Advances in thresholding and multiple comparisons

Thomas Nichols, Ph.D.
Department of Statistics &
Warwick Manufacturing Group
University of Warwick

OHBM Advanced fMRI Course
8 June, 2014
Overview

• Why threshold?
• Assessing statistic images
• Measuring false positives
• Practical solutions
Thresholding

Where’s the signal?

High Threshold
\( t > 5.5 \)
- Good Specificity
- Poor Power
  (risk of false negatives)

Med. Threshold
\( t > 3.5 \)

Low Threshold
\( t > 0.5 \)
- Poor Specificity
  (risk of false positives)
- Good Power

...but why threshold?!
Don’t threshold, model the signal!

- Signal location?
  - Estimates and CI’s on (x,y,z) location
- Signal magnitude?
  - CI’s on % change
- Spatial extent?
  - Estimates and CI’s on activation volume
  - Robust to choice of cluster definition

...but this requires an explicit spatial model
Blue-sky inference: What we need

• Explicit spatial models
  – No routine methods exist
    • High-dimensional mixture modeling problem
    • Activations don’t look like Gaussian blobs

• Some encouraging initial efforts…

Gershman et al. (2011). *NI*, 57(1), 89-100.

• **ADVT**: Mon, 8:00, Hall 1,
  Session “Methodological Advances in Lesion Symptom Mapping”
  Talk “Spatial Bayesian Modelling of Binary Lesion data”, T. Nichols
Real-life inference: What we get (typically)

- Signal **location**
  - Local maximum — *no inference*

- Signal **magnitude**
  - Local maximum intensity — P-values (& CI’s)

- Spatial **extent**
  - Cluster volume — P-value, no CI’s
    - Sensitive to blob-defining-threshold
Assessing Statistic Images...
Ways of assessing statistic images

• Standard methods
  – Voxel
  – Cluster
  – Set
  – Peak (new)
Voxel-level Inference

- Retain voxels above $\alpha$-level threshold $u_\alpha$
- Gives best spatial specificity
  - The null hyp. at a single voxel can be rejected
Cluster-level Inference

- Two step-process
  - Define clusters by arbitrary threshold $u_{\text{clus}}$
  - Retain clusters larger than $\alpha$-level threshold $k_\alpha$

![Diagram showing cluster threshold and significance](image)
Cluster-level Inference

• Typically better sensitivity
• Worse spatial specificity
  – The null hyp. of entire cluster is rejected
  – Only means that one or more of voxels in cluster active
Set-level Inference

- Count number of blobs $c$
  - Minimum blob size $k$
- Worst spatial specificity
  - Only can reject global null hypothesis

Here $c = 1$; only 1 cluster larger than $k$
Peak-level Inference

• Identify all the local maxima
  – Ignore all smaller than $u_{peak}$
• Retain peaks by height

$u_\alpha$ $u_{peak}$

Significant peak

Not a significant peak
Peak-level Inference

• “Topological inference” – interpretable with boundless Point Spread Function (see Chumbley & Friston, NI, 2009)

• Cumbersome – only making inference at a sprinkling of locations

\[ u_{\alpha}, u_{\text{peak}} \]

Significant peak

Not a significant peak
Test Statistics for Assessing Statistic Images...
Sometimes, Different Possible Ways to Test...

<table>
<thead>
<tr>
<th>Image Feature</th>
<th>Test Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voxel</td>
<td>1. Statistic image value</td>
</tr>
<tr>
<td>Cluster</td>
<td>1. Cluster size in voxels</td>
</tr>
<tr>
<td></td>
<td>2. Cluster size in RESELs</td>
</tr>
<tr>
<td></td>
<td>3. Combination, Joint Peak-Cluster</td>
</tr>
<tr>
<td></td>
<td>4. Combination, Cluster Mass</td>
</tr>
<tr>
<td></td>
<td>5. Combination, Threshold-Free Cluster Enhancement</td>
</tr>
<tr>
<td>Set</td>
<td>1. Cluster count</td>
</tr>
<tr>
<td>Peak</td>
<td>1. Statistic image value</td>
</tr>
</tbody>
</table>
Sometimes, Different Possible Ways to Test...

<table>
<thead>
<tr>
<th>Image Feature</th>
<th>Test Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voxel</td>
<td>1. Statistic image value</td>
</tr>
</tbody>
</table>
| Cluster       | 1. Cluster size in voxels  
                | 2. Cluster size in RESELs  
                | 3. Combination, Joint Peak-Cluster  
                | 4. Combination, Cluster Mass  
                | 5. Combination, Threshold-Free Cluster Enhancement |
| Set           | 1. Cluster count |
| Peak          | 1. Statistic image value |
Combining Cluster Size with Intensity Information

- **Peak-Height combining** Poline *et al.*, NeuroImage 1997
  - Minimum $P_{\text{extent}}$ & $P_{\text{height}}$
    - Take better of two P-values; (use RFT to correct for taking minimum)
  - Can catch small, intense clusters

- **Cluster mass** Bullmore *et al.*, IEEE Trans Med Img 1999
  - Integral $M$ above threshold
    - More powerfully combines peak & height (Hayasaka & Nichols, NI 2004)

- Both are still cluster inference methods!
The Pesky Cluster Forming Threshold $u_c$

- Cluster inference is highly sensitive to cluster-forming threshold $u_c$
  - Set too low, one big blob
  - Set too high, miss all the signal
Threshold-Free Cluster Enhancement (TFCE)

- A cluster-informed voxel-wise statistic
- Consider cluster mass voxel-wise, for every $u_c$!
  - For a given voxel, sum up all clusters ‘below’
    - For all possible $u_c$, add up all clusters that contain that voxel
  - But this would give low $u_c$’s too much weight
    - Low $u_c$’s give big clusters just by chance

Smith & Nichols, NI 2009
Threshold-Free Cluster Enhancement (TFCE)

- A cluster-informed voxel-wise statistic
- Consider cluster mass voxel-wise, for every $u_c$!
  - For a given voxel, sum up all clusters ‘below’
    - For all possible $u_c$, add up all clusters that contain that voxel
  - But this would give low $u_c$’s too much weight
    - Low $u_c$’s give big clusters just by chance
  - Solution: Down-weight according to $u_c$!
Threshold-Free Cluster Enhancement (TFCE)

- TFCE Statistic for voxel \( v \)

\[
TFCE(v) = \int_0^{t(v)} h^H e(h)^E dh \approx \sum_{0, \delta, 2\delta, \ldots, t(v)} h^H e(h)^E \delta
\]

- Parameters H & E control balance between cluster & height information
  - H=2 & E=1/2 as motivated by theory
TFCE Redux

- Avoids choice of cluster-forming threshold $u_c$
- Generally more sensitive than cluster-wise
- But yet less specific
  - Inference is on some cluster for some $u_c$
  - “Support” of effect could extend far from significant voxels
- Implementation
  - Currently only FSL’s randomise
Multiple comparisons...
Multiple Comparisons Problem

• Which of 100,000 voxels are sig.?
  – $\alpha=0.05 \Rightarrow 5,000$ false positive voxels

• Which of (random number, say) 100 clusters significant?
  – $\alpha=0.05 \Rightarrow 5$ false positives clusters
MCP Solutions: Measuring False Positives

- **Familywise Error Rate (FWER)**
  - Familywise Error
    - Existence of one or more false positives
  - FWER is probability of familywise error

- **False Discovery Rate (FDR)**
  - FDR = E(V/R)
  - R voxels declared active, V falsely so
    - Realized false discovery rate: V/R
Random field theory...
FWER MCP Solutions: Random Field Theory

- Euler Characteristic $\chi_u$
  - Topological Measure
    - #blobs - #holes
  - At high thresholds, just counts blobs
  - $\text{FWER} = \mathbb{P}(\text{Max voxel} \geq u \mid H_0)$
    - $= \mathbb{P}(\text{One or more blobs} \mid H_0)$
    - $\approx \mathbb{P}(\chi_u \geq 1 \mid H_0)$
    - $\approx \mathbb{E}(\chi_u \mid H_0)$
Random Field Theory
Smoothness Parameterization

- $E(\chi_{\mu})$ depends on $|\Lambda|^{1/2}$
  - $\Lambda$ roughness matrix:

$$
\Lambda = \text{Var} \left( \frac{\partial G}{\partial (x, y, z)} \right) = 
\begin{pmatrix}
\text{Var} \left( \frac{\partial G}{\partial x} \right) & \text{Cov} \left( \frac{\partial G}{\partial x}, \frac{\partial G}{\partial y} \right) & \text{Cov} \left( \frac{\partial G}{\partial x}, \frac{\partial G}{\partial z} \right) \\
\text{Cov} \left( \frac{\partial G}{\partial y}, \frac{\partial G}{\partial x} \right) & \text{Var} \left( \frac{\partial G}{\partial y} \right) & \text{Cov} \left( \frac{\partial G}{\partial y}, \frac{\partial G}{\partial z} \right) \\
\text{Cov} \left( \frac{\partial G}{\partial z}, \frac{\partial G}{\partial x} \right) & \text{Cov} \left( \frac{\partial G}{\partial z}, \frac{\partial G}{\partial y} \right) & \text{Var} \left( \frac{\partial G}{\partial z} \right)
\end{pmatrix} =
\begin{pmatrix}
\lambda_{xx} & \lambda_{xy} & \lambda_{xz} \\
\lambda_{yx} & \lambda_{yy} & \lambda_{yz} \\
\lambda_{zx} & \lambda_{zy} & \lambda_{zz}
\end{pmatrix}
$$

- Smoothness parameterized as Full Width at Half Maximum
  - FWHM of Gaussian kernel needed to smooth a white noise random field to roughness $\Lambda$

$$
|\Lambda|^{1/2} = \frac{(4 \log 2)^{3/2}}{\text{FWHM}_x \text{FWHM}_y \text{FWHM}_z}.
$$
Random Field Theory
Smoothness Parameterization

- **RESELS**
  - Resolution Elements
  - 1 RESEL = FWHM$_x$ $\times$ FWHM$_y$ $\times$ FWHM$_z$
  - RESEL Count $R$
    - $R = \lambda(\Omega) \sqrt{|\Lambda|} = (4 \log 2)^{3/2} \lambda(\Omega) / (\text{FWHM}_x \times \text{FWHM}_y \times \text{FWHM}_z)$
    - Volume of search region in units of smoothness
    - Eg: 10 voxels, 2.5 FWHM 4 RESELS

- **Beware RESEL misinterpretation**
  - RESEL are not “number of independent ‘things’ in the image”
Random Field Theory
Smoothness Estimation

- Smoothness est’d from standardized residuals
  - Variance of gradients
  - Yields resels per voxel (RPV)

- RPV image
  - Local roughness est.
  - Can transform in to local smoothness est.
    - FWHM Img = (RPV Img)^{-1/D}
    - Dimension $D$, e.g. $D=2$ or 3

- Est. smoothness also needed for AlphaSim

```
spm_imcalc_ui('RPV.img', ... 'FWHM.img','i1.^(-1/3)')
```
Random Field Theory
Limitations

• Sufficient smoothness
  – FWHM smoothness $3-4 \times$ voxel size ($Z$)
  – More like $\sim 10 \times$ for low-df $T$ images

• Smoothness estimation
  – Estimate is biased when images not sufficiently smooth

• Multivariate normality
  – Virtually impossible to check

• Several layers of approximations

• Stationary required for cluster size results
Real Data

• fMRI Study of Working Memory
  – Item Recognition
    • Active: View five letters, 2s pause, view probe letter, respond
    • Baseline: View XXXXX, 2s pause, view Y or N, respond

• Second Level RFX
  – Difference image, A-B constructed for each subject
  – One sample t test
Real Data: RFT Result

- Threshold
  - $S = 110,776$
  - $2 \times 2 \times 2$ voxels
    - $5.1 \times 5.8 \times 6.9$ mm FWHM
  - $u = 9.870$

- Result
  - 5 voxels above the threshold
  - 0.0063 minimum FWE-corrected $p$-value
Permutation...
Nonparametric Permutation Test

- **Parametric methods**
  - Assume distribution of statistic under null hypothesis

- **Nonparametric methods**
  - Use *data* to find distribution of statistic under null hypothesis
  - Any statistic!
Permutation Test & Exchangeability

- Exchangeability is fundamental
  - Def: Distribution of the data unperturbed by permutation
  - Under H0, exchangeability justifies permuting data
  - Allows us to build permutation distribution
- fMRI scans not exchangeable over time!
  - Even if no signal, autocorrelation structures data
- Subjects are exchangeable
  - Under Ho, each subject’s “active” “control” labels can be flipped
  - Equivalently, under Ho flip the sign of each subject’s contrast images
Controlling FWE: Permutation Test

• Parametric methods
  – Assume distribution of $\text{max}$ statistic under null hypothesis

• Nonparametric methods
  – Use data to find distribution of $\text{max}$ statistic under null hypothesis
  – Again, any max statistic!
Permutation Test
Smoothed Variance $t$

- Collect max distribution
  - To find threshold that controls FWER
- Consider smoothed variance $t$ statistic
Permutation Test
Smoothed Variance $t$

- Collect max distribution
  - To find threshold that controls FWER
- Consider smoothed variance $t$ statistic
Permutation Test Example

- Permute!
  - \(2^{12} = 4,096\) ways to flip 12 A/B labels
  - For each, note maximum of \(t\) image

![Permutation Distribution](image1.png)
![Orthogonal Slice Overlay Thresholded \(t\)](image2.png)
**Permutation**

$u_{\text{Perm}} = 7.67$

58 sig. vox.

$t_{11}$ Statistic, Nonparametric Threshold

---

**RFT & Bonferroni**

$u_{\text{RF}} = 9.87$

$u_{\text{Bonf}} = 9.80$

5 sig. vox.

$5.1 \times 5.8 \times 6.9$ mm FWHM noise smoothness

$t_{11}$ Statistic, RF & Bonf. Threshold

---

**Permutation & Sm.Var.**

378 sig. vox.

Test Level vs. $t_{11}$ Threshold

---

Smoothed Variance $t$ Statistic, Nonparametric Threshold
Reliability with Small Groups

• Consider n=50 group study
  – Event-related Odd-Ball paradigm, Kiehl, et al.

• Analyze all 50
  – Analyze with SPM and SnPM, find FWE thresh.

• Randomly partition into 5 groups 10
  – Analyze each with SPM & SnPM, find FWE thresh

• Compare reliability of small groups with full
  – With and without variance smoothing
SPM $t_{11}$: 5 groups of 10 vs all 50
5% FWE Threshold

T > 10.93
SPM\{$T_g$\}
10 subj

T > 11.04
SPM\{$T_g$\}
10 subj

T > 11.01
SPM\{$T_g$\}
10 subj

T > 10.69
SPM\{$T_g$\}
10 subj

T > 10.10
SPM\{$T_g$\}
10 subj

T > 4.66
SPM\{$T_{49}$\}
all 50
SnPM $t$: 5 groups of 10 vs. all 50
5% FWE Threshold

Arbitrary thresh of 9.0
SnPM SmVar $t$: 5 groups of 10 vs. all 50
5% FWE Threshold

T > 4.69

T > 5.04

T > 4.57

T > 4.84

T > 4.64

Arbitrary thresh of 9.0
False Discovery Rate...
MCP Solutions: Measuring False Positives

• Familywise Error Rate (FWER)
  – Familywise Error
    • Existence of one or more false positives
  – FWER is probability of familywise error

• False Discovery Rate (FDR)
  – FDR = E(V/R)
  – R voxels declared active, V falsely so
    • Realized false discovery rate: V/R
False Discovery Rate Illustration:

Noise

Signal

Signal+Noise
Control of Per Comparison Rate at 10%

<table>
<thead>
<tr>
<th></th>
<th>6.7%</th>
<th>10.4%</th>
<th>14.9%</th>
<th>9.3%</th>
<th>16.2%</th>
<th>13.8%</th>
<th>14.0%</th>
<th>10.5%</th>
<th>12.2%</th>
<th>8.7%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of Null Pixels that are False Positives</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Control of Familywise Error Rate at 10%

<table>
<thead>
<tr>
<th></th>
<th>11.3%</th>
<th>11.3%</th>
<th>12.5%</th>
<th>10.8%</th>
<th>11.5%</th>
<th>10.0%</th>
<th>10.7%</th>
<th>11.2%</th>
<th>10.2%</th>
<th>9.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occurrence of Familywise Error</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Control of False Discovery Rate at 10%

<table>
<thead>
<tr>
<th></th>
<th>6.7%</th>
<th>10.4%</th>
<th>14.9%</th>
<th>9.3%</th>
<th>16.2%</th>
<th>13.8%</th>
<th>14.0%</th>
<th>10.5%</th>
<th>12.2%</th>
<th>8.7%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of Activated Pixels that are False Positives</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Benjamini & Hochberg Procedure

- Select desired limit \( q \) on FDR
- Order p-values, \( p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(V)} \)
- Let \( r \) be largest \( i \) such that

\[
p_{(i)} \leq \frac{i}{V} \times q
\]

- Reject all hypotheses corresponding to \( p_{(1)}, \ldots, p_{(r)} \).
- Threshold is adaptive to signal in the data

\[ \text{JRSS-B (1995)} \]
\[ 57:289-300 \]
Real Data: FDR Example

• Threshold
  - Indep/PosDep
    \( u = 3.83 \)
  - Arb Cov
    \( u = 13.15 \)

• Result
  - 3,073 voxels above Indep/PosDep \( u \)
  - <0.0001 minimum FDR-corrected p-value

FDR Threshold = 3.83
3,073 voxels
FWER Perm. Thresh. = 9.87
7 voxels
### Changes in SPM Inference

#### Before SPM8

<table>
<thead>
<tr>
<th></th>
<th>&lt; SPM8</th>
<th>Uncorrected</th>
<th>FDR</th>
<th>FWE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voxel-wise</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Cluster-wise</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
</tbody>
</table>

#### SPM8

<table>
<thead>
<tr>
<th></th>
<th>≥ SPM8</th>
<th>Uncorrected</th>
<th>FDR</th>
<th>FWE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voxel-wise</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>≠</td>
</tr>
<tr>
<td>Cluster-wise</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Peak-wise</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
</tbody>
</table>

- SPM 8 placed new emphasis on peak inference, removed voxel-wise FDR
  - FWE Voxel-wise & Peak-wise equivalent
  - FDR Voxel-wise & Peak-wise **not** equivalent!
- To get voxel FDR, edit `spm_defaults.m` or do `global defaults; defaults.stats.topoFDR=0;`
Cluster FDR: Example Data

Level 5% **Voxel-FWE**

- P = 0.001 cluster-forming thresh
  - k_{FWE} = 241, 5 clusters

Level 5% **Voxel-FDR**

- P = 0.001 cluster-forming thresh
  - k_{FDR} = 138, 6 clusters

Level 5% **Cluster-FWE**

- P = 0.01 cluster-forming thresh
  - k_{FWE} = 1132, 4 clusters

Level 5% **Cluster-FDR**

- P = 0.01 cluster-forming thresh
  - k_{FDR} = 1132, 4 clusters
Conclusions

• Thresholding is not modeling!
  – Just inference on a feature of a statistic image

• Many features to choose from
  – Voxel-wise, cluster-wise, peak-wise…

• FWER
  – Very specific, not very sensitive

• FDR
  – Voxel-wise: Less specific, more sensitive
  – Cluster-, Peak-wise: Similar to FWER
References


