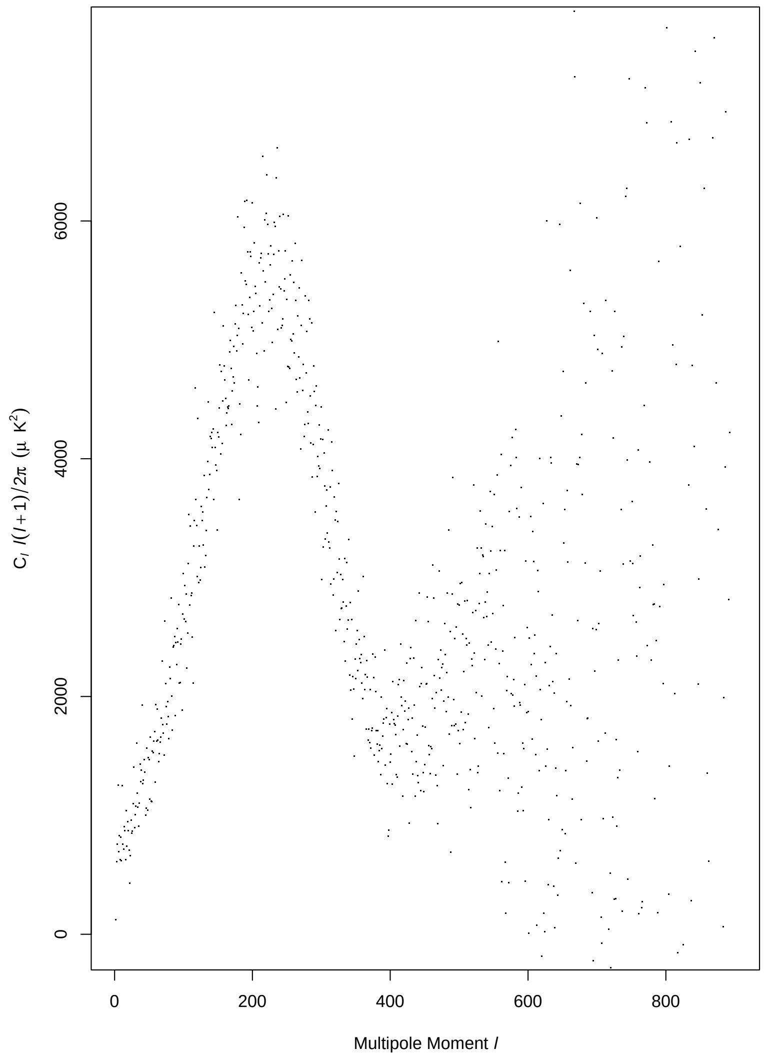
1. Inferences about Functions
   - CMB Spectrum Example
   - The Statistical Problem in General Form
   - Criteria for Effective Inferences
2. Nonparametric Confidence Balls
   - Features and Extensions
3. Keeping Your Eyes on the Ball
   - Parametric Probes
   - Model Checking
   - Confidence Catalogs

Road Map: “Balls”

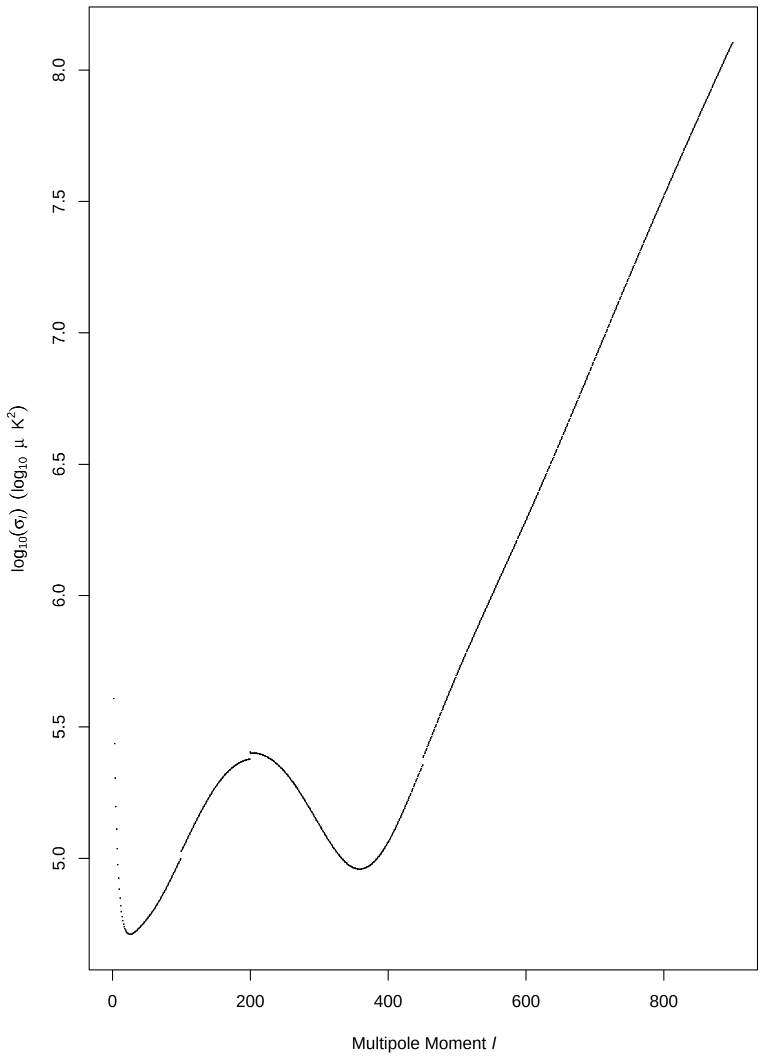
1. Inferences about Functions
   - CMB Spectrum Example
   - The Statistical Problem in General Form
   - Criteria for Effective Inferences
2. Nonparametric Confidence Balls
   - Features and Extensions
3. Keeping Your Eyes on the Ball
   - Parametric Probes
   - Model Checking
   - Confidence Catalogs

CMB Power Spectrum: WMAP Data



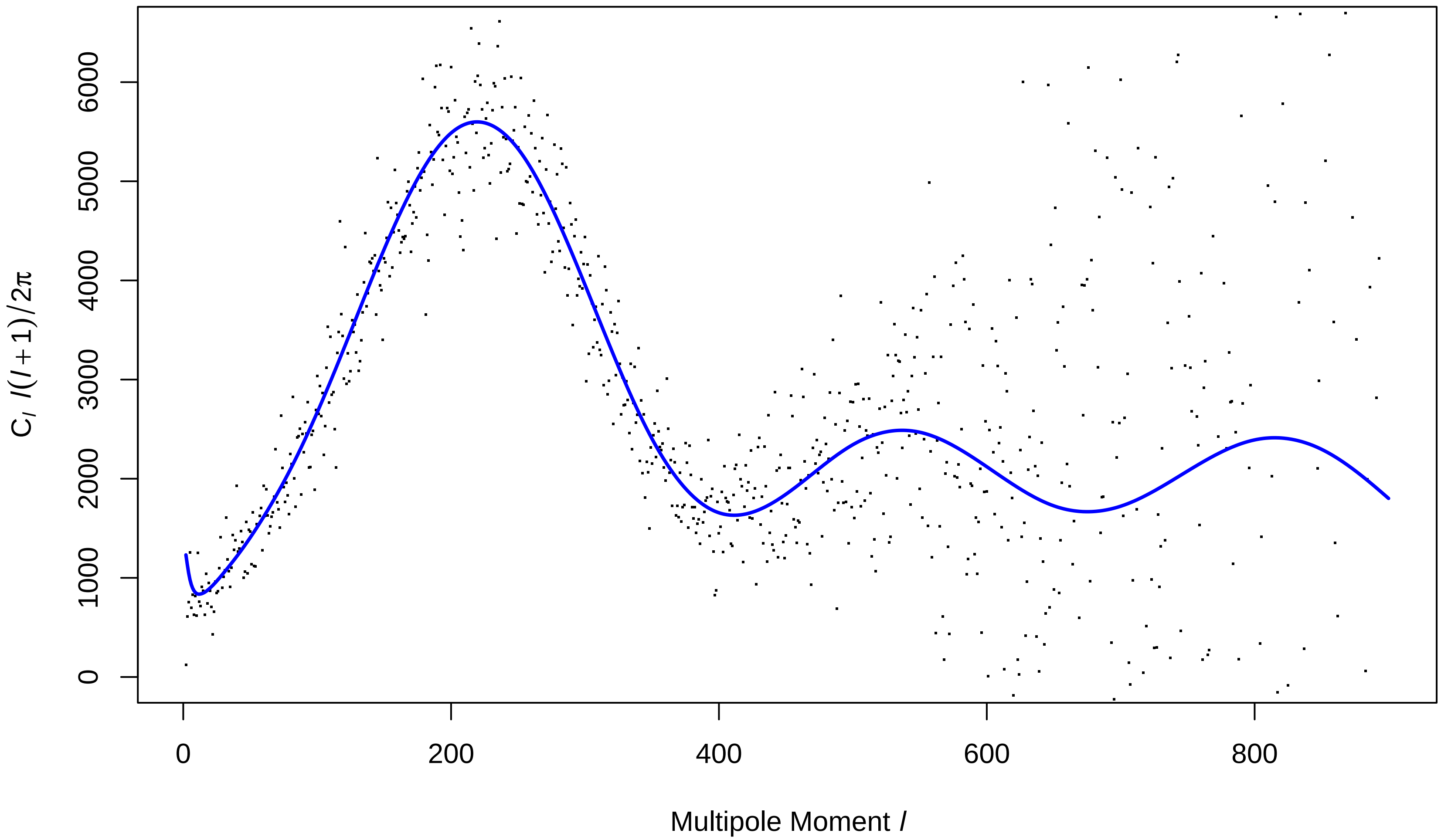


CMB Power Spectrum: WMAP Variances



Noise is correlated and heteroskedastic

CMB Power Spectrum: Models



- 11(7)-dimensional model maps cosmological parameters to spectra.
- Ultimate goal: inferences about these cosmological parameters.
- Subsidiary goal: identify location, height, widths of peaks

The Statistical Problem

- Observe noisy samples of an unknown function.

  Data of the form

  $$Y_i = f(x_i) + \epsilon_i, \qquad i = 1, \ldots, n,$$

  where f is a function on $[0, 1]$ and ϵ is a possibly correlated vector of (Gaussian) noise.

- We assume f lies in some pre-specified space of functions $\mathcal{F}$, such as a Besov ball.
- Assume for the moment that the noise covariance is known.
- Goal: Make inferences about (often complicated) functionals of f.

Approaches to Function Inference

- Common
  - Estimate plus Goodness of Fit
  - Pointwise confidence bands
  - Confidence intervals on pre-specified features

- Another Idea

A. Generate a confidence set for the *whole* object.

B. Restrict by imposing constraints, if desired.

C. Probe confidence set to address specific questions of interest.

What Do We Want from an Inference?

- Frequentist Confidence Set $\mathcal{C}$

$$\min_f \mathsf{P}\Big\{\mathcal{C} \ni f\Big\} \geq 1 - \alpha. \tag{1}$$

- Bayesian Posterior Region $\mathcal{B}$

$$\mathsf{P}\Big\{f \in \mathcal{B} \mid \text{Data}\Big\} \geq 1 - \alpha. \tag{2}$$

- Can have (2) hold and yet have

$$\min_f \mathsf{P}\Big\{\mathcal{B} \ni f\Big\} \approx 0 \tag{3}$$

  in nonparametric problems.

Road Map: “Balls”

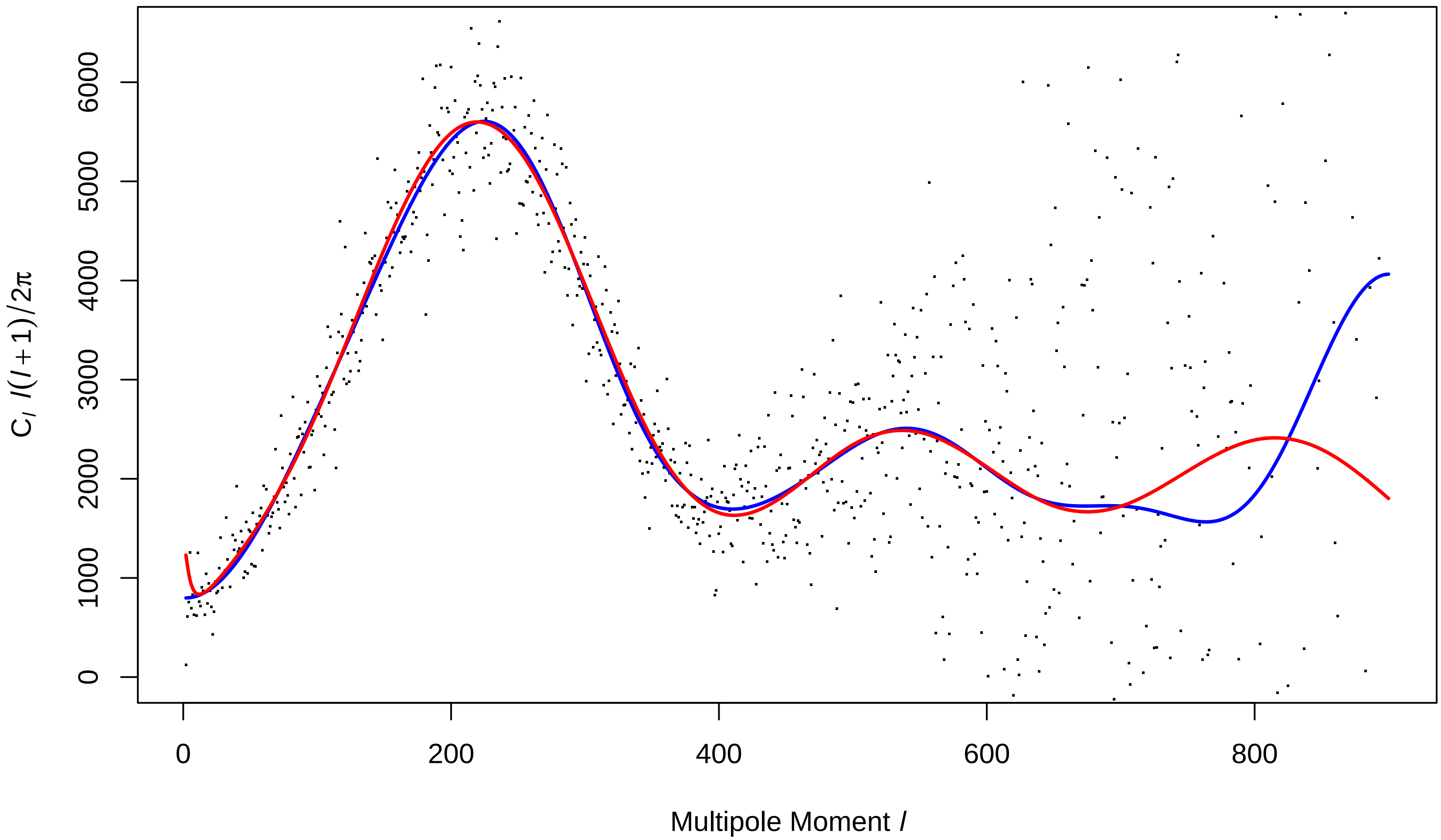
1. Inferences about Functions
   - CMB Spectrum Example
   - The Statistical Problem in General Form
   - Criteria for Effective Inferences
2. Nonparametric Confidence Balls
   - Features and Extensions
3. Keeping Your Eyes on the Ball
   - Parametric Probes
   - Model Checking
   - Confidence Catalogs

Our Approach

- Construct (asymptotic) confidence set for f
  - that is uniform $\sup_{f \in \mathcal{F}} \left| \mathsf{P}\left\{ \mathcal{C}_n \ni f \right\} - (1 - \alpha) \right| \to 0$,
  - that provides post-hoc protection: we can constrain or probe the ball to address any set of questions.

- Construction based on Stein-Beran-Dümbgen pivot method.
- Extended to wavelet bases (GW, 2003b), weighted loss functions (GW, 2003c), and density estimation (GJW, 2003).
- Confidence set takes form of ball (or ellipsoid)

CMB: Center of Ball vs Concordance Model

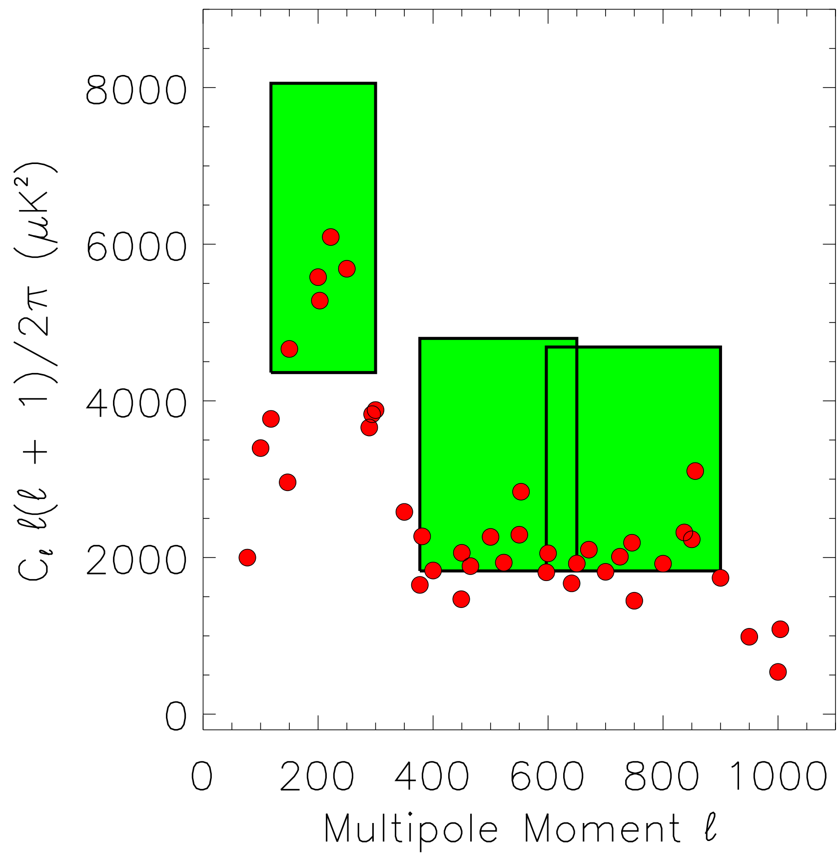


Road Map: "Balls"

1. Inferences about Functions
   - CMB Spectrum Example
   - The Statistical Problem in General Form
   - Criteria for Effective Inferences
2. Nonparametric Confidence Balls
   - Features and Extensions
3. Keeping Your Eyes on the Ball
   - Parametric Probes
   - Model Checking
   - Confidence Catalogs

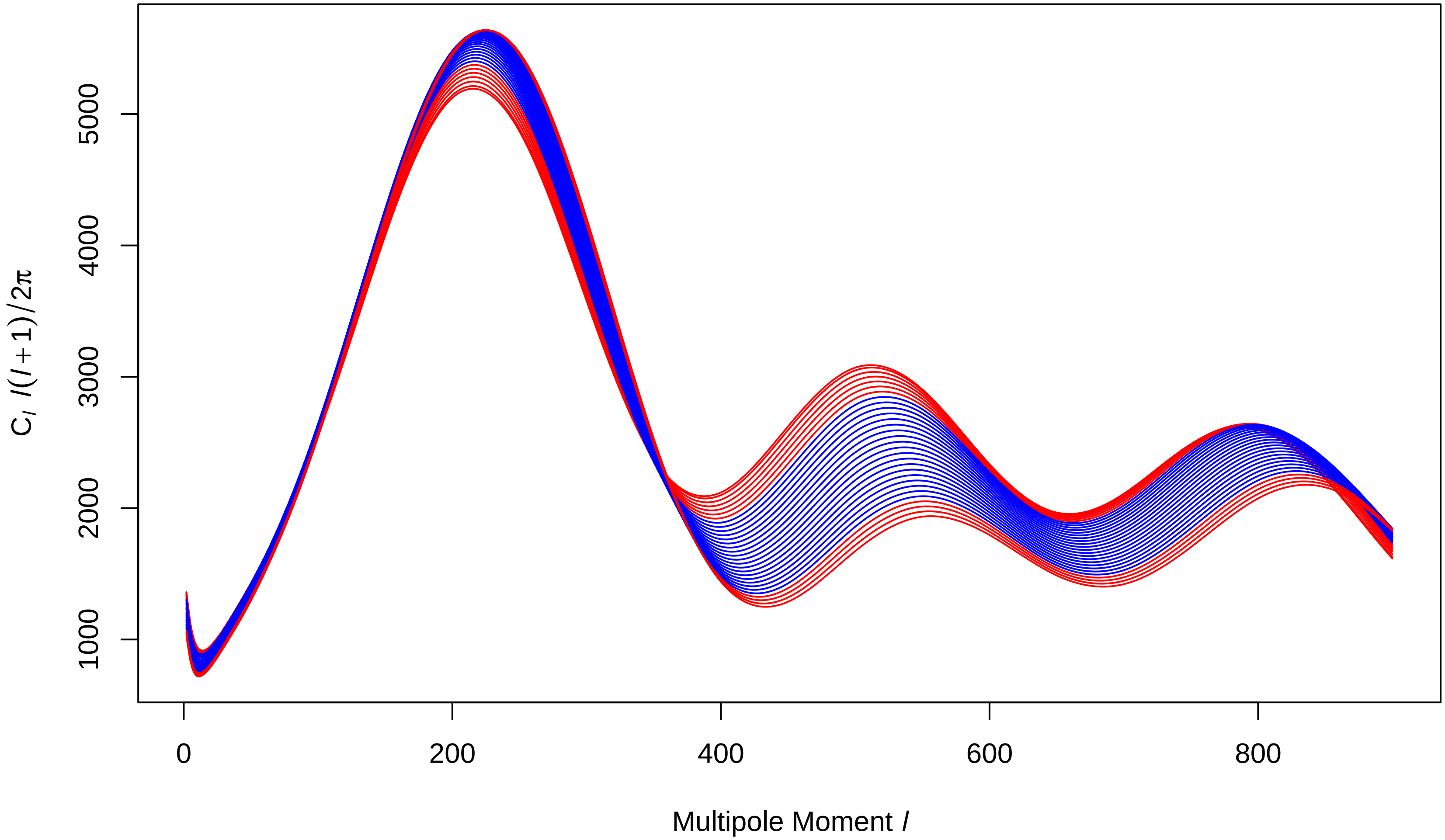
Eyes on the Ball I: Parametric Probes

- Peak Heights
- Peak Locations
- Ratios of Peak Heights



Eyes on the Ball I: Parametric Probes (cont'd)

Varied baryon fraction in CMBFAST keeping $\Omega_{\mathrm{total}} = 1$

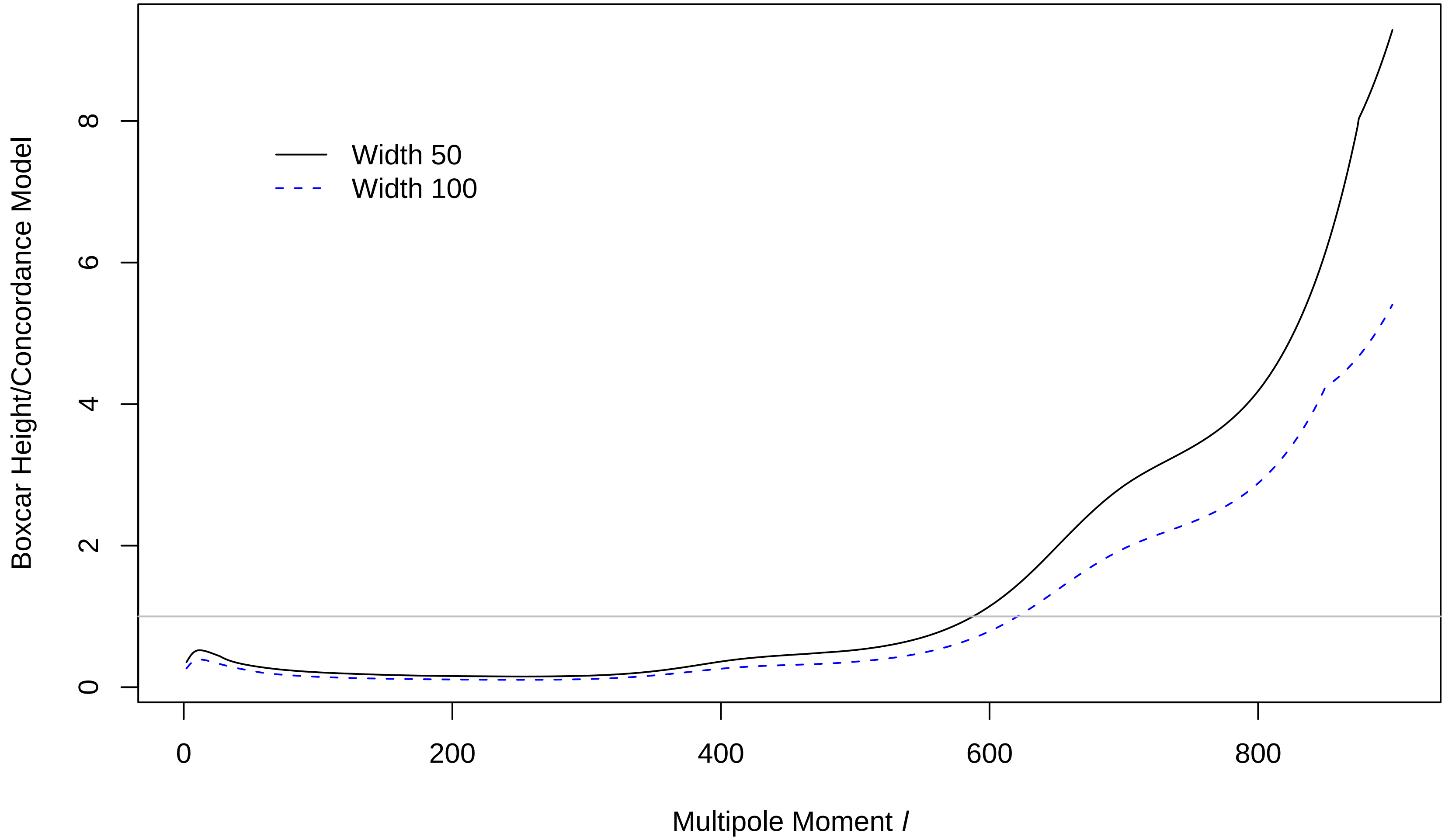


Range [0.034,0.0586] in ball

Eyes on the Ball I: Parametric Probes (cont'd)

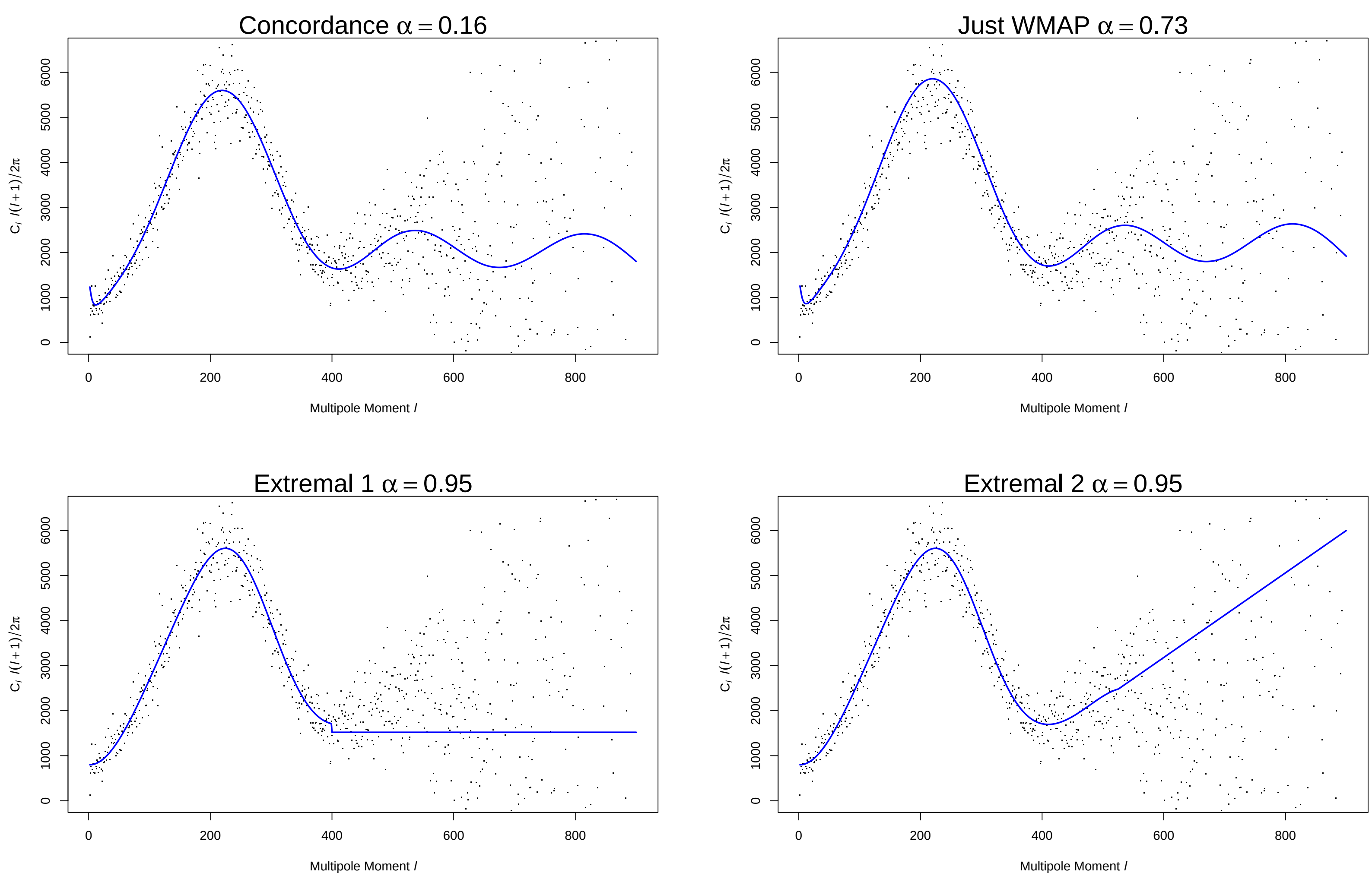
Probe from center with boxcars of given width centered at each ℓ.

Maximum boxcar height in 95% ball, relative to Concordance Model



Eyes on the Ball II: Model Checking

Inclusion in the confidence ball provides simultaneous goodness-of-fit tests for parametric (or other) models.



Eyes on the Ball III: Confidence Catalogs

- Our confidence set construction does not impose constraints based on prior knowledge.

  Instead: form ball first and impose constraints at will.
- Raises the possibility of viewing inferences *as a function* of prior assumptions.

  The confidence ball creates a mapping from prior assumptions to inferences; we call this a confidence catalog.
- Ex: Constraints on peak curvature over range defined by reasonable parametric models.

Take-Home Points: "Balls"

- Uniformity makes the asymptotic approximations more useful.
- Post-hoc protection allows snooping. Can make inferences about any set of functionals with simultaneous validity.
- Nonparametric approach provides check on physical models.

  Embedding parametric model in constrained nonparametric model gives flexibility when model is uncertain.
- Beginning with a confidence set on the whole object makes it easy to compare different sets of assumptions.