Appendix: “Envelopes”

Appendix

The Benjamini-Hochberg Procedure (cont'd)

- Let $\widehat{G}_m$ be the empirical cdf of P^m under the mixture model. Ignoring ties, $\widehat{G}_m(P_{(i)}) = i/m$, so BH equivalent to

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

  as Storey (2002) first noted.

- One can think of this 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 $0 \leq t \leq 1$,

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

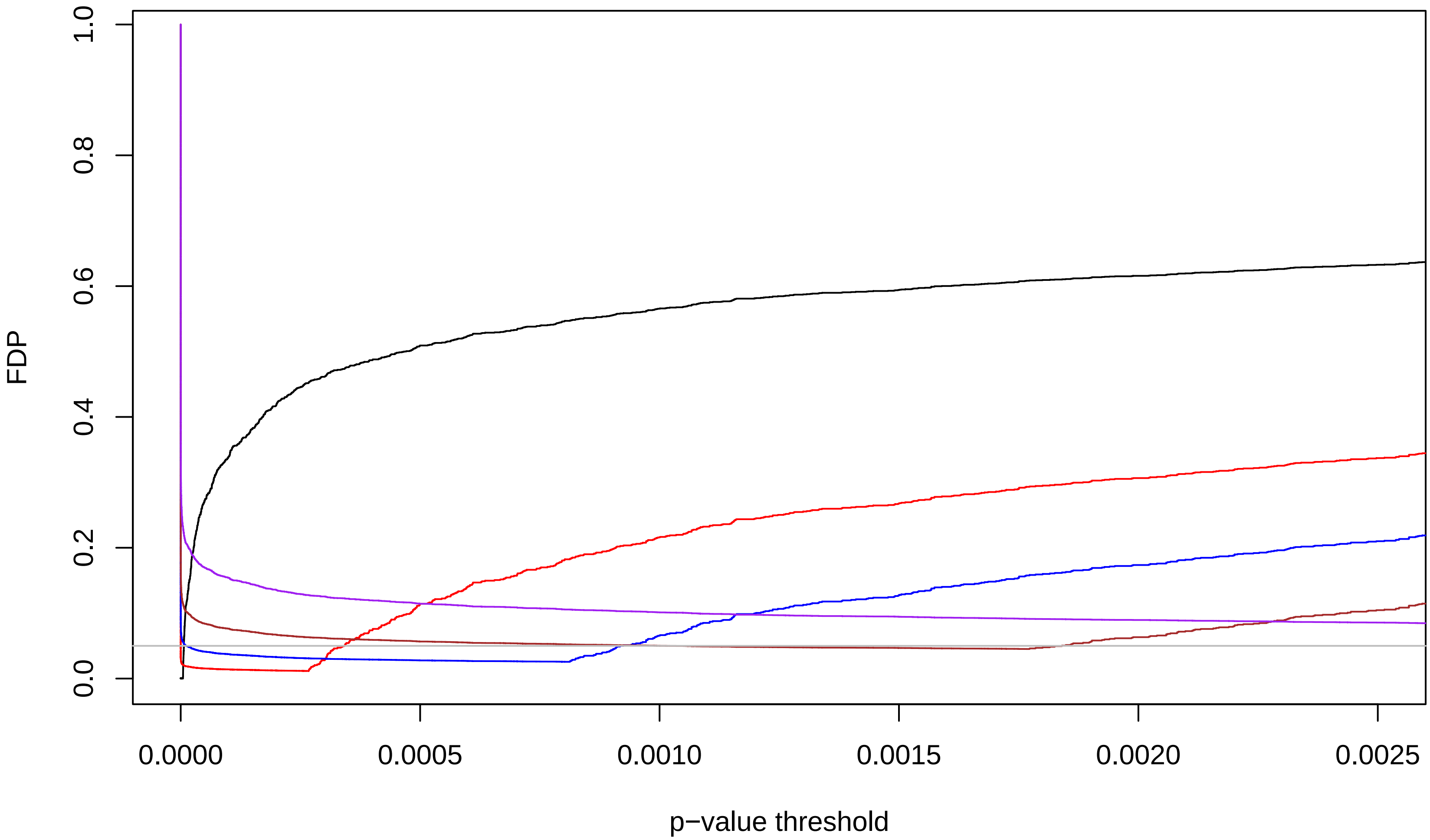
$$\mathsf{E}(\mathsf{FNP}(t)) = a\,\frac{1-F(t)}{1-G(t)} + O\left((a+(1-a)t)^m\right).$$

- Ignoring $O()$ terms and choosing t to minimize $\mathsf{E}(\mathsf{FNP}(t))$ subject to $\mathsf{E}(\mathsf{FDP}(t)) \leq \alpha$, 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).

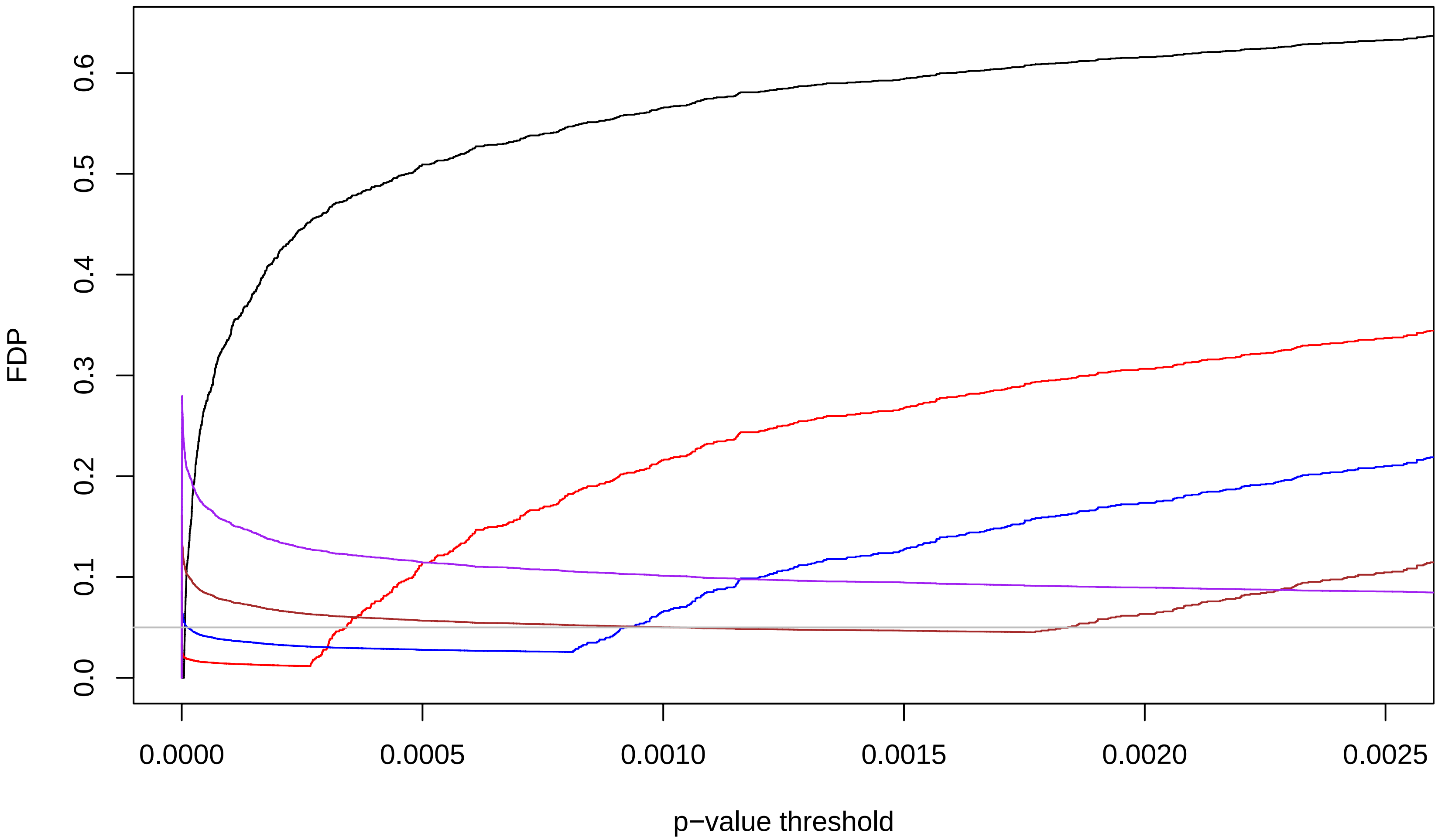
Results: $P_{(k)}$ 90% Confidence Envelopes

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



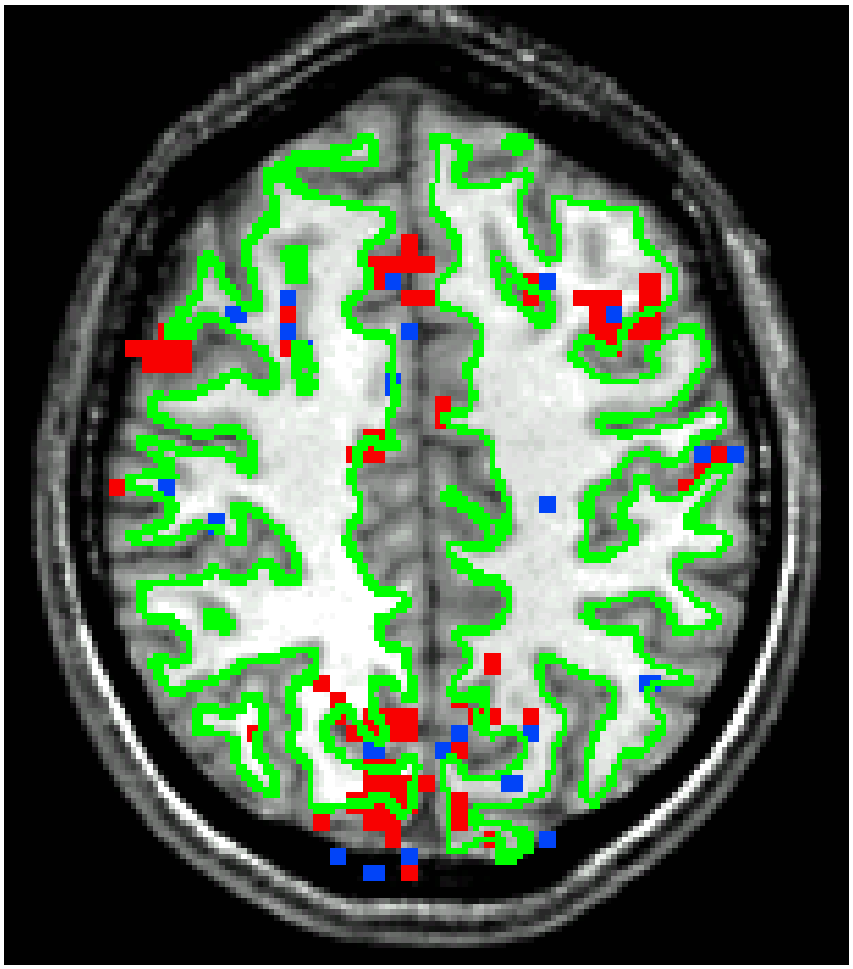
Results: $P_{(k)}$ 90% Modified Envelopes

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



Results: (0.05,0.9) Threshold versus BH

Sample Slice



Computing $P_{(k)}$ Envelopes

- Let q_{mkj} denote the α quantile of the Beta$(k, m-j+1)$ for $k \le j \le m$.

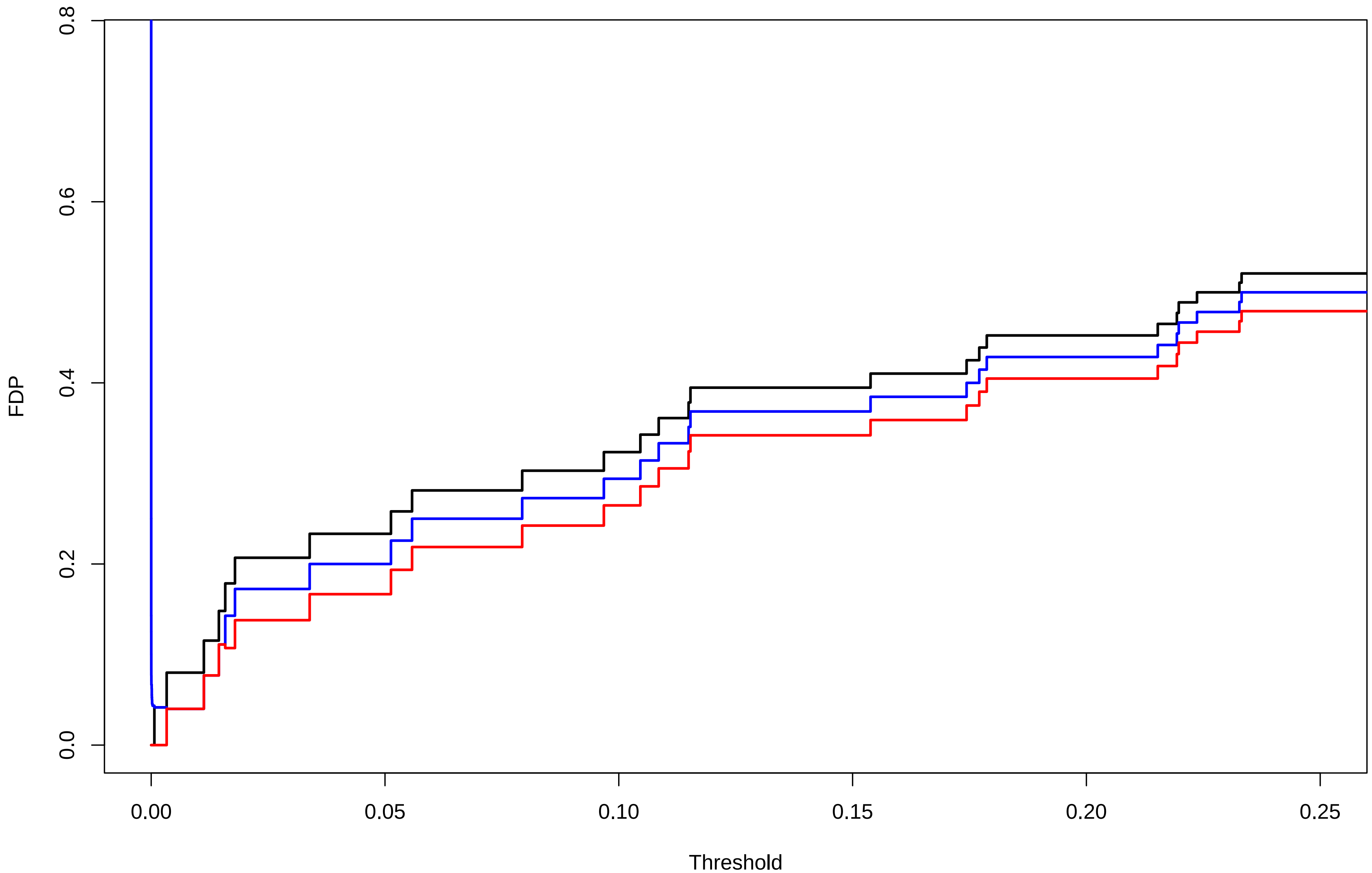
- Let J_k be the index of the smallest $P_{(j)}$ which is $\ge q_{mkj}$.

- The confidence envelope for the $P_{(k)}$-test is achieved by the configuration of nulls (0) and alternatives (1) in the ordered p-values.

$$\underbrace{0 \cdots 0}_{k-1} \overbrace{1 \cdots 1}^{J_k - k} 0 \cdots 0$$

$$\overline{\mathsf{FDP}}_k(t) = \begin{cases} 1 & \text{if } t \le \frac{k-1}{m} \\ \frac{k-1}{m\widehat{G}(t)} & \text{if } \frac{k-1}{m} < t \le \frac{J_k}{m} \\ 1 - \frac{J_k - k + 1}{m\widehat{G}(t)} & \text{if } t > \frac{J_k}{m} \end{cases}$$

Computing $P_{(k)}$ Envelopes (cont'd)



Choice Among $P_{(k)}$ Tests

- For any k, let $V_k = J_k - k$.
- In any pairwise comparison of $P_{(k)}$ and $P_{(k')}$ tests with $k < k'$, there are only three possible orderings:

  A. $P_{(k)}$ dominates everwhere if $V_k \geq V_{k'}$,

  B. $P_{(k')}$ dominates everywhere if $V_{k'} > V_k \left[1 + \frac{k' - k}{k - 1}\right] + \frac{k' - k}{k - 1}$,

  C. Otherwise, the two profiles cross at $J_{k'}$ with value $(k' - 1)/J_{k'}$.

- The result for any k can be put in terms of Uniform hitting times for a boundary of the form $G(q_{mkj}) \approx G(\tilde{q}_{mk}/(m - j + 1))$.

  The distribution of these hitting times can be computed exactly (with difficulty) via Steck's equality.

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
$$\mathsf{FDP}(t) = \frac{\lambda(S_0 \cap \{s \in S : X(s) \geq t\})}{\lambda(\{s \in S : X(s) \geq t\})},$$
where λ is usually Lebesgue measure.

- As 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.

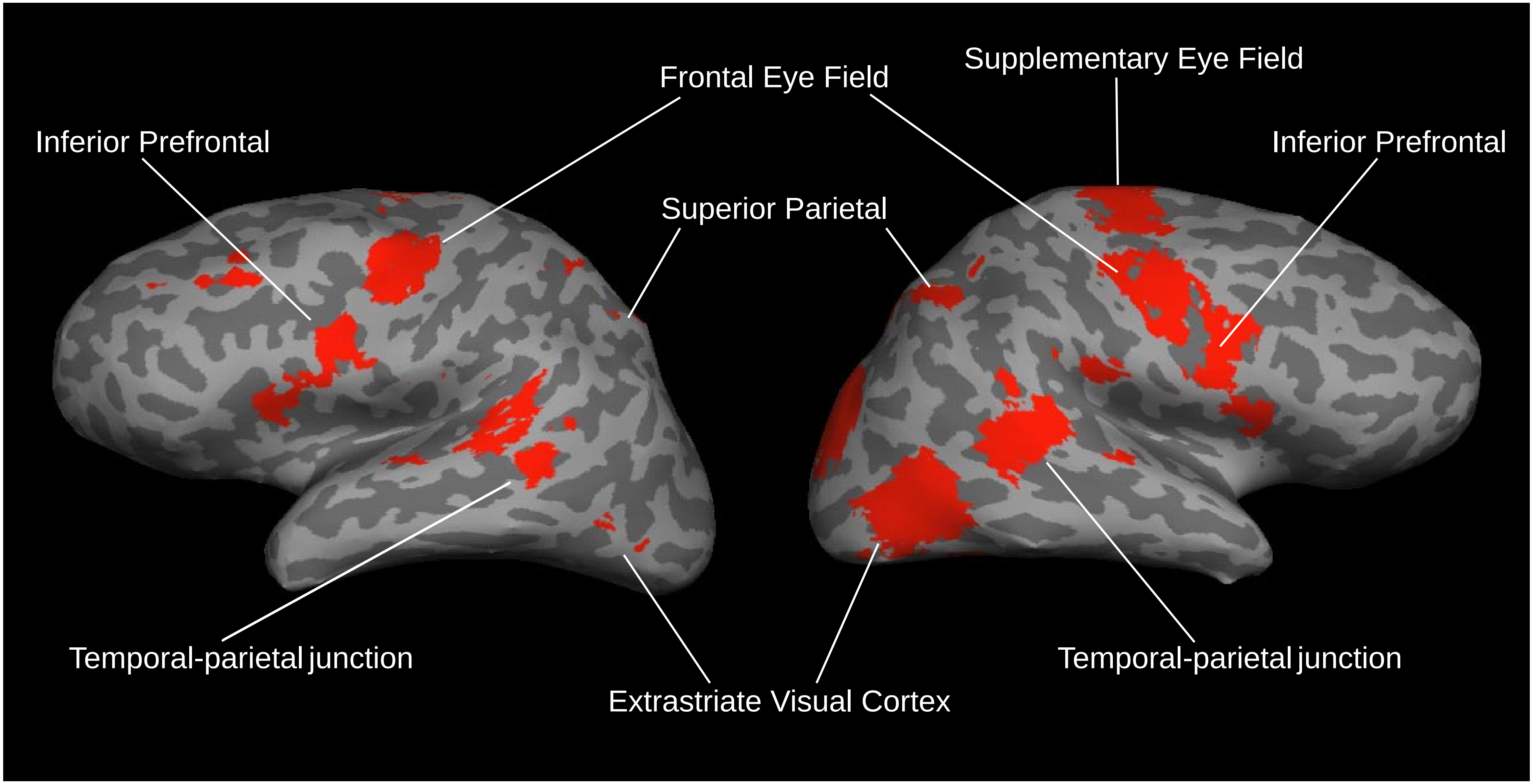
- For our purposes, approximation based on the expected Euler characteristic of the field's level sets will not work because the Euler characteristic is non-monotone for non-convex sets.

  (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.

Algorithm for Confidence Superset

1. Compute all realized values of the test statistics $x(S_j)$
2. Sort these in decreasing order $x_{(1)} \geq \cdots \geq x_{(N)}$.

   Let $S_{(j)}$ be the partition element corresponding to $x_{(j)}$.
3. For $k = 1, \ldots, N$ do the following:

   a. Set $V_k = \bigcup_{j=k}^{N} S_{(j)}$.

   b. Compute $p(x_{(k)}, V_k)$.

   c. If $p(x_{(k)}, V_k) \geq \alpha$: STOP and set $V^* = V_k$.

   d. If $p(x_{(k)}, V_k) < \alpha$: increase k by 1 and GOTO 3a.

Gaussian Fields

- Assume $S = [0,1]^d$ and that X is a zero-mean, homogeneous Gaussian field with covariance

$$\mathrm{Cov}(X(r), X(s)) = \sigma^2 \rho(r - s),$$

  that gives X almost surely continuous sample paths.

  Example: $\rho(u) = 1 - u^T C^{-2} u + o(\|u\|^2)$ for some matrix C.

- The key challenge here is to approximate the p-values $p(z, A)$.

  One approximation is based on the expected Euler characteristic of the field's level sets.

Gaussian Fields (cont'd)

- For our purposes, this will not work because the Euler characteristic approximation is non-monotone for non-convex sets.

  Note also that for non-convex sets, not all terms in the Euler approximation are accurate.

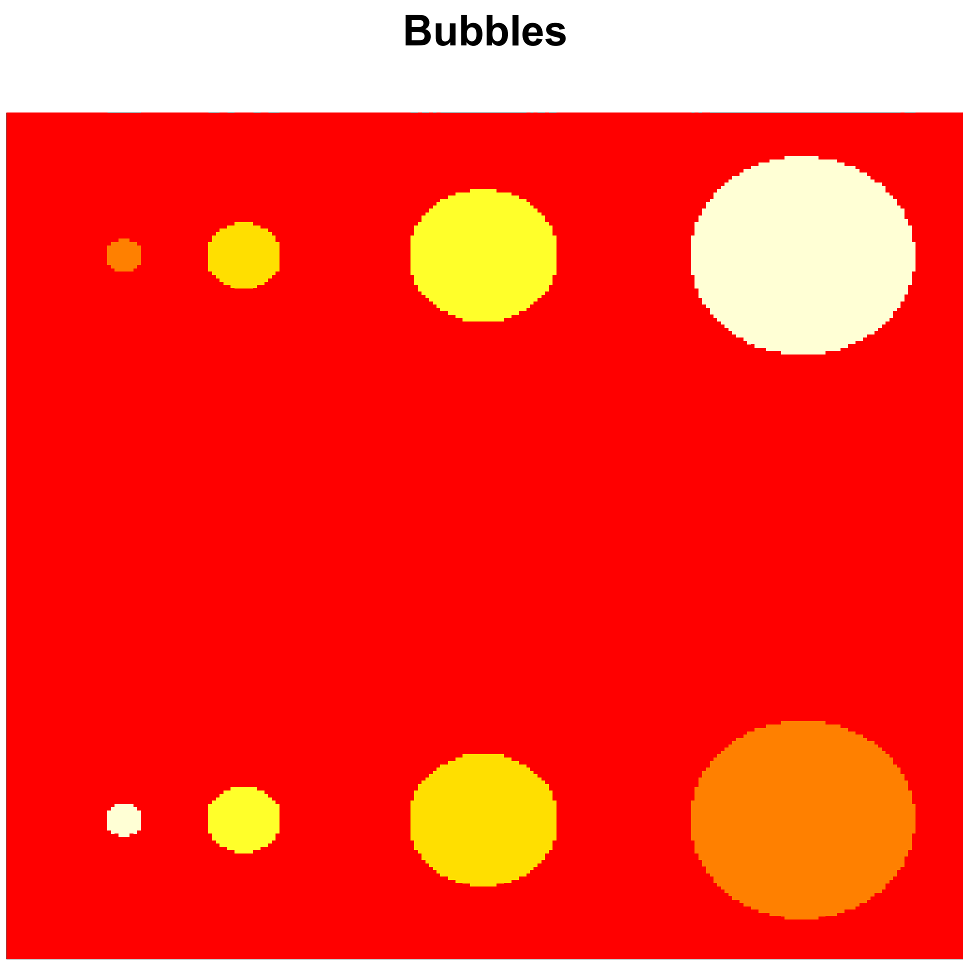
- Instead we use a result of Piterbarg (1996) to obtain

$$p(z, A) = \mathsf{P}\left\{ \sup_{s \in A} \frac{X(s)}{\sigma} \geq \frac{z}{\sigma} \right\} \simeq \frac{\pi^{-\frac{d}{2}}}{|\det C|} \lambda(A) \left(\frac{z}{\sigma}\right)^d \left[1 - \Phi\left(\frac{z}{\sigma}\right)\right],$$

  for C as in the quadratic form above.

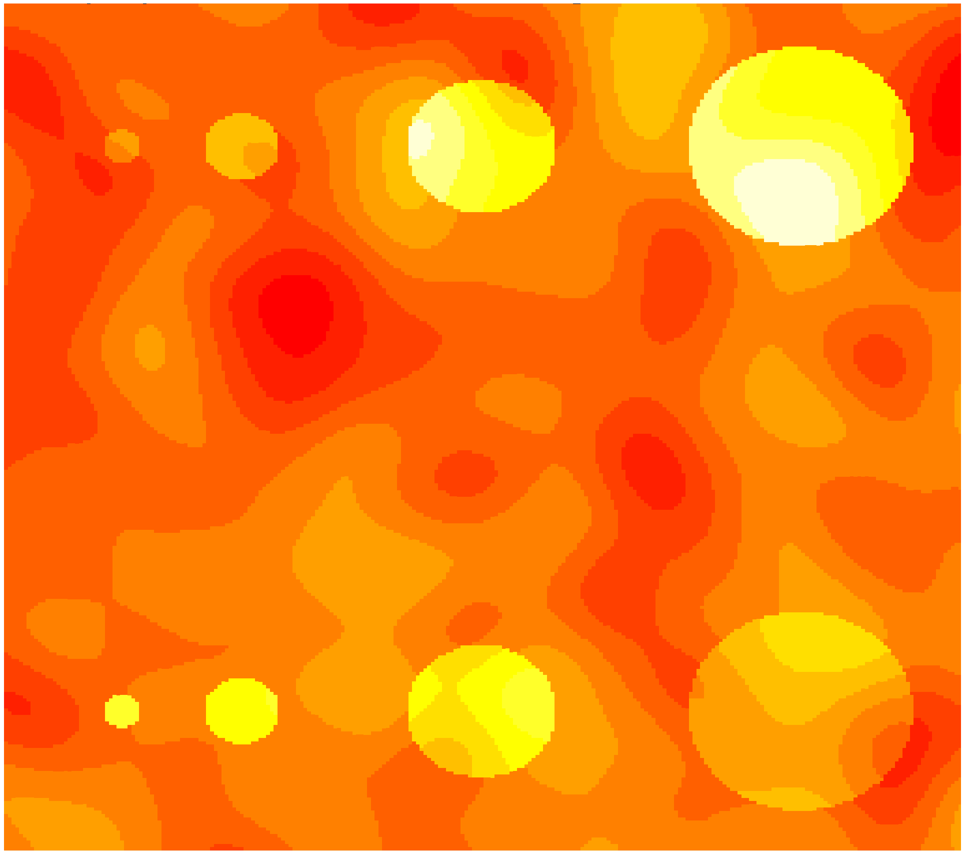
- Simulations over a wide variety of S_0s and covariance structures show that coverage of U rapidly converges to the target level.

Gaussian Fields: Example



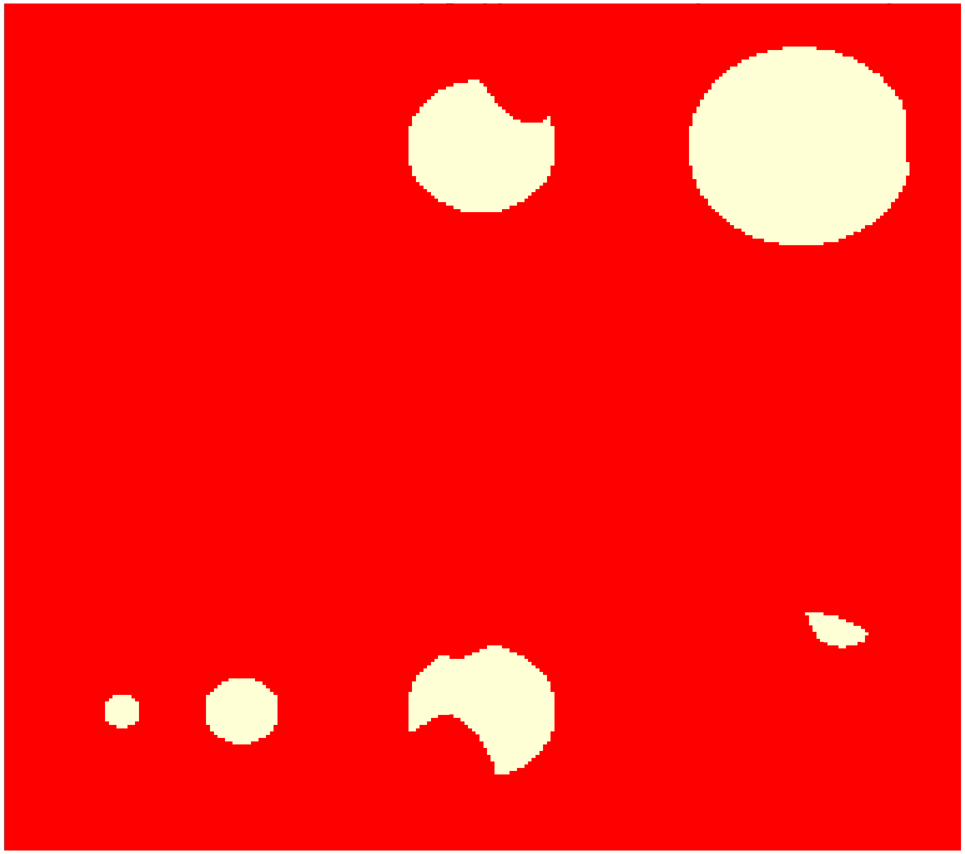
Gaussian Fields: Example (cont'd)

Bubbles + noise

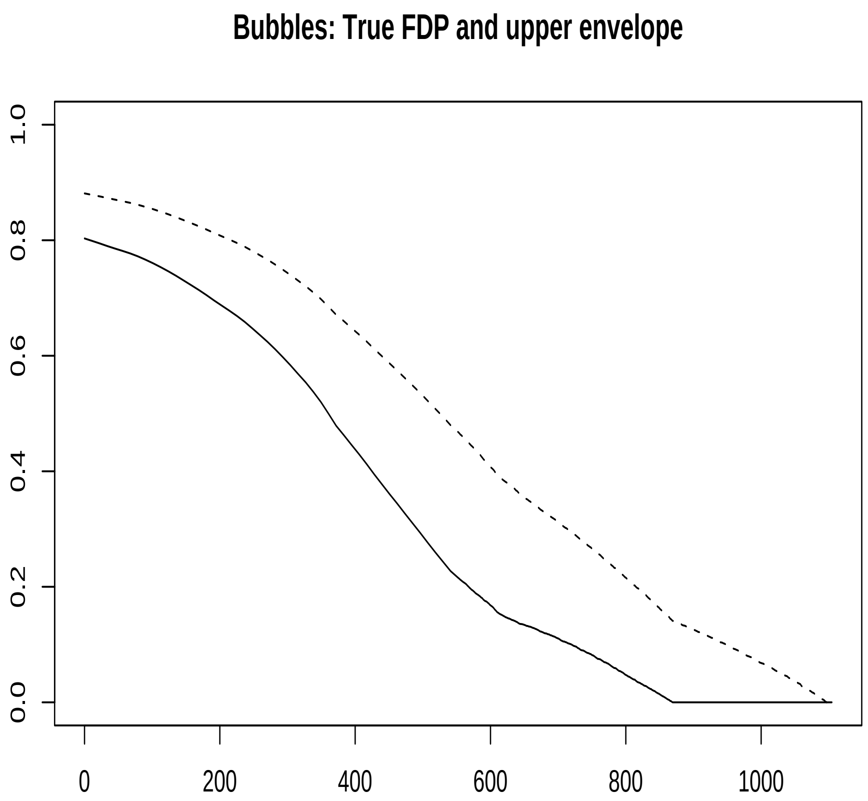


Gaussian Fields: Example (cont'd)

Bubbles: confidence bound



Gaussian Fields: Example (cont'd)



Appendix: “Balls”

Stein-Beran-Dümbgen Pivot Method

- Convert function space problem $Y = f + \epsilon$ into sequence space problem.

  Let $\phi_1, \phi_2, \ldots$ be an orthonormal basis for $[0, 1]$ and let $\mu_j = \int f \phi_j$.

  Define

$$Z_j = \frac{1}{n} \sum_{i=1}^{n} Y_i \phi_j(x_i) \;\approx\; \mu_j + \frac{1}{\sqrt{n}} \epsilon_j.$$

- Estimate μ by $\widehat{\mu}(\lambda)$ for some possibly vector-valued tuning parameter.
- Let $L_n(\lambda)$ be the (unobserved) loss as a function of λ. For example, $L_n(\lambda) = \sum_j (\widehat{\mu}_j(\lambda) - \mu_j)^2$.
- Let $S_n(\lambda)$ be an (asymptotically) unbiased estimate of risk.

Pivot Method (cont'd)

1. Show that the *pivot process* $B_n(\lambda) = \sqrt{n}(L_n(\lambda) - S_n(\lambda))$ has a Gaussian limit process.
2. For $\widehat{\lambda}_n$ minimizing $S_n(\lambda)$, show $B_n(\widehat{\lambda}_n)$ has a Gaussian limit.
3. Find a consistent estimator $\widehat{\tau}_n^2$ for variance of latter limit.
4. Conclude that

$$\mathcal{D}_n = \left\{ \mu \colon \frac{L_n(\widehat{\lambda}_n) - S_n(\widehat{\lambda}_n)}{\widehat{\tau}_n/\sqrt{n}} \leq z_\alpha \right\}$$
$$= \left\{ \mu \colon \sum_{\ell=1}^{n} (\widehat{\mu}_\ell(\widehat{\lambda}_n) - \mu_\ell)^2 \leq \frac{\widehat{\tau}_n\, z_\alpha}{\sqrt{n}} + S_n(\widehat{\lambda}_n) \right\}$$

is an asymptotic $1 - \alpha$ confidence set for μ.

Pivot Method (cont'd)

5. It follows that

$$\mathcal{A}_n = \left\{ \sum_{\ell=1}^{n} \mu_\ell \phi_\ell(\cdot) : \ \mu \in \mathcal{D}_n \right\}$$

   is an asymptotic $1 - \alpha$ confidence set for $f_n = \sum_{\ell=1}^{n} \mu_\ell \phi_\ell$.

6. With appropriate function-space assumptions, can dilate $\mathcal{A}_n$ to a set $\mathcal{C}_n$ that is a uniform confidence set for f.

Pivot Method: Extension

- extend to invariant loss...