The past and future of Random Field Theory for neuroimaging inference

Thomas Nichols, Ph.D.
Department of Statistics & WMG
University of Warwick
http://warwick.ac.uk/tenichols

January 11, 2016
Overview

• Multiple Comparisons Problem (MCP)
  – Which of my 100,000 voxels are “active”

• Controlling MCP with FWE methods
  – Random Field Theory
  – Permutation

• Evaluations
  – Real data & simulations

• Conclusions
Functional Magnetic Resonance Imaging (fMRI)

- Magnetic properties of blood vary
  - Blue blood $\rightarrow$ Red blood
  - Paramagnetic $\rightarrow$ Diamagnetic

- BOLD
  - Blood Oxygenation Level Dependent effect
  - $\uparrow$ Blood flow $\uparrow$ fMRI Signal

[Image of brain scan and graph]
fMRI Perspective

- 4-Dimensional Data
  - 1,000 multivariate observations, each with 100,000 elements
  - 100,000 time series, each with 1,000 observations
- Usual approach is the time-series perspective
Hypothesis Testing in fMRI

- Massively Univariate Modeling
  - Fit model at each voxel
  - Create statistic images of effect
- Which of 100,000 voxels are significant?
  - $\alpha=0.05 \Rightarrow 5,000$ false positives!

$t > 0.5 \quad t > 1.5 \quad t > 2.5 \quad t > 3.5 \quad t > 4.5 \quad t > 5.5 \quad t > 6.5$
Multiple Comparisons Problem (MCP)

• Standard Hypothesis Test
  – Controls Type I error of each test, at say 5%
  – “Type I Error” only defined for single test

• Must control false positive rate over image
  – What false positive rate?
  – Chance of 1 or more Type I errors
  – Chance of 50 or more?
  – Expected fraction of false positives?
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)
  – R voxels declared active, V falsely so
    • Observed false discovery rate: V/R
  – FDR = E(V/R)
FWER MCP Solutions

- Bonferroni
- Maximum Distribution Methods
  - Random Field Theory
  - Permutation
FWER MCP Solutions: Bonferroni

• Based on truncation of Boole’s formula

\[
\text{FWER} = P(\text{FWE}) \\
= P\left( \bigcup_i \{T_i \geq u\} \mid H_0 \right) \\
\leq \sum_i P( T_i \geq u \mid H_0)
\]

• Corrected Threshold
  – Use P-value threshold \( \alpha = 0.05/V \)
    • to test \( V \) voxels
  – Or statistic threshold \( u_\alpha : P( T_i \geq u_\alpha \mid H_0) = \alpha \)

• Corrected P-value
  – \( \min\{ \text{P-value} \times V, 1 \} \)
FWER MCP Solutions

- Bonferroni
- Maximum Distribution Methods
  - Random Field Theory
  - Permutation
FWER MCP Solutions: Controlling FWER w/ Max

• FWER & distribution of maximum

\[
\text{FWER} = P(\text{FWE}) \\
= P\left( \bigcup_i \{T_i \geq u\} \mid H_0 \right) \\
= P\left( \max_i T_i \geq u \mid H_0 \right)
\]

• 100(1-\(\alpha\))\%ile of max distn controls FWER

\[
\text{FWER} = P\left( \max_i T_i \geq u_\alpha \mid H_0 \right) = \alpha
\]

– where

\[
u_\alpha = F_{\max}(1-\alpha)
\]
FWER MCP Solutions

• Bonferroni
• Maximum Distribution Methods
  – Random Field Theory
  – Permutation
FWER MCP Solutions: Random Field Theory

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

No holes
Never more than 1 blob
RFT Details: Expected Euler Characteristic

\[ E(\chi_u) \approx \lambda(\Omega) \sqrt{|\Lambda|} (u^2 - 1) \exp(-u^2/2) / (2\pi)^2 \]

- \( \Omega \rightarrow \) Search region \( \Omega \subset \mathbb{R}^3 \)
- \( \lambda(\Omega) \rightarrow \) volume
- \( \sqrt{|\Lambda|} \rightarrow \) roughness

**Assumptions**
- Multivariate Normal
- Stationary*
- ACF twice differentiable at 0

* Stationary
- Only cluster results need stationary
- Most accurate when stat. holds

* Only very upper tail approximates
1 - \( F_{\max}(u) \)
Random Field Theory
Smoothness Parameterization

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

$$\Lambda = \text{Var} \left( \frac{\partial G}{\partial(x, y, z)} \right) = \begin{pmatrix} \text{Var} \left( \frac{\partial G}{\partial x} \right) & \text{Cov} \left( \frac{\partial G}{\partial x}, \frac{\partial G}{\partial y} \right) & \text{Cov} \left( \frac{\partial G}{\partial x}, \frac{\partial G}{\partial z} \right) \\ \text{Cov} \left( \frac{\partial G}{\partial y}, \frac{\partial G}{\partial x} \right) & \text{Var} \left( \frac{\partial G}{\partial y} \right) & \text{Cov} \left( \frac{\partial G}{\partial y}, \frac{\partial G}{\partial z} \right) \\ \text{Cov} \left( \frac{\partial G}{\partial z}, \frac{\partial G}{\partial x} \right) & \text{Cov} \left( \frac{\partial G}{\partial z}, \frac{\partial G}{\partial y} \right) & \text{Var} \left( \frac{\partial G}{\partial z} \right) \end{pmatrix}$$

- 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$_x$ × FWHM$_y$ × FWHM$_z$
  - RESEL Count $R$
    - $R = \lambda(\Omega) \sqrt{|\Lambda|}$  \(\leftarrow\) *The only data-dependent part of $E(\chi_u)$*
    - Volume of search region in units of smoothness
    - Eg: 10 voxels, 2.5 voxel FWHM smoothness, 4 RESELS

- *Wrong* RESEL interpretation
  - “Number of independent ‘things’ in the image”
Random Field Intuition

• Corrected P-value for voxel value $t$

$$P_c = P(\text{max } T > t) 
\approx E(\chi_i) 
\approx \lambda(\Omega) |\Lambda|^{1/2} t^2 \exp(-t^2/2)$$

• Statistic value $t$ increases
  – $P_c$ decreases (but only for large $t$)

• Search volume increases
  – $P_c$ increases (more severe MCP)

• Roughness increases (Smoothness decreases)
  – $P_c$ increases (more severe MCP)
RFT Details:
Super General Formula

- General form for expected Euler characteristic
  - $\chi^2, F, \& t$ fields • restricted search regions • $D$ dimensions •

$$E[\chi_u(\Omega)] = \sum_d R_d(\Omega) \rho_d(u)$$

$R_d(\Omega)$: $d$-dimensional Minkowski functional of $\Omega$
- function of dimension, space $\Omega$ and smoothness:

- $R_0(\Omega) = \chi(\Omega)$ Euler characteristic of $\Omega$
- $R_1(\Omega) = \text{resel diameter}$
- $R_2(\Omega) = \text{resel surface area}$
- $R_3(\Omega) = \text{resel volume}$

$\rho_d(\Omega)$: $d$-dimensional EC density of $Z(x)$
- function of dimension and threshold, specific for RF type:

E.g. Gaussian RF:

- $\rho_0(u) = 1 - \Phi(u)$
- $\rho_1(u) = (4 \ln2)^{1/2} \exp(-u^2/2) / (2\pi)$
- $\rho_2(u) = (4 \ln2) \exp(-u^2/2) / (2\pi)^{3/2}$
- $\rho_3(u) = (4 \ln2)^{3/2} (u^2 -1) \exp(-u^2/2) / (2\pi)^2$
- $\rho_4(u) = (4 \ln2)^2 (u^3 -3u) \exp(-u^2/2) / (2\pi)^{5/2}$
Random Field Theory
Cluster Size Tests

- Expected Cluster Size
  - $E(S) = E(N)/E(L)$
  - $S$ cluster size
  - $N$ suprathreshold volume
    $\lambda(\{T > u_{clus}\})$
  - $L$ number of clusters
- $E(N) = \lambda(\Omega) P(T > u_{clus})$
- $E(L) \approx E(\chi_u)$
  - Assuming no holes
Random Field Theory
Cluster Size Distribution

- Gaussian Random Fields (Nosko, 1969)
  \[ S^{2/D} \sim \text{Exp} \left( \frac{E(N)}{\Gamma(D/2+1)E(L)} \right)^{-2/D} \]
  - D: Dimension of RF

- \( t \) Random Fields (Cao, 1999)
  - \( B \): Beta dist
  - \( U' \) s: \( \chi^2 \)' s
  - \( c \) chosen s.t.
    \[ E(S) = \frac{E(N)}{E(L)} \]
Random Field Theory
Cluster Size Corrected P-Values

• Previous results give uncorrected P-value
• Corrected P-value
  – Bonferroni
    • Correct for expected number of clusters
    • Corrected $P^c = \text{E}(L) \ P^{uncorr}$
  – Poisson Clumping Heuristic (Adler, 1980)
    • Corrected $P^c = 1 - \exp( -\text{E}(L) \ P^{uncorr} )$
Random Field Theory
Strengths & Weaknesses

• Closed form results for $E(\chi_u)$
  – $Z$, $t$, $F$, Chi-Squared Continuous RFs
• Results depend only on volume & smoothness

• Smoothness assumed known
• Sufficient smoothness required
  – Results are for continuous random fields
  – Smoothness estimate becomes biased
• Multivariate normality
• Several layers of approximations
FWER MCP Solutions

- Bonferroni
- Maximum Distribution Methods
  - Random Field Theory
  - 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!
Controlling FWER: Permutation Test

- Parametric methods
  - Assume distribution of \( \max \) statistic under null hypothesis

- Nonparametric methods
  - Use data to find distribution of \( \max \) statistic under null hypothesis
  - Again, any max statistic!
Permutation Test & Exchangeability

• Exchangeability is fundamental
  – Def: Distribution of the data unperturbed by permutation
  – Under $H_0$, exchangeability justifies permuting data
  – Allows us to build permutation distribution

• Subjects are exchangeable
  – Under Ho, each subject’s A/B labels can be flipped

• fMRI scans not exchangeable under $H_0$
Permutation Test & Exchangeability

- fMRI scans are not exchangeable
  - Permuting disrupts order, temporal autocorrelation

- **Intra**subject fMRI permutation test
  - Must decorrelate data, model before permuting
  - What is correlation structure?
    - Usually must use parametric model of correlation
    - E.g. Use wavelets to decorrelate
      - Bullmore et al 2001, HBM 12:61-78

- **Inter**subject fMRI permutation test
  - Create difference image for each subject
  - For each permutation, flip sign of some subjects
Real Data Example

- fMRI Study of Working Memory
  - 12 subjects, block design Marshuetz et al (2000)
  - 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, smoothed variance $t$ test
Permutation Test Example

• Permute!
  – $2^{12} = 4,096$ ways to flip 12 A/B labels
  – For each, note maximum of $t$ image
Permutation Test
Example

• Compare with Bonferroni
  – $\alpha = 0.05/110,776$

• Compare with parametric RFT
  – 110,776 $2\times2\times2$mm voxels
  – 5.1×5.8×6.9mm FWHM smoothness
  – 462.9 RESELs
Stat. RF & Bonf. Threshold

$t_{11}$ Statistic, Nonparametric Threshold

$u^\text{Perm} = 7.67$

58 sig. vox.

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

$u^\text{RF} = 9.87$
$u^\text{Bonf} = 9.80$

5 sig. vox.

Smoothed Variance $t$ Statistic, Nonparametric Threshold

Test Level vs. $t_{11}$ Threshold
Does this Generalize?  
RFT vs Bonf. vs Perm.

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>RF</th>
<th>Bonf</th>
<th>Perm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verbal Fluency</td>
<td>4</td>
<td>4701.32</td>
<td>42.59</td>
<td>10.14</td>
</tr>
<tr>
<td>Location Switching</td>
<td>9</td>
<td>11.17</td>
<td>9.07</td>
<td>5.83</td>
</tr>
<tr>
<td>Task Switching</td>
<td>9</td>
<td>10.79</td>
<td>10.35</td>
<td>5.10</td>
</tr>
<tr>
<td>Faces: Main Effect</td>
<td>11</td>
<td>10.43</td>
<td>9.07</td>
<td>7.92</td>
</tr>
<tr>
<td>Faces: Interaction</td>
<td>11</td>
<td>10.70</td>
<td>9.07</td>
<td>8.26</td>
</tr>
<tr>
<td>Item Recognition</td>
<td>11</td>
<td>9.87</td>
<td>9.80</td>
<td>7.67</td>
</tr>
<tr>
<td>Visual Motion</td>
<td>11</td>
<td>11.07</td>
<td>8.92</td>
<td>8.40</td>
</tr>
<tr>
<td>Emotional Pictures</td>
<td>12</td>
<td>8.48</td>
<td>8.41</td>
<td>7.15</td>
</tr>
<tr>
<td>Pain: Warning</td>
<td>22</td>
<td>5.93</td>
<td>6.05</td>
<td>4.99</td>
</tr>
<tr>
<td>Pain: Anticipation</td>
<td>22</td>
<td>5.87</td>
<td>6.05</td>
<td>5.05</td>
</tr>
</tbody>
</table>
Monte Carlo Evaluations

• What’s going wrong?
  – Normality assumptions?
  – Smoothness assumptions?

• Use Monte Carlo Simulations
  – Normality strictly true
  – Compare over range of smoothness, df

• Previous work
  – Gaussian ($Z$) image results well-validated
  – $t$ image results hardly validated at all!
Monte Carlo Evaluations Challenges

• Accurately simulating $t$ images
  – Cannot directly simulate smooth $t$ images
  – Need to simulate $\nu$ smooth Gaussian images
    \( (\nu = \text{degrees of freedom}) \)

• Accounting for all sources of variability
  – Most M.C. evaluations use known smoothness
  – Smoothness not known
  – We estimated it residual images
Monte Carlo Evaluations

- Simulated One Sample T test
  - 32x32x32 Images (32,767 voxels)
  - Smoothness: 0, 1.5, 3, 6,12 FWHM
  - Degrees of Freedom: 9, 19, 29
  - Realizations: 3000

- Permutation
  - 100 relabelings
  - Threshold: 95%ile of permutation dist\(n\) of maximum

- Random Field
  - Threshold: \(\{ u : E(\chi_u | H_0) = 0.05 \} \)
Monte Carlo Evaluations

- Voxel-wise (intensity) Results
  Equivalent Independent Elements?
- Cluster-wise (extent) Results
Familywise Error Thresholds

- RF & Perm adapt to smoothness
- Perm & Truth close
- Bonferroni close to truth for low smoothness
Familywise Rejection Rates

- Bonf good on low df, smoothness
- Bonf bad for high smoothness
- RF only good for high df, high smoothness
- Perm exact more
Familywise Rejection Rates

- Smoothness estimation is not (sole) problem
Monte Carlo Evaluations

- Voxel-wise (intensity) Results
  Equivalent Independent Elements?
- Cluster-wise (extent) Results
Equivalent Independent Elements

- RFT methods not “RESEL Bonferroni”
  - Consider corrected P-values $P^c$ for statistic $t$
    
    $P^c_{\text{Bonf}} \propto V \times e^{t^2/2} t^{-1}$
    $P^c_{\text{RFT}} \propto R \times e^{t^2/2} t^2$

    - No “equivalent” $V$ for all thresholds $t$

- But this assumes RFT works
  - What if there were an equivalent number of independent spatial elements
Equivalent Independent Elements

• FWE control with \( \max_i T_i \)
  
  \[- F_{\max_i T_i}(t) = \prod_i F_{T_i}(t) = (F_T(t))^V \]
  
  \[= (F_T(t))^{\theta V} \text{ for some } \theta? \]

• In terms of P-values
  
  \[- \max_i T_i > t \iff \min_i P_i < \gamma \]
  
  \[- F_{\min_i P_i}(\gamma) = 1-(1-F_P(\gamma))^{\theta V} = 1-(1-\gamma)^{\theta V} \]

• Use simulations to ask...
  
  \[- \text{Is there an } \theta \text{ such that } F_{\min_i P_i}(\gamma) \text{ behaves like the minimum of } \theta V \text{ independent voxels?} \]
Simulations: Min P CDF’s
Min P CDFs

- Higher threshold (smaller P) doesn’t help
- For low / moderate smoothness, equivalent independent approach promising

Cumulative Probability / FWE Corrected P-Value

$df=9$ FWHM=12

$\alpha_{corr} = 0.05$
Monte Carlo Evaluations

• Voxel-wise (intensity) Results
  Equivalent Independent Elements?
• Cluster-wise (extent) Results
Familywise Cluster Size Threshold

- RF & Perm adapt to smoothness
- RFT not bad above 3 FWHM sm.
Familywise Rejection Rates

- Interesting that gets worse with larger df.
FWE Corrected p-values

- For df=9 biased smoothness estimation improves rejection rate
Performance Summary

• Bonferroni
  – Not adaptive to smoothness
  – Not so conservative for low smoothness

• Random Field
  – Adaptive
  – Conservative for low smoothness & df
  – Not so bad for cluster size inference

• Permutation
  – Adaptive (Exact)
Understanding Performance Differences

- **RFT Troubles**
  - Multivariate Normality assumption
    - True by simulation
  - Smoothness estimation
    - Not much impact
  - Smoothness
    - You need lots, more at low df
  - High threshold assumption
    - Doesn’t improve for $\alpha_0$ less than 0.05
Massive Empirical Evaluation

- Monte Carlo doesn’t capture weirdness of real data
- In last 5 years, explosion of open resting fMRI data repositories
  - Suddenly null (task) fMRI data is plentiful
First-Level (single subject) fMRI

- Eklund (2012) analyzed 1,484 resting fMRI datasets from public repositories
- Fed through standard SPM pipeline, with 8 different “pretend” paradigms

<table>
<thead>
<tr>
<th>Paradigm</th>
<th>Activity periods (s)</th>
<th>Rest periods (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>B2</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>B3</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>B4</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>E1</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>E2</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>E3</td>
<td>1–4 (R)</td>
<td>3–6 (R)</td>
</tr>
<tr>
<td>E4</td>
<td>3–6 (R)</td>
<td>4–8 (R)</td>
</tr>
</tbody>
</table>
Computed Familywise Error (FWE) Rates

- Many settings had awful FWE!
  - Block worse than event; fast TR worse than slow
Massive Empirical Evaluation – Take II

- Previous result only for first level fMRI
- 2\textsuperscript{nd} level fMRI doesn’t depend on 1\textsuperscript{st} level P-values
- Data quality also an issue
Massive Empirical Evaluation – Take II

• Same fcon1000 repository, just 2 largest sites: Beijing & Cambridge

• Second level analyses
  – 1-sample t-test: \( n = 20, 40 \)
  – 2-sample t-test: \( n_1 = n_2 = 10, 20 \)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Values used</th>
</tr>
</thead>
<tbody>
<tr>
<td>fMRI data</td>
<td>Beijing (198 subjects), Cambridge (198 subjects)</td>
</tr>
<tr>
<td>Block activity paradigms</td>
<td>B1 (10 s on off), B2 (30 s on off)</td>
</tr>
<tr>
<td>Event activity paradigms</td>
<td>E1 (2 s activation, 6 s rest), E2 (1 - 4 s activation, 3 - 6 s rest, randomized)</td>
</tr>
<tr>
<td>Smoothing</td>
<td>4, 6, 8, 10 mm FWHM</td>
</tr>
<tr>
<td>Analysis type</td>
<td>One sample t-test (group activation), two sample t-test (group difference)</td>
</tr>
<tr>
<td>Number of subjects</td>
<td>20, 40</td>
</tr>
<tr>
<td>Inference level</td>
<td>Voxel, cluster</td>
</tr>
<tr>
<td>Cluster defining threshold</td>
<td>( p = 0.01 (z = 2.3), p = 0.001 (z = 3.1) )</td>
</tr>
</tbody>
</table>
Massive Group fMRI Evaluation

Voxel-wise

- Voxel-wise inference OK
  - Sometimes very conservative!
Massive Group fMRI Evaluation
Cluster-wise CFT $p=0.01$

- Cluster-wise a catastrophe!
  - Rarely valid at cluster forming threshold

(CFT) $p=0.01$ – default CFT in FSL
Massive Group fMRI Evaluation

Cluster-wise CFT $p=0.001$

- Cluster-wise CFT $p=0.001$ better
  - Valid $\approx 50\%$ time, depending on design
Massive Group fMRI Evaluation
“ad hoc”, CFT $p=0.001$ K$>10$

- Authors often use “folk” multiple testing method
- $P=0.001$ and only clusters of 10 voxels or more
- This has 50-80% FWE
Massive Group fMRI Eval: What’s going wrong?

• RFT Assumptions
  – Gaussian errors
  – Spatial ACF has 2 derivatives at origin
  – For cluster-size only
    • Spatial ACF has Gaussian shape
    • CFT “sufficient” high
    • Stationary (spatially homogeneous smoothness)
What’s wrong with FSL’s FLAME1

• Univariate P-values are conservative
  – Nothing to do with RFT!
  – Turns out to be artifact of using completely null data ($\sigma_{BTW} = 0$)
  – Using non-null ($\sigma_{BTW} > 0$) data for same-vs-same 2-group comparison, resolves this. But then results similar (not shown) to FSL OLS 😐
Massive Group fMRI Eval: Spatial ACF

- Much heavier tails than Gaussian pdf!
Massive Group fMRI Eval: Spatial Dist$^n$ of False Clusters

• Great smoothness in “default mode” areas
What always works? Permutation!

• How does this compare on real (non-null) data?

Usually, would say “non-parametric so much less powerful”

In light of evaluations, “non-parametric valid, parametric inflated significance”
Conclusions

• Gaussian Monte Carlo results show…
  – $t$ random field results conservative for
    • Low $df$ & smoothness
    • $9 \, df \leq 12 \text{ FWHM}; \quad 19 \, df < 10 \text{ FWHM}$
  – Nonparametric methods perform well overall

• Real data evaluations
  – RFT Voxel-wise OK, but conservative
  – Cluster-wise $P=0.01$ invalid danger danger danger
  – Cluster-wise $P=0.001$ – sometimes OK, sometimes invalid

• Permutation embarrassingly parallelizable, GPU friendly
  – See BROCCOLI

• Standard tools available!
  – FSL: randomise  SPM: SnPM toolbox
References
