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

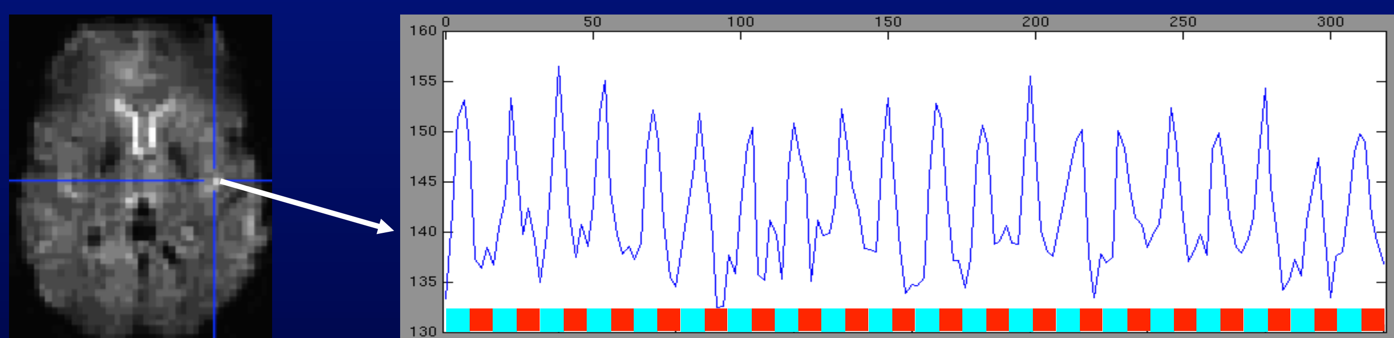
- Multiple Comparisons Problem (MCP)
  - Which of my 100,000 voxels are “active”
- Controlling MCP with FWE methods
  - Random Field Theory
  - Permutation
- Evaluations
  - Real data & simulations
- Conclusions

Functional Magnetic Resonance Imaging (fMRI)

- Magnetic properties of blood vary
  - Blue blood → Red blood
  - Paramagnetic → Diamagnetic
- BOLD
  - Blood Oxygenation Level Dependent effect
  - ↑ Blood flow ↑ fMRI Signal

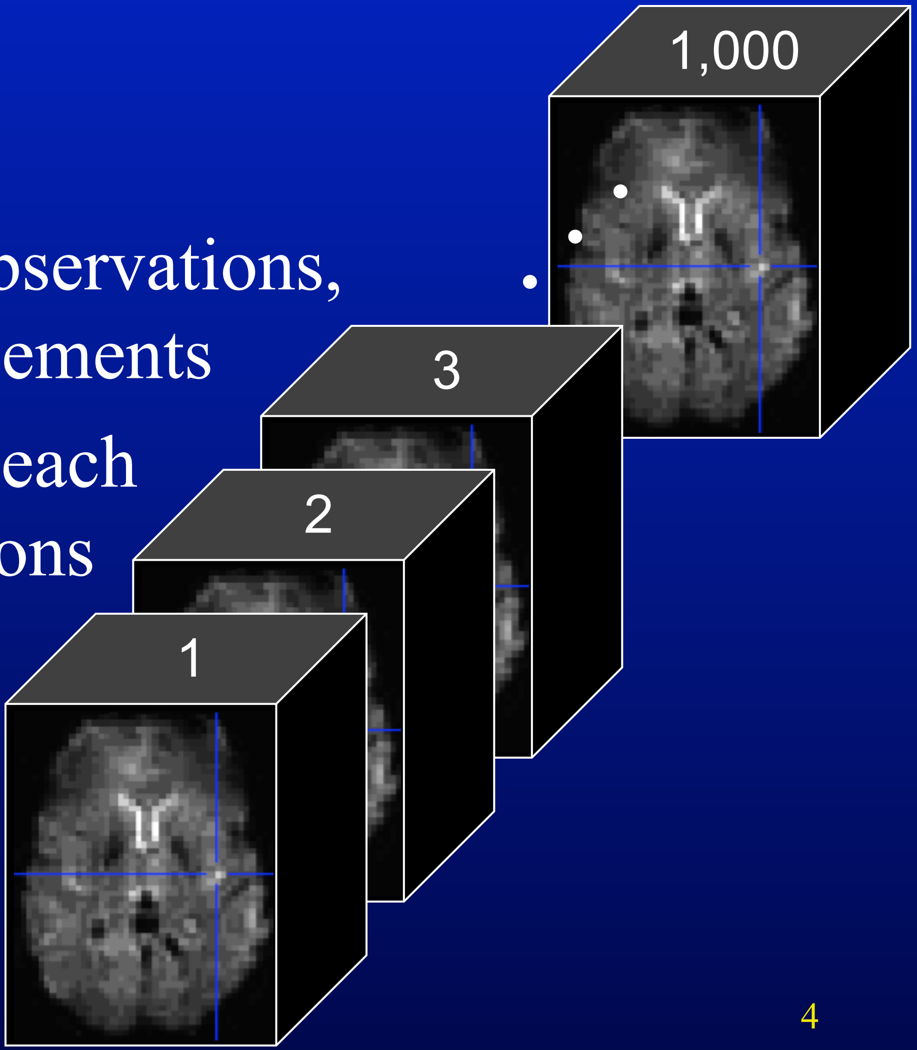
Tap fingers

Rest

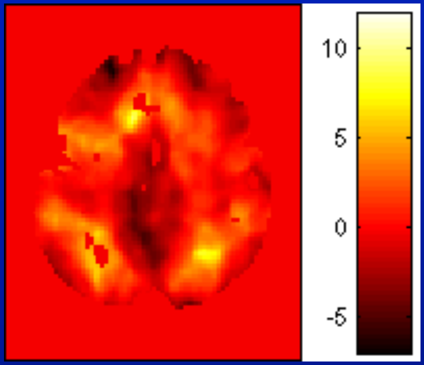


fMRI Perspective

- 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

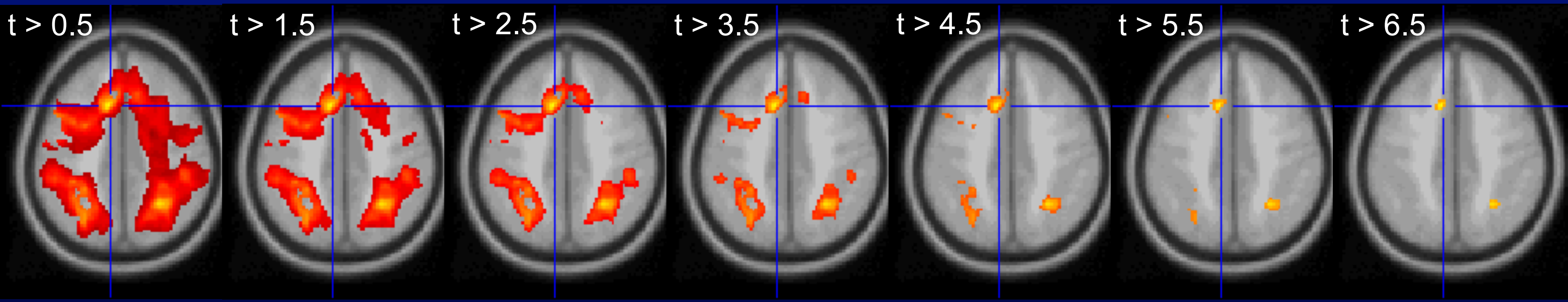


Hypothesis Testing in fMRI



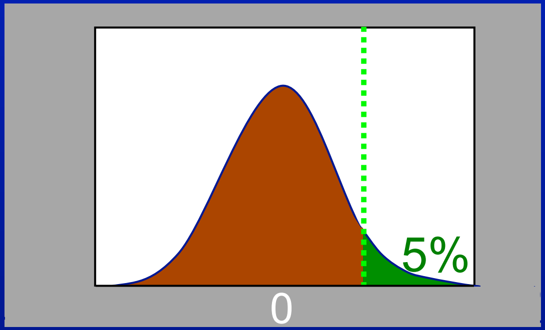
- Massively Univariate Modeling
  - Fit model at each voxel
  - Create statistic images of effect
- Which of 100,000 voxels are significant?
  - α=0.05 $\Rightarrow$ 5,000 false positives!

Must we threshold?



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

$$\begin{aligned} \text{FWER} &= \text{P(FWE)} \\ &= \text{P}(\cup_i \{T_i \geq u\} \mid H_o) \\ &\leq \Sigma_i \, \text{P}(T_i \geq u \mid H_o) \end{aligned}$$

- Corrected Threshold
  - Use P-value threshold $\alpha = 0.05/V$
    - to test V voxels
  - Or statistic threshold u_α : $\text{P}(T_i \geq u_\alpha \mid H_o) = \alpha$
- 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

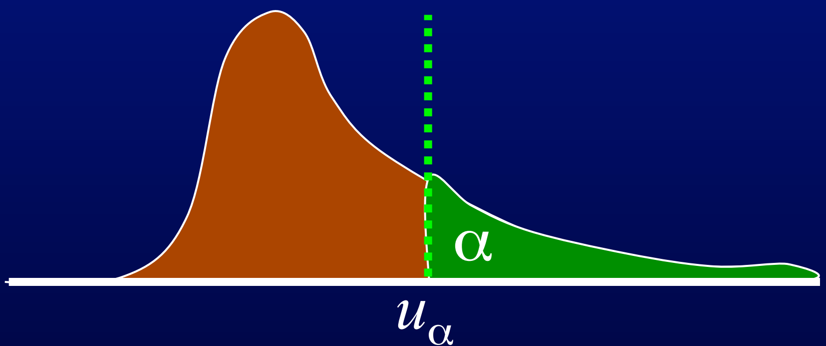
$$\begin{aligned}\text{FWER} &= \text{P(FWE)} \\ &= \text{P}(\cup_i \{T_i \geq u\} \mid H_o) \\ &= \text{P}(\max_i T_i \geq u \mid H_o)\end{aligned}$$

- 100(1-α)%ile of max dist^n controls FWER

$$\text{FWER} = \text{P}(\max_i T_i \geq u_\alpha \mid H_o) = \alpha$$

  – where

$$u_\alpha = F^{-1}_{\max}(1-\alpha)$$



FWER MCP Solutions

- Bonferroni
- Maximum Distribution Methods
  - Random Field Theory
  - Permutation

FWER MCP Solutions: Random Field Theory

- Euler Characteristic χ_u
  - Topological Measure
    - #blobs - #holes
  - At high thresholds, just counts blobs
  - FWER = P(Max voxel ≥ u | H_o)

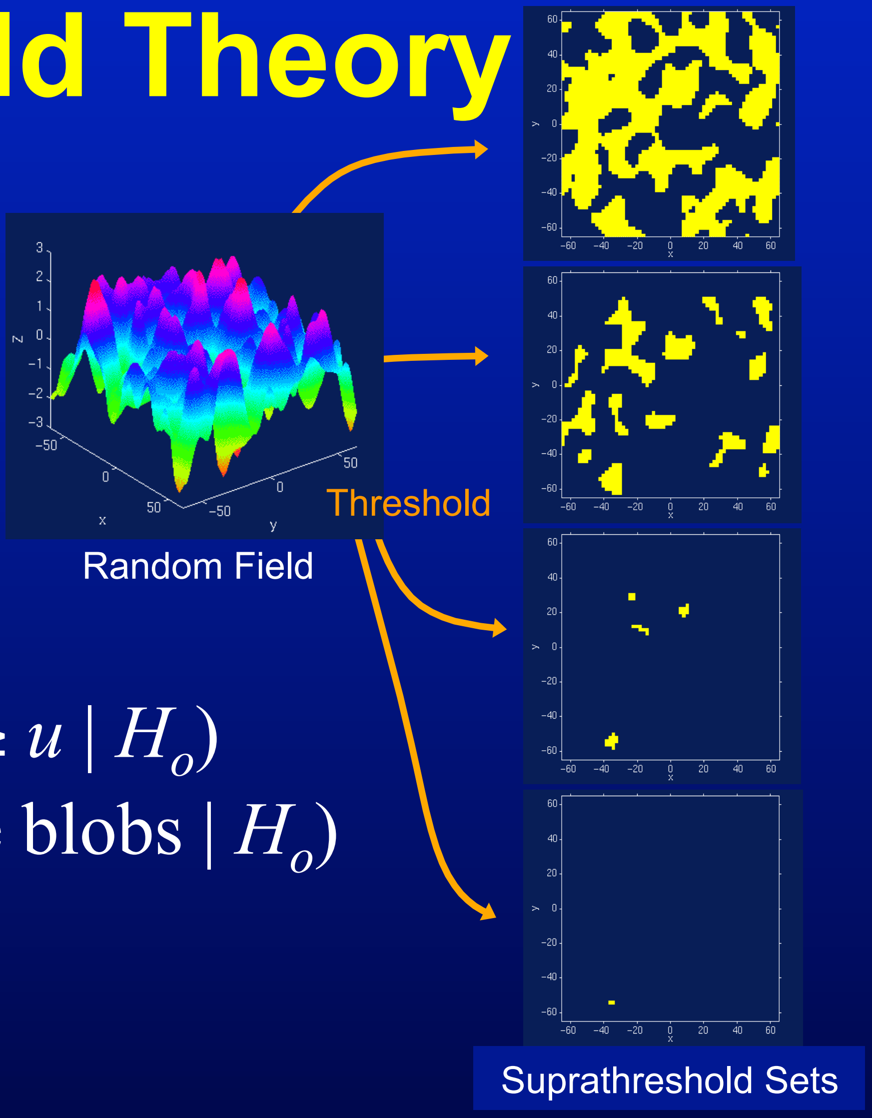
    = P(One or more blobs | H_o)

    ≈ P(χ_u ≥ 1 | H_o)

    ≈ E(χ_u | H_o)

No holes

Never more than 1 blob



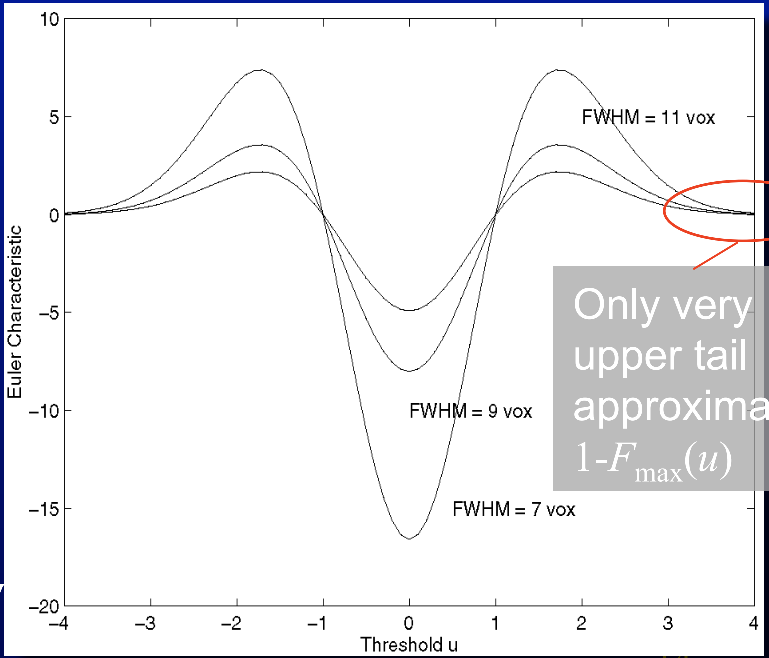
RFT Details: Expected Euler Characteristic

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

- Ω → Search region $\Omega \subset \mathcal{R}^3$
- $\lambda(\Omega)$ → volume
- $|\Lambda|^{1/2}$ → roughness

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

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



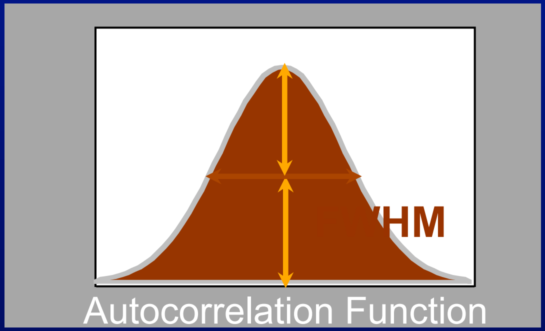
Only very upper tail approximates $1\text{-}F_{max}(u)$

Random Field Theory Smoothness Parameterization

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

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

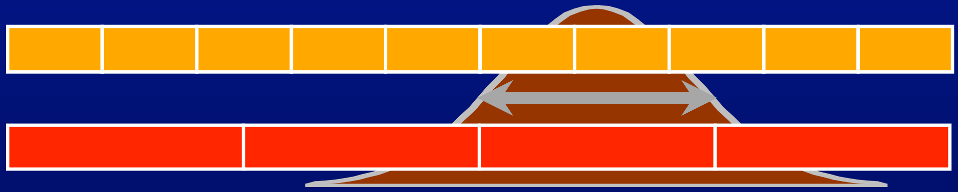
- Smoothness parameterized as Full Width at Half Maximum
  - FWHM of Gaussian kernel needed to smooth a white noise random field to roughness Λ



$$|\Lambda|^{1/2} = \frac{(4\log 2)^{3/2}}{\mathrm{FWHM}_x\mathrm{FWHM}_y\mathrm{FWHM}_z}.$$

Random Field Theory Smoothness Parameterization

- RESELS – Resolution Elements
  - 1 RESEL = $FWHM_x \times FWHM_y \times FWHM_z$
  - RESEL Count R
    - $R = \lambda(\Omega) \sqrt{|\Lambda|}$ ← *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"
    - Nichols & Hayasaka, 2003, Stat. Meth. in Med. Res.

Random Field Intuition

- Corrected P-value for voxel value t

$$
\begin{aligned}
P^c &= \mathrm{P}(\max T > t) \\
&\approx \mathrm{E}(\chi_t) \\
&\approx \lambda(\Omega)\, |\Lambda|^{1/2}\, t^2 \exp(-t^2/2)
\end{aligned}
$$

- 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
  - χ^2, F, & t fields • restricted search regions • D dimensions •

$$\mathrm{E}[\chi_u(\Omega)] = \Sigma_d \, R_d(\Omega) \, \rho_d(u)$$

$R_d(\Omega)$: d-dimensional Minkowski functional of Ω

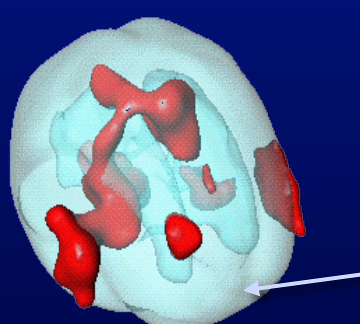
– function of dimension, space Ω and smoothness:

$R_0(\Omega)$ = $\chi(\Omega)$ Euler characteristic of Ω

$R_1(\Omega)$ = resel diameter

$R_2(\Omega)$ = resel surface area

$R_3(\Omega)$ = resel volume



Ω

$\rho_d(\Omega)$: d-dimensional EC density of $Z(\underline{x})$

– function of dimension and threshold, specific for RF type:

E.g. Gaussian RF:

$$\rho_0(u) = 1 - \Phi(u)$$

$$\rho_1(u) = (4 \ln 2)^{1/2} \exp(-u^2/2) / (2\pi)$$

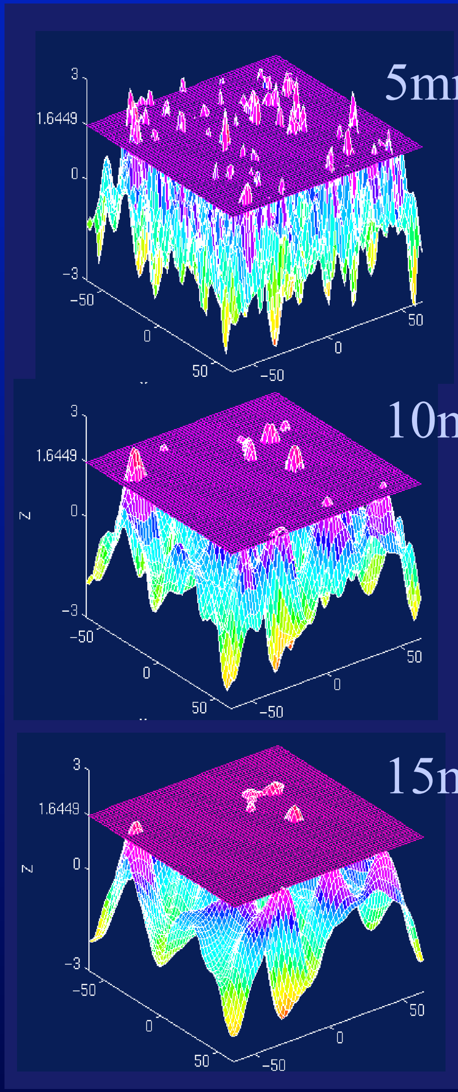
$$\rho_2(u) = (4 \ln 2) \exp(-u^2/2) / (2\pi)^{3/2}$$

$$\rho_3(u) = (4 \ln 2)^{3/2} (u^2 - 1) \exp(-u^2/2) / (2\pi)^2$$

$$\rho_4(u) = (4 \ln 2)^2 (u^3 - 3u) \exp(-u^2/2) / (2\pi)^{5/2}$$

Random Field Theory Cluster Size Tests

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



5mm FWHM

10mm FWHM

15mm FWHM

Random Field Theory Cluster Size Distribution

- Gaussian Random Fields (Nosko, 1969)

$$S^{2/D} \sim Exp\left(\left[\frac{E(N)}{\Gamma(D/2+1)E(L)}\right]^{-2/D}\right)$$

  - D: Dimension of RF
- t Random Fields (Cao, 1999)
  - B: Beta dist^n
  - U' s: χ^2' s
  - c chosen s.t. E(S) = E(N) / E(L)

$$S \sim cB^{1/2}\left[\frac{U_0^D}{\Pi_{b=0}^D U_b}\right]^{2/D}$$

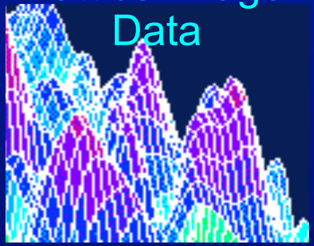
Random Field Theory
Cluster Size Corrected P-Values

- Previous results give uncorrected P-value
- Corrected P-value
  - Bonferroni
    - Correct for expected number of clusters
    - Corrected $P^c = E(L)\, P^{uncorr}$
  - Poisson Clumping Heuristic (Adler, 1980)
    - Corrected $P^c = 1 - \exp(-E(L)\, P^{uncorr})$

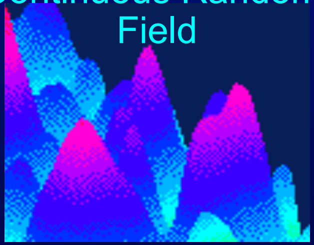
Random Field Theory Strengths & Weaknesses

- Closed form results for E(χ_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



≈

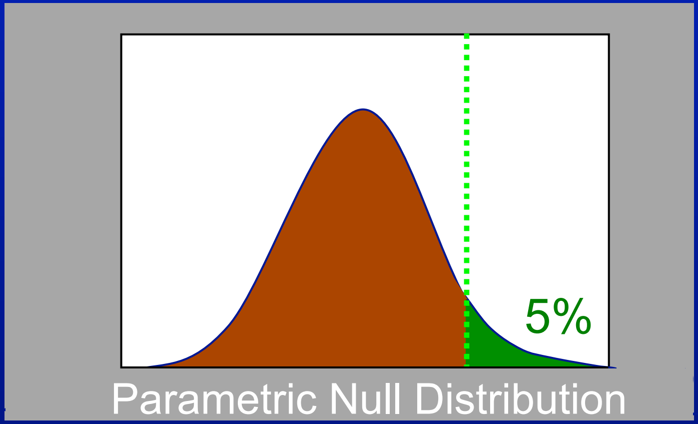


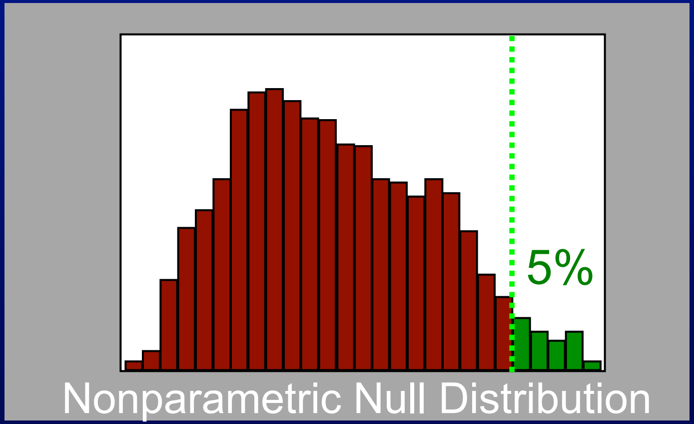
FWER MCP Solutions

- Bonferroni
- Maximum Distribution Methods
  - Random Field Theory
  - Permutation

Nonparametric Permutation Test

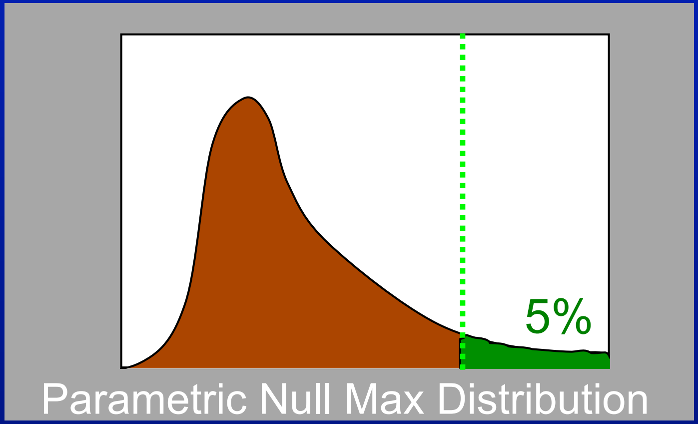
- Parametric methods
  - Assume distribution of statistic under null hypothesis
- Nonparametric methods
  - Use *data* to find distribution of statistic under null hypothesis
  - Any statistic!

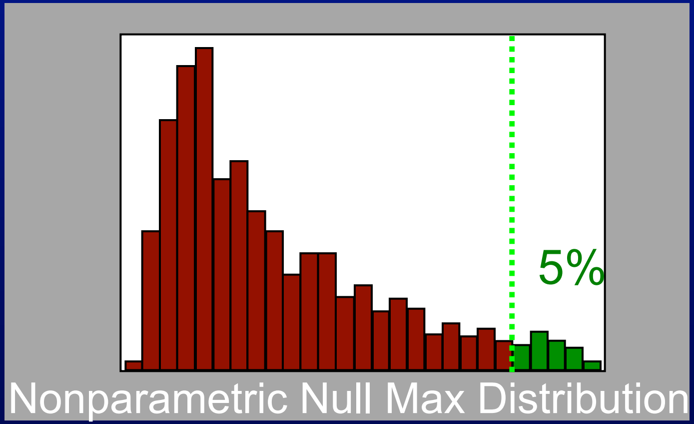




Controlling FWER: Permutation Test

- 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

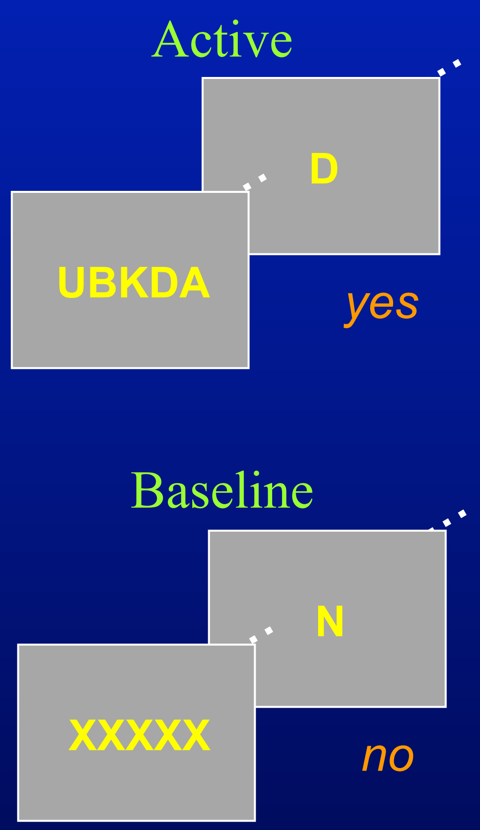
- Exchangeability is fundamental
  - Def: Distribution of the data unperturbed by permutation
  - Under H_0, exchangeability justifies permuting data
  - Allows us to build permutation distribution
- Subjects are exchangeable
  - Under Ho, each subject's A/B labels can be flipped
- fMRI scans not exchangeable 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

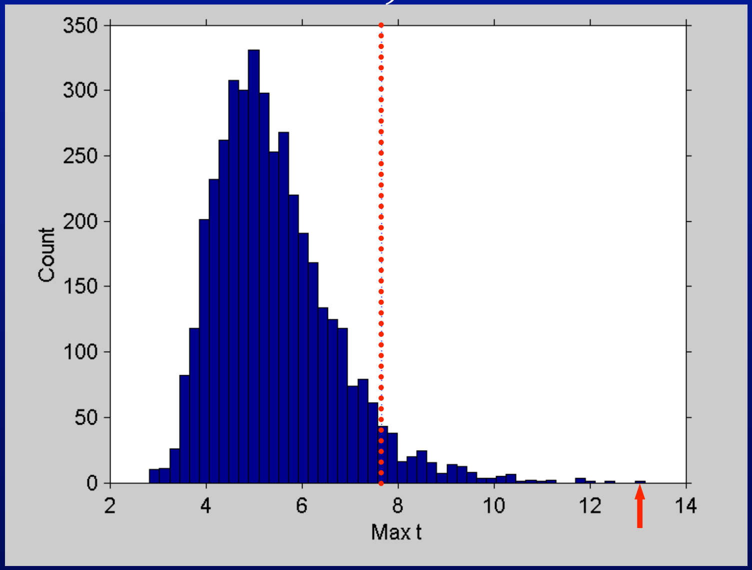
Real Data Example

- 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

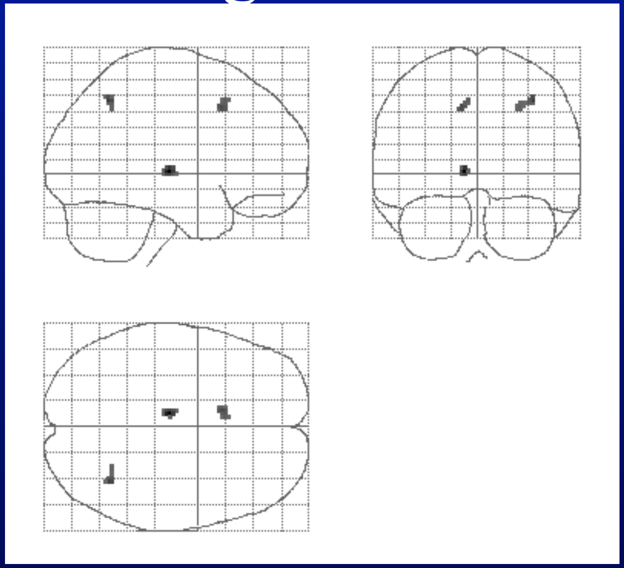


Permutation Test Example

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



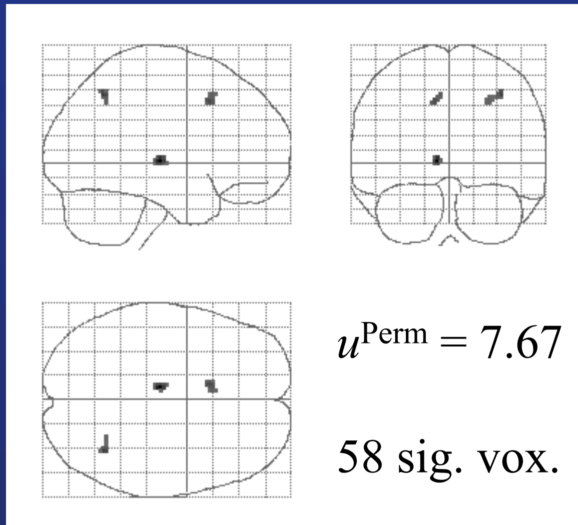
Permutation Distribution Maximum *t*



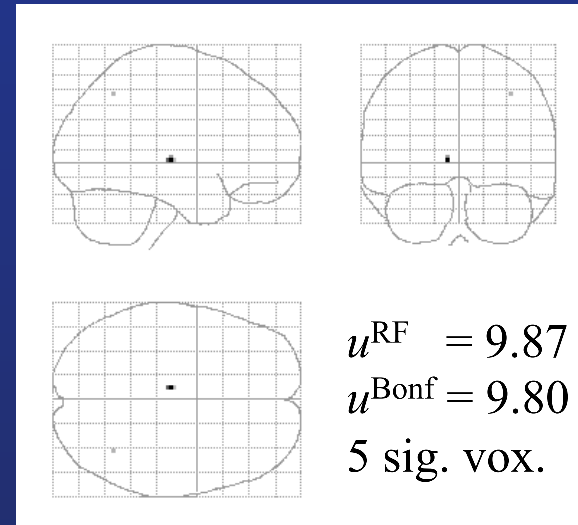
Maximum Intensity Projection Thresholded *t*

Permutation Test Example

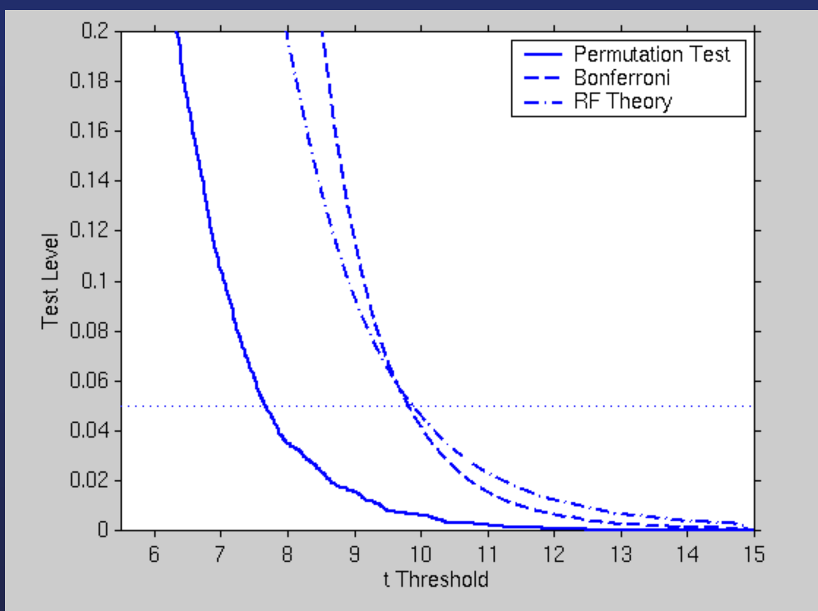
- Compare with Bonferroni
  - $\alpha = 0.05/110{,}776$
- Compare with parametric RFT
  - 110,776 $2\times2\times2$mm voxels
  - $5.1\times5.8\times6.9$mm FWHM smoothness
  - 462.9 RESELs



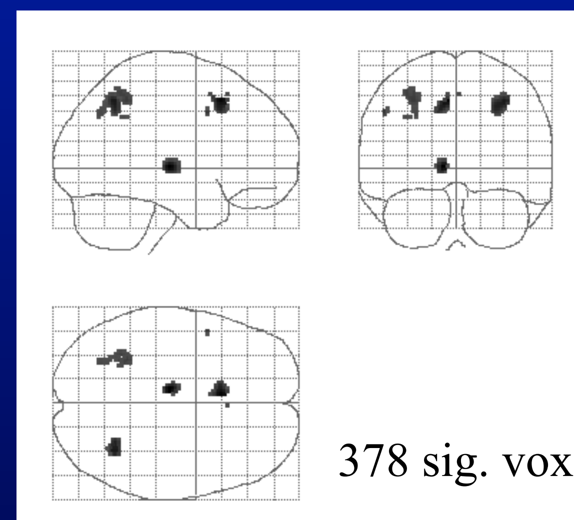
t_{11} Statistic, Nonparametric Threshold



t_{11} Statistic, RF & Bonf. Threshold



Test Level vs. t_{11} Threshold



Smoothed Variance *t* Statistic, Nonparametric Threshold

Does this Generalize?
RFT vs Bonf. vs Perm.

		t Threshold (0.05 Corrected)		
	df	RF	Bonf	Perm
Verbal Fluency	4	4701.32	42.59	10.14
Location Switching	9	11.17	9.07	5.83
Task Switching	9	10.79	10.35	5.10
Faces: Main Effect	11	10.43	9.07	7.92
Faces: Interaction	11	10.70	9.07	8.26
Item Recognition	11	9.87	9.80	7.67
Visual Motion	11	11.07	8.92	8.40
Emotional Pictures	12	8.48	8.41	7.15
Pain: Warn ing	22	5.93	6.05	4.99
Pain: Anticipation	22	5.87	6.05	5.05

Monte Carlo Evaluations

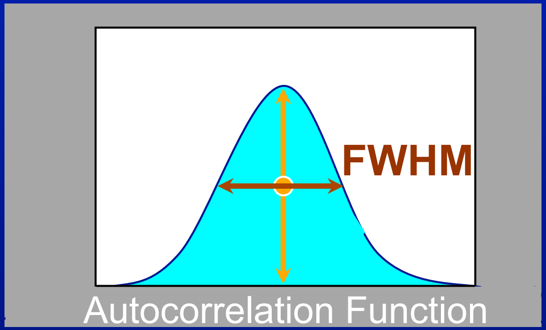
- What's going wrong?
  - Normality assumptions?
  - Smoothness assumptions?
- Use Monte Carlo Simulations
  - Normality strictly true
  - Compare over range of smoothness, df
- Previous work
  - Gaussian (Z) image results well-validated
  - t image results hardly validated at all!

Monte Carlo Evaluations Challenges

- Accurately simulating *t* images
  - Cannot directly simulate smooth *t* images
  - Need to simulate ν smooth Gaussian images
    (ν = 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



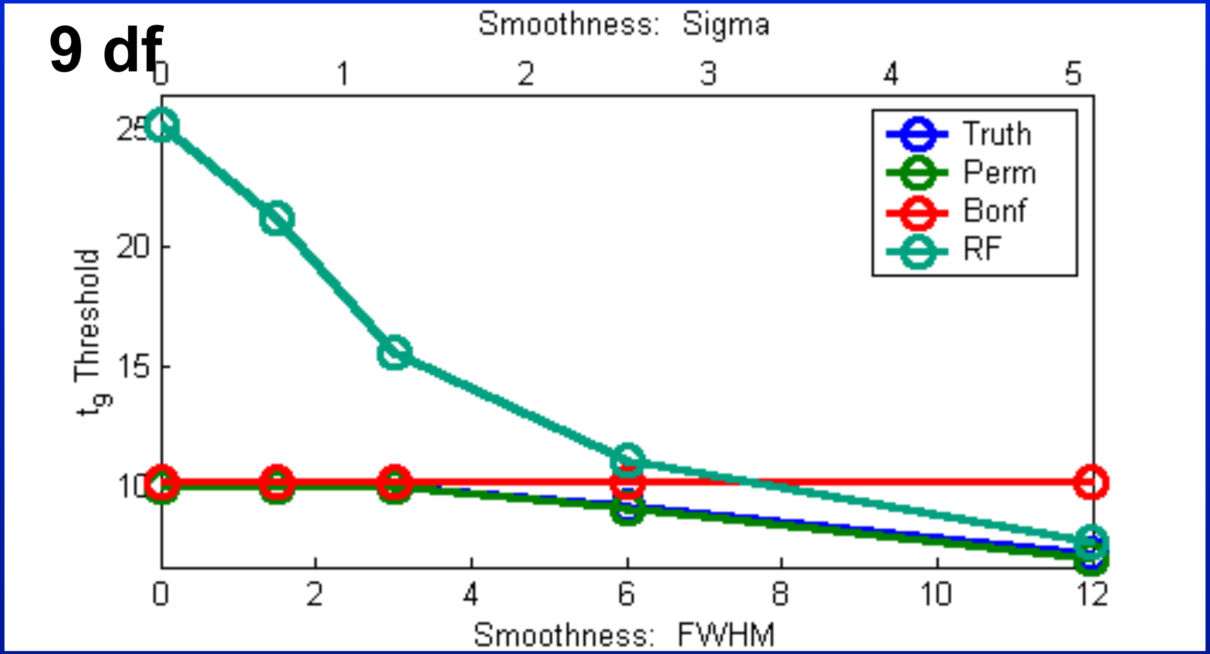
- Permutation
  - 100 relabelings
  - Threshold: 95%ile of permutation dist^n of maximum
- Random Field
  - Threshold: $\{ u : E(\chi_u \mid H_o) = 0.05 \}$

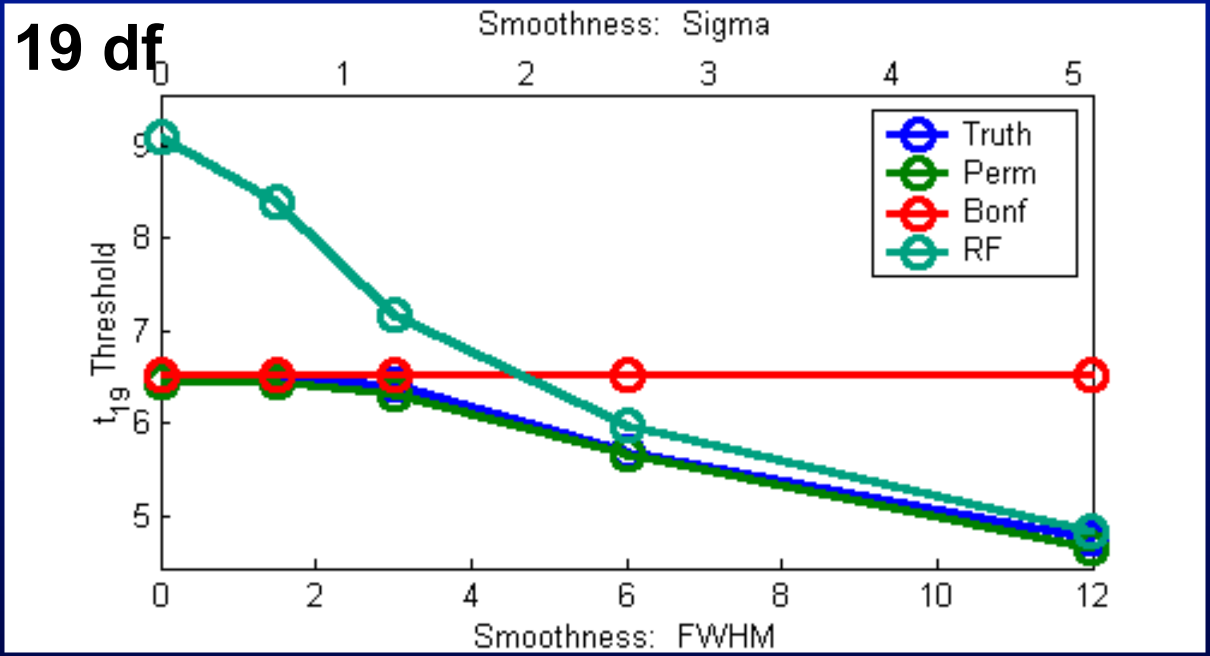
Monte Carlo Evaluations

- 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

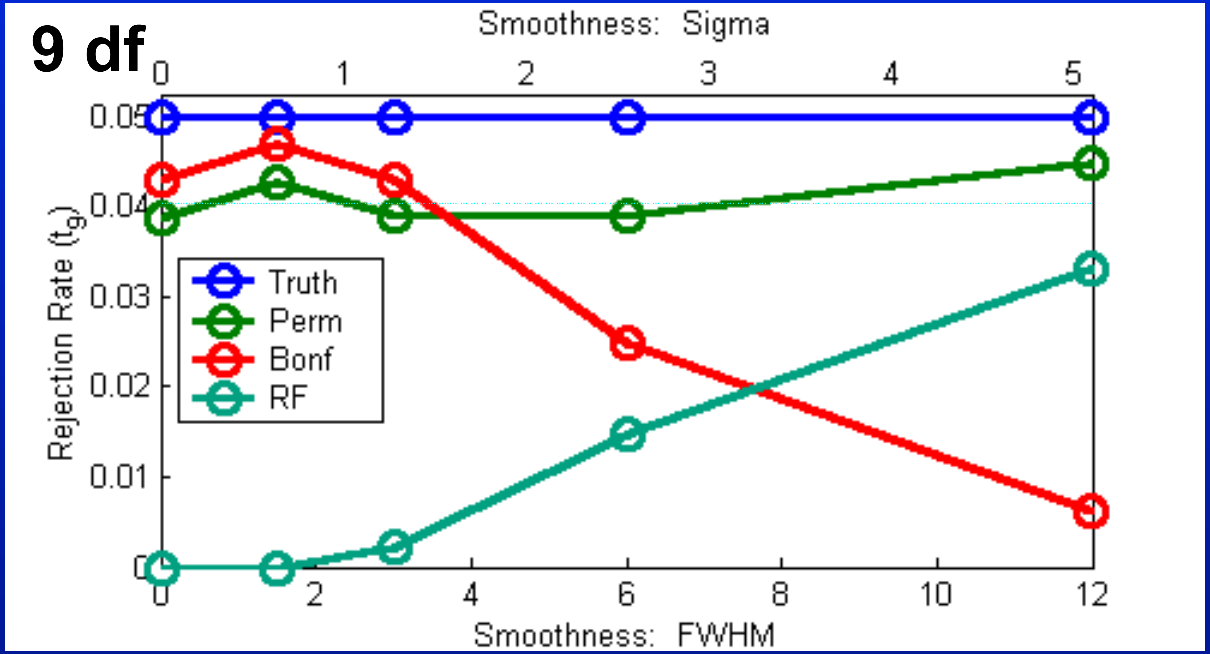


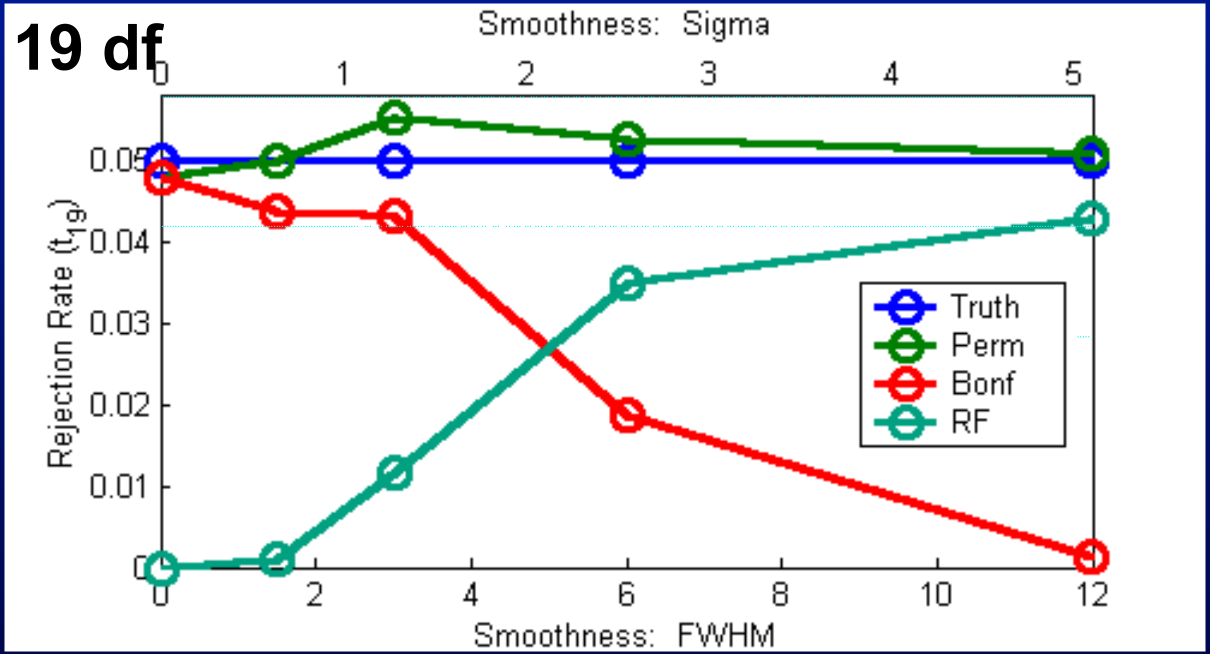


more

Familywise Rejection Rates

- Bonf good on low df, smoothness
- Bonf bad for high smoothness
- RF only good for high df, high smoothness
- Perm exact

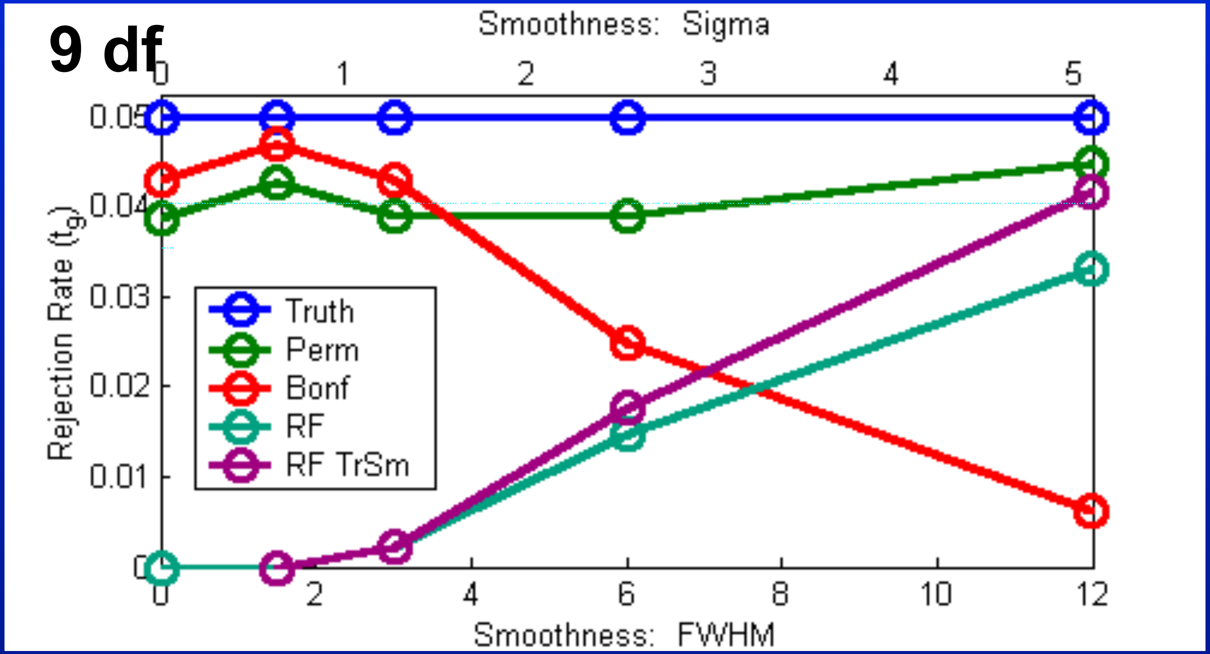


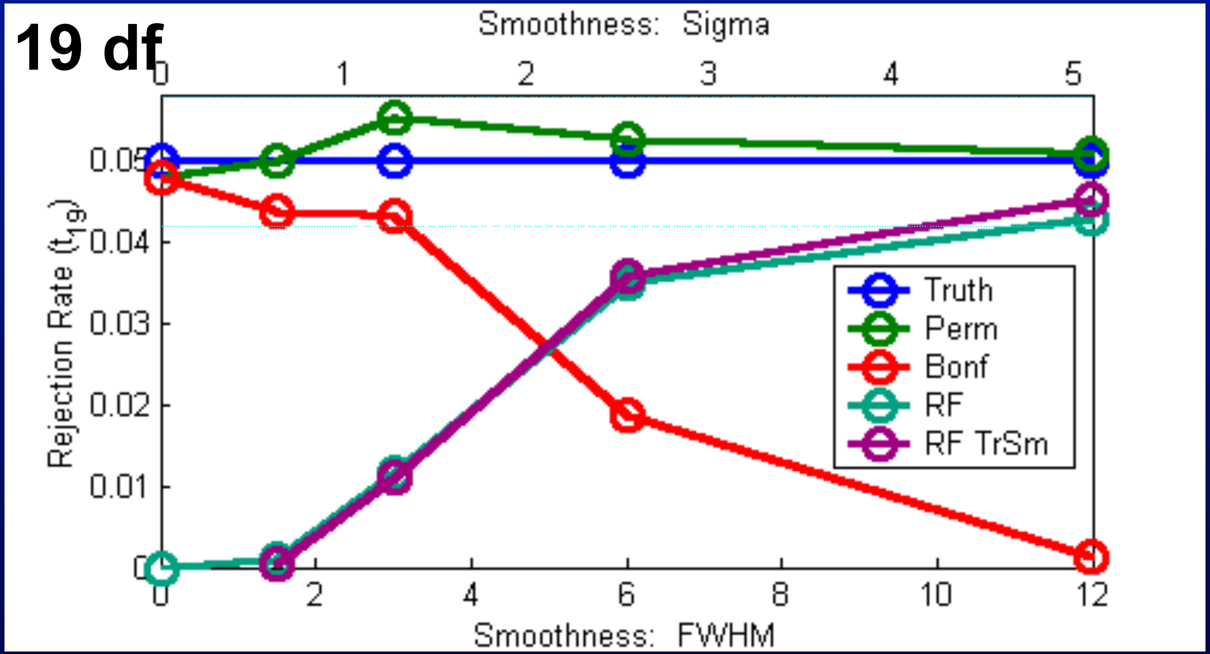


more

Familywise Rejection Rates

- Smoothness estimation is not (sole) problem





cont

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

$$P^c_{\text{Bonf}} \propto V \times e^{t^2/2} t^{-1}$$
$$P^c_{\text{RFT}} \propto R \times e^{t^2/2} t^{2}$$

V – # of voxels

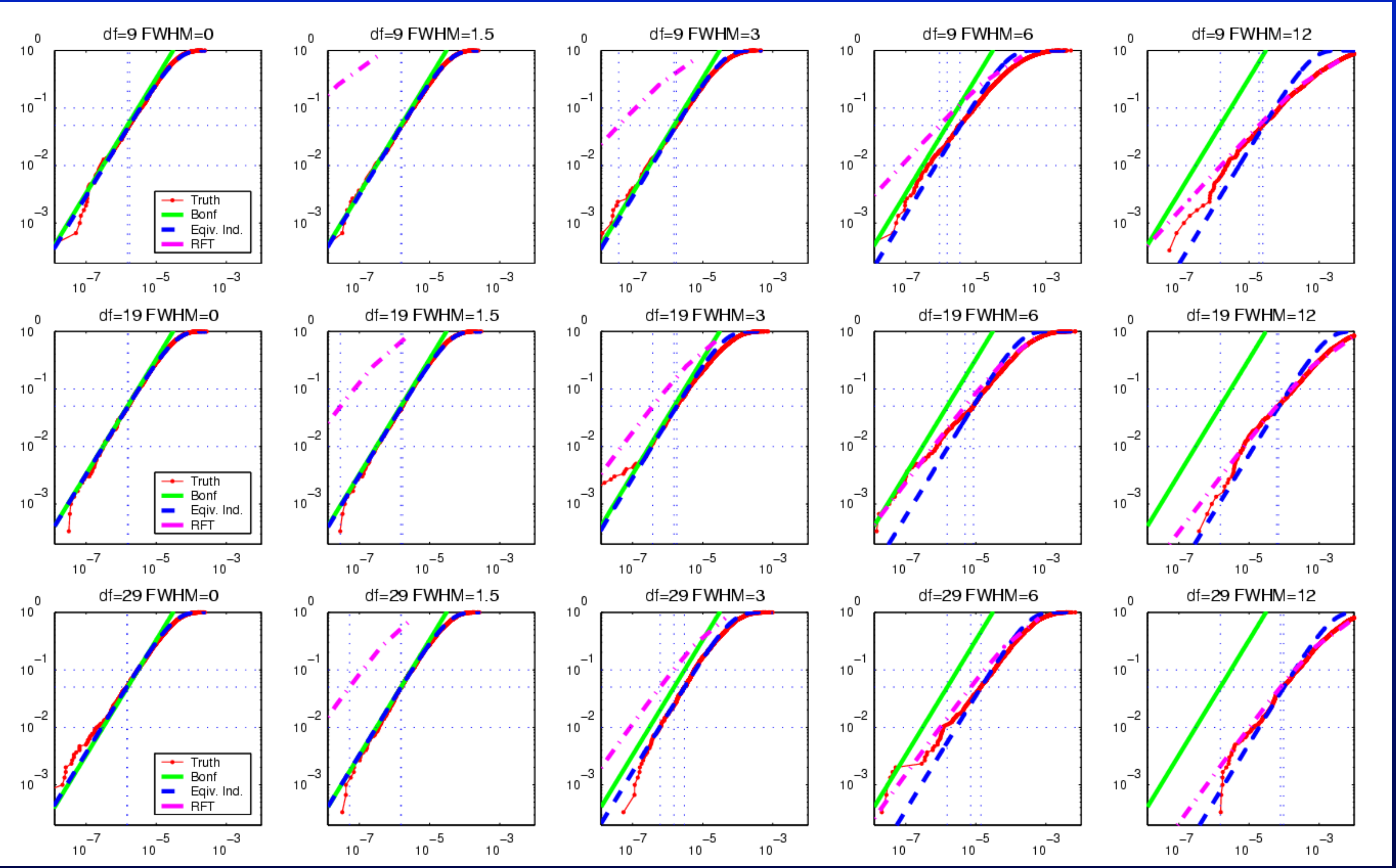
R – # of RESELs

  - 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

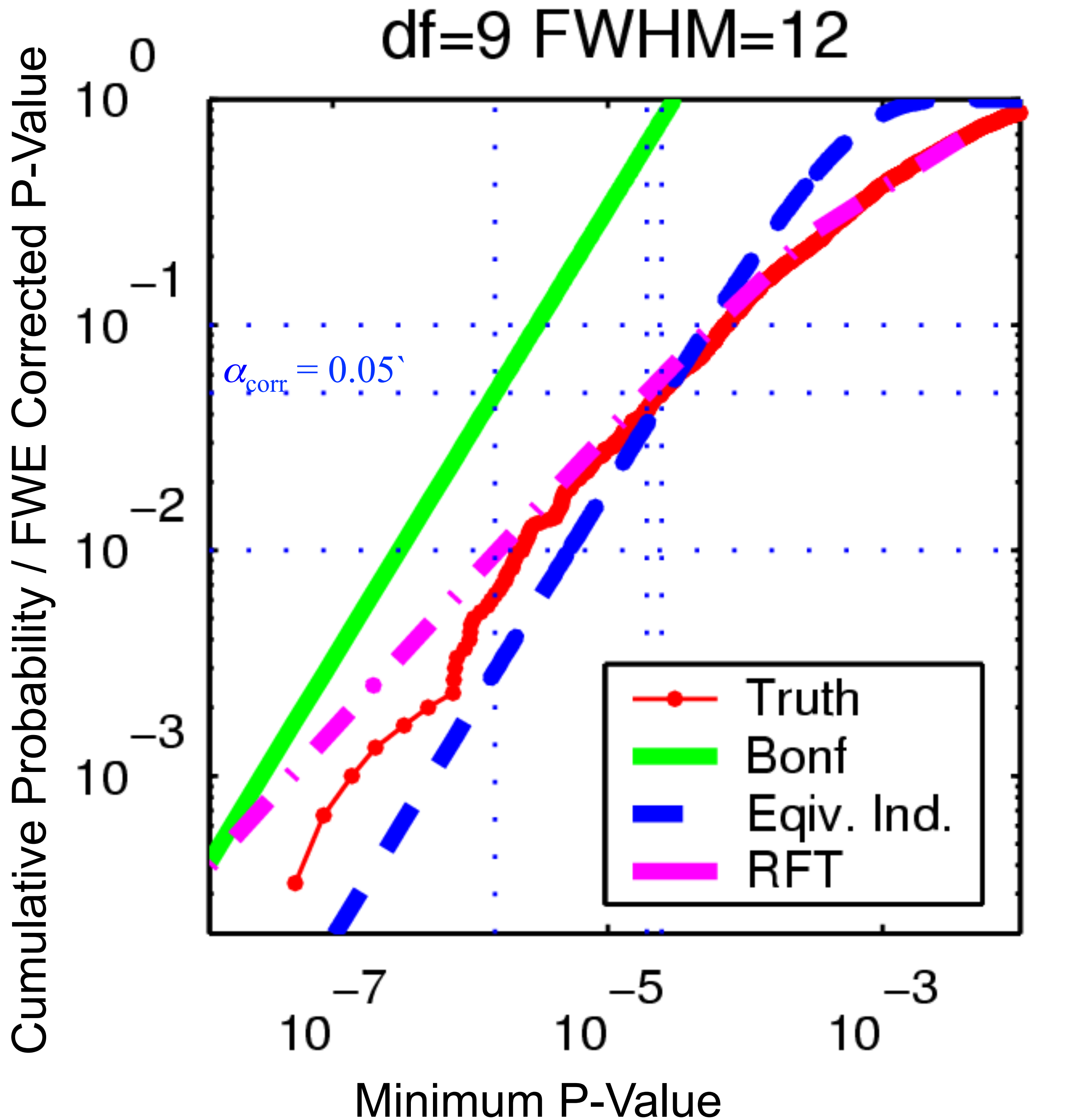
- FWE control with $\max_i T_i$
  - $F_{\max_i T_i}(t) = \Pi_i F_{T_i}(t) = (F_T(t))^V$
    $= (F_T(t))^{\theta V}$ for some θ?
- In terms of P-values
  - $\max_i T_i > t \leftrightarrow \min_i P_i < \gamma$
  - $F_{\min_i P_i}(\gamma) = 1\text{-}(1\text{-}F_P(\gamma))^{\theta V} = 1\text{-}(1\text{-}\gamma)^{\theta V}$
- Use simulations to ask...
  - Is there an θ such that $F_{\min_i P_i}(\gamma)$ behaves like the minimum of θV 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

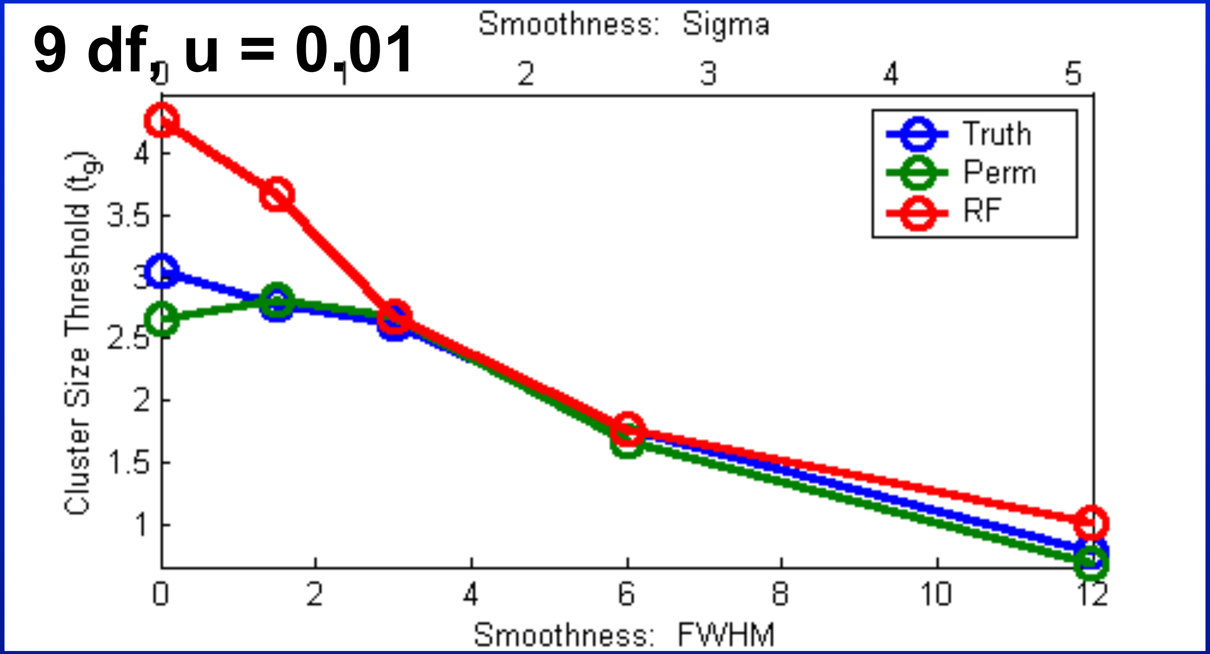


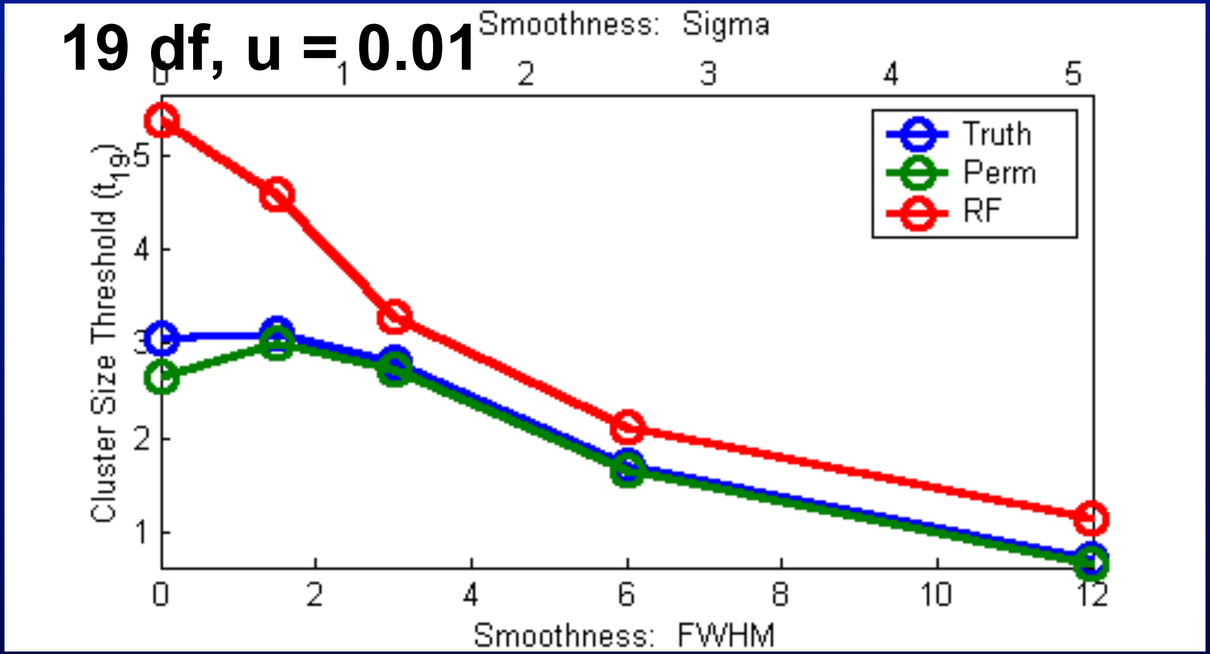
Monte Carlo Evaluations

- 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.

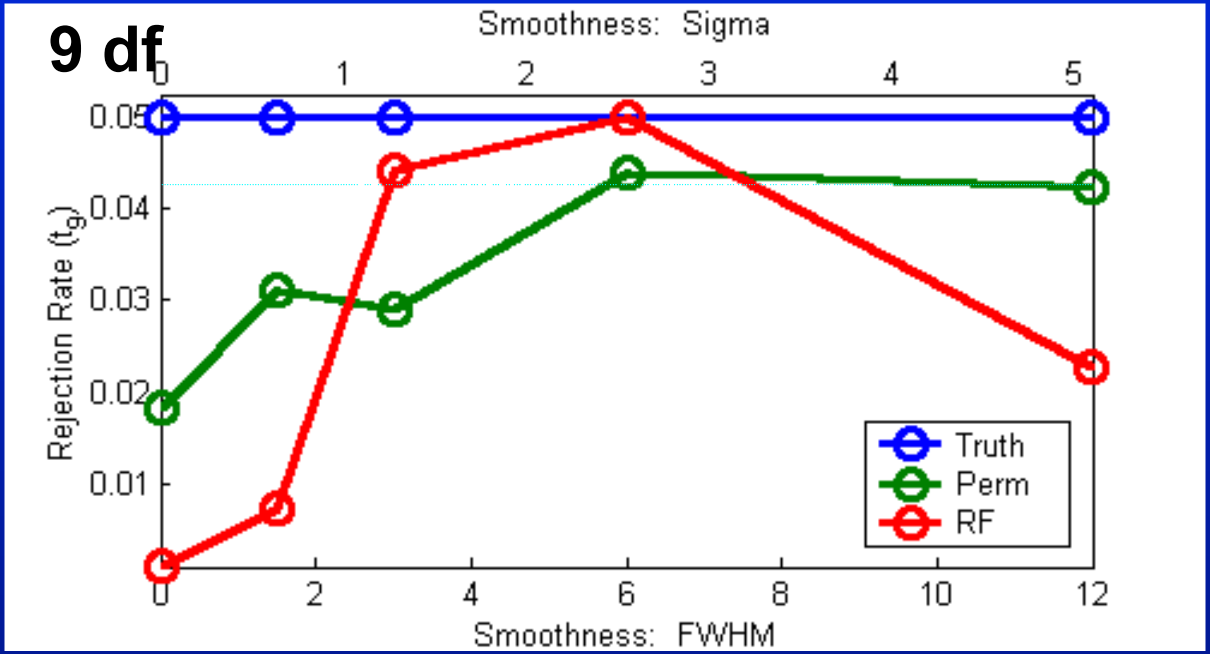


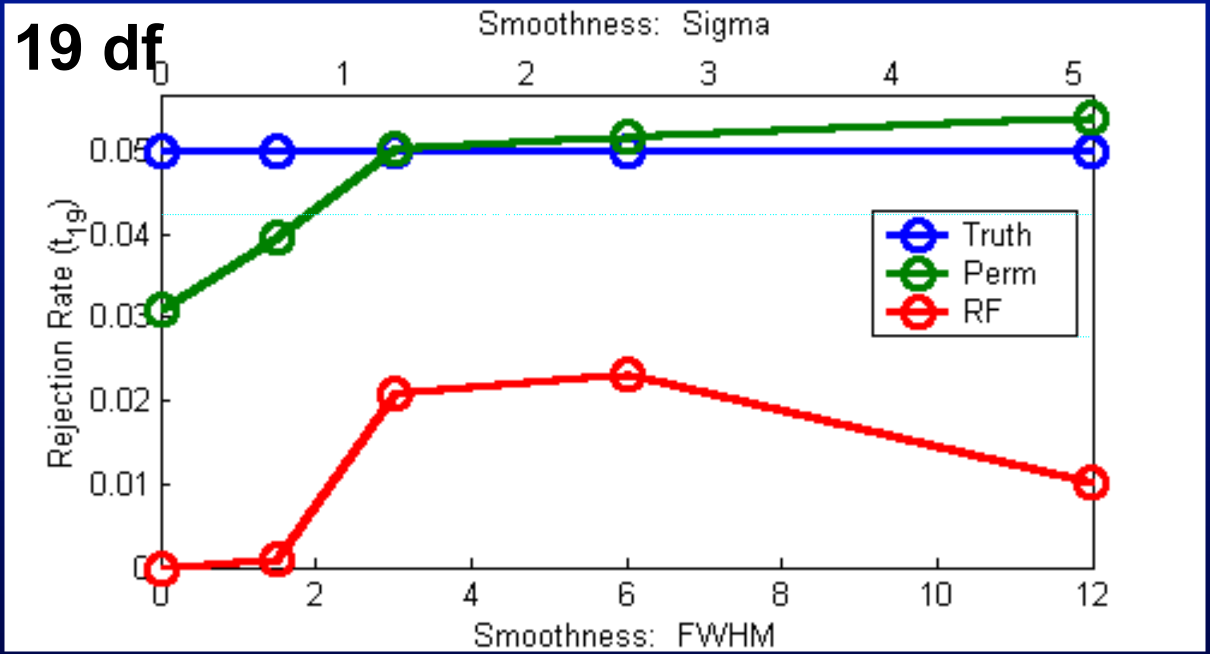


SKIP

Familywise Rejection Rates

- Interesting that gets worse with larger df.

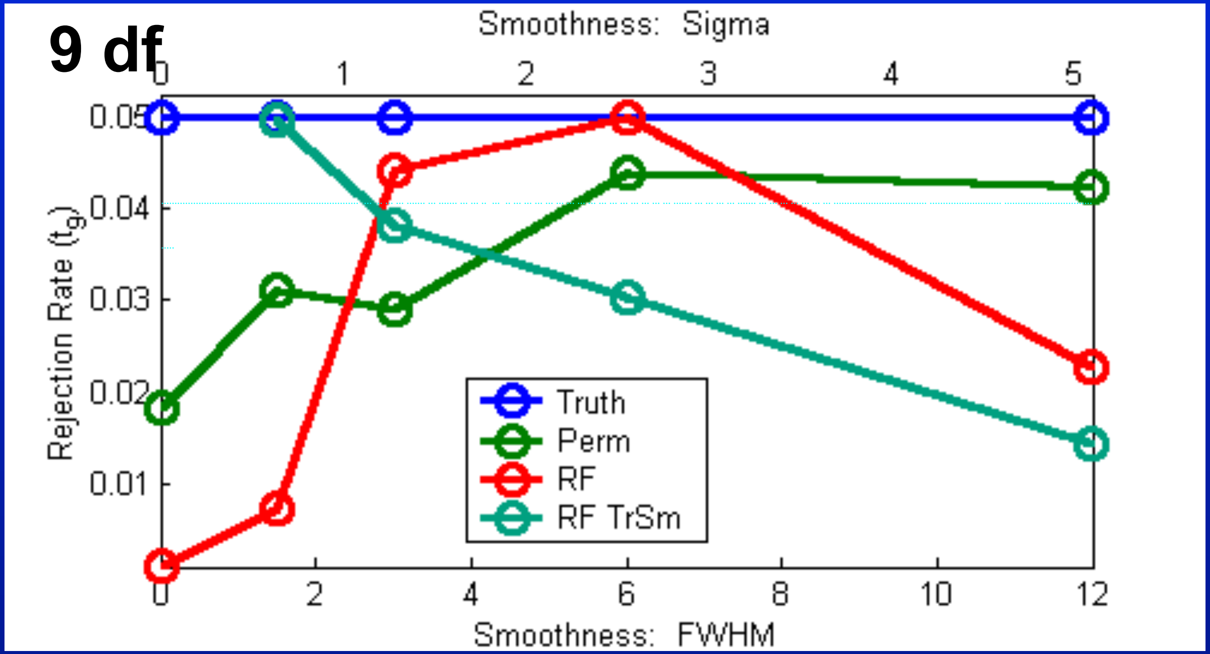


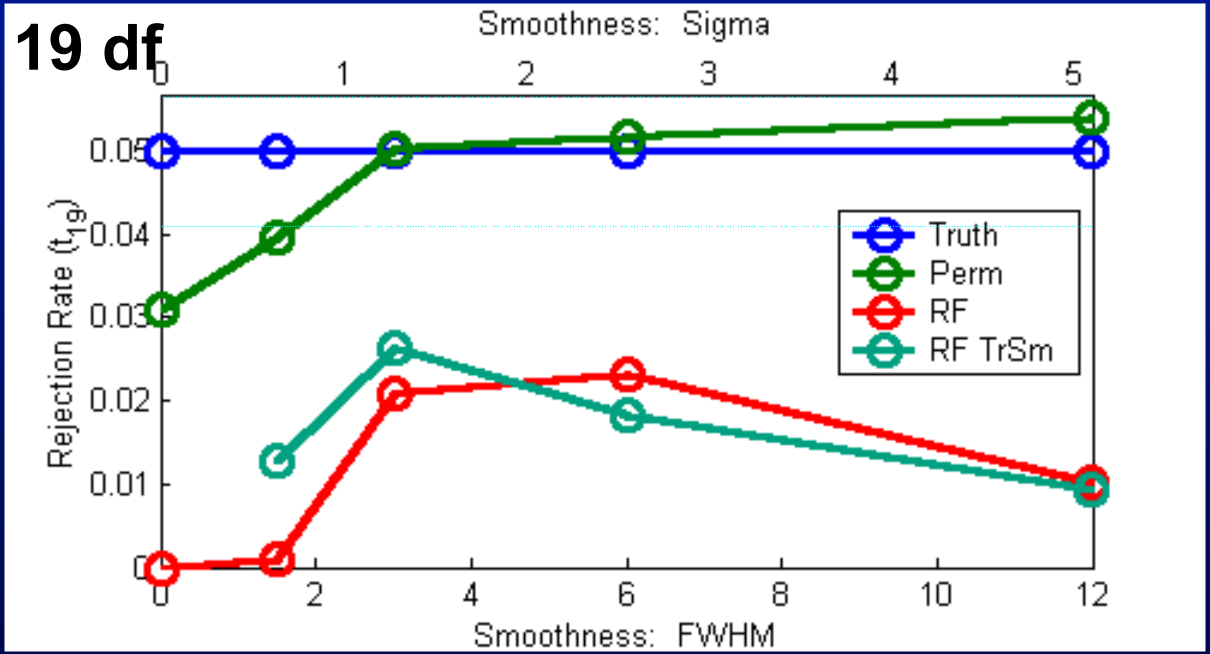


more

FWE Corrected p-values

- For df=9 biased smoothness estimation improves rejection rate





cont

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 α_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

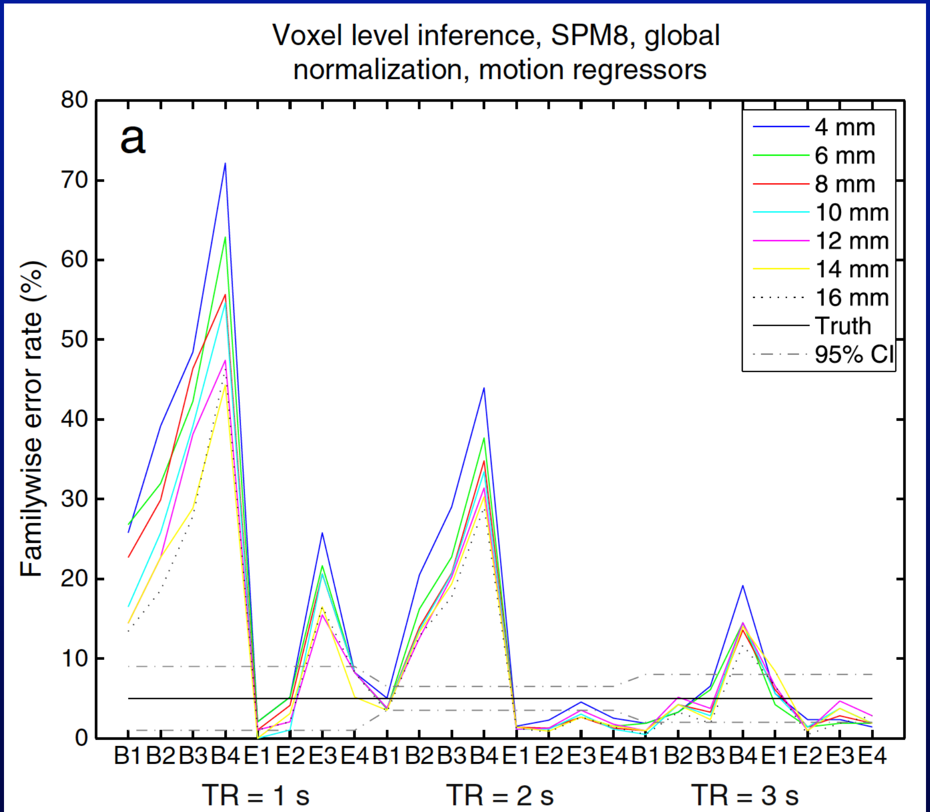
- Eklund (2012) analyzed **1,484** resting fMRI datasets from public repositories
- Fed through standard SPM pipeline, with 8 different "pretend" paradigms

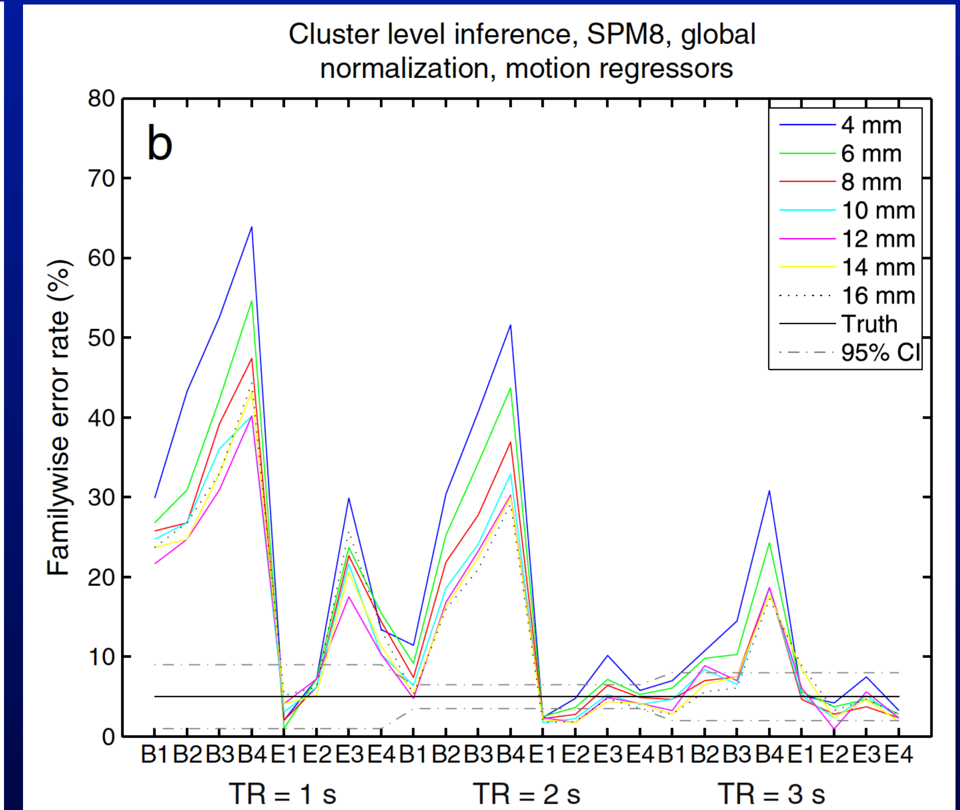
Paradigm	Activity periods (s)	Rest periods (s)
B1	10	10
B2	15	15
B3	20	20
B4	30	30
E1	2	6
E2	4	8
E3	1–4 (R)	3–6 (R)
E4	3–6 (R)	4–8 (R)

Eklund et al. (2012). Does parametric fMRI analysis with SPM yield valid results? An empirical study of 1484 rest datasets. *NeuroImage*, 61(3), 565–78.

Computed Familywise Error (FWE) Rates

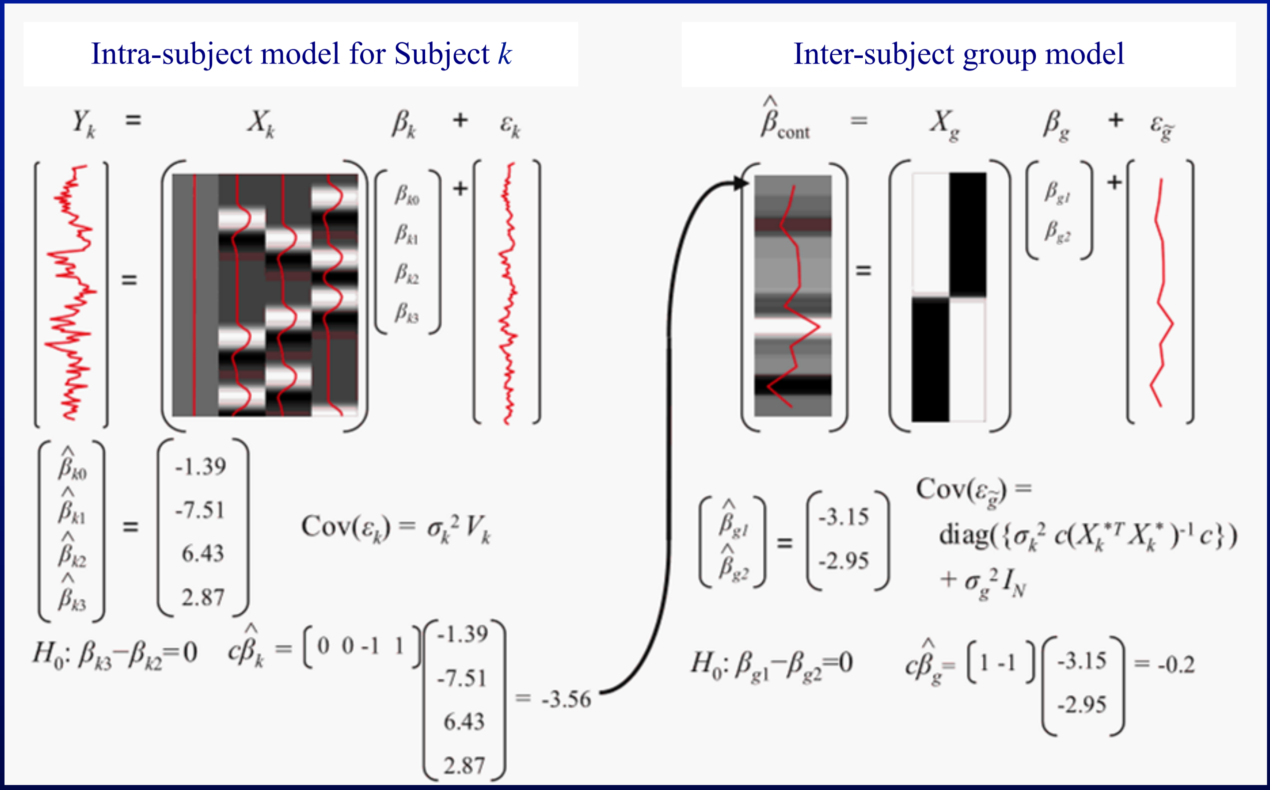
- Many settings had awful FWE!
  - Block worse than event; fast TR worse that slow





Massive Empirical Evaluation – Take II

- Previous result only for first level fMRI
- 2^{nd} level fMRI doesn't depend on 1^{st} level P-values
- Data quality also an issue



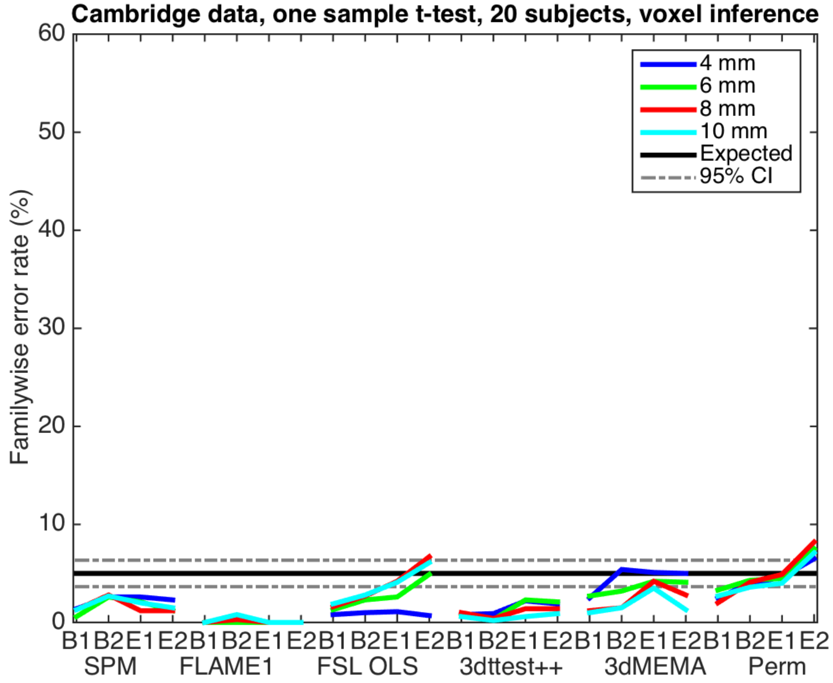
Massive Empirical Evaluation – Take II

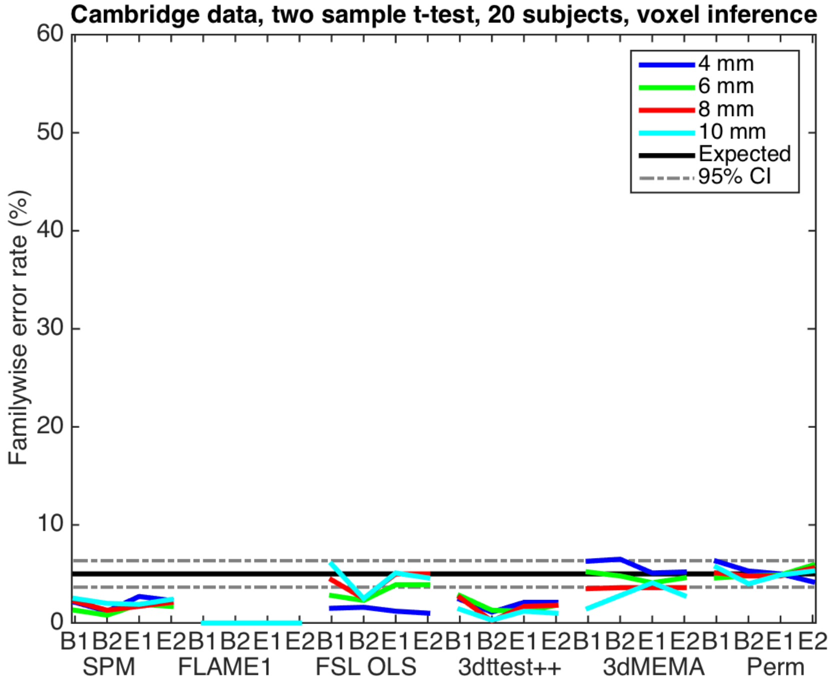
- Same fcon1000 repository, just 2 largest sites: Beijing & Cambridge
- Second level analyses
  - 1-sample t-test: n = 20, 40
  - 2-sample t-test: $n_1 = n_2 = 10$, 20

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

Massive Group fMRI Evaluation
Voxel-wise

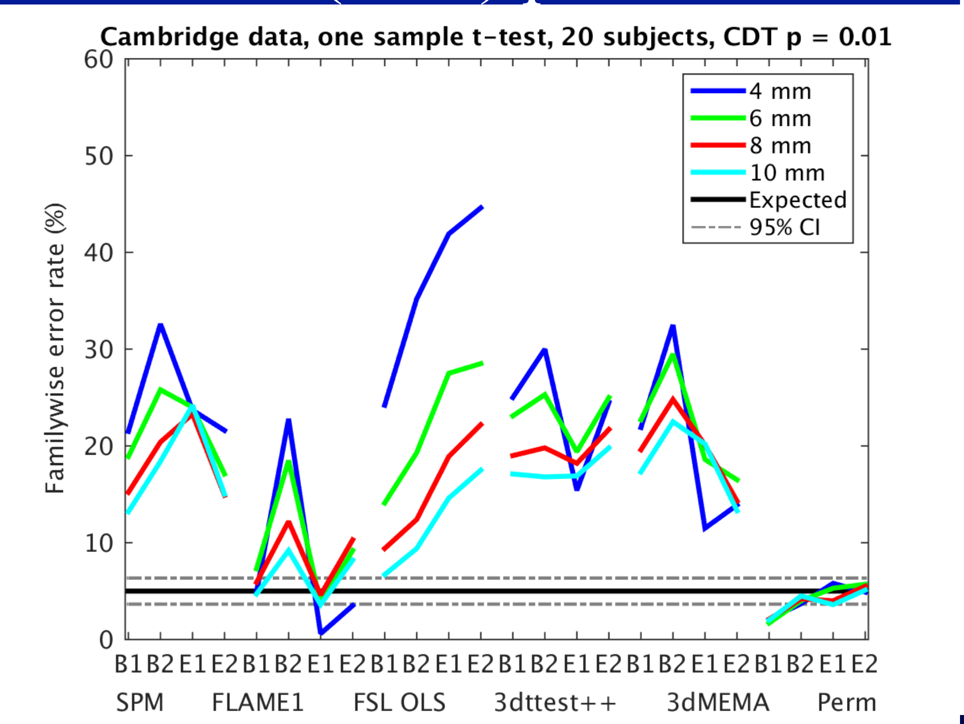
- Voxel-wise inference OK
  - Sometimes very conservative!

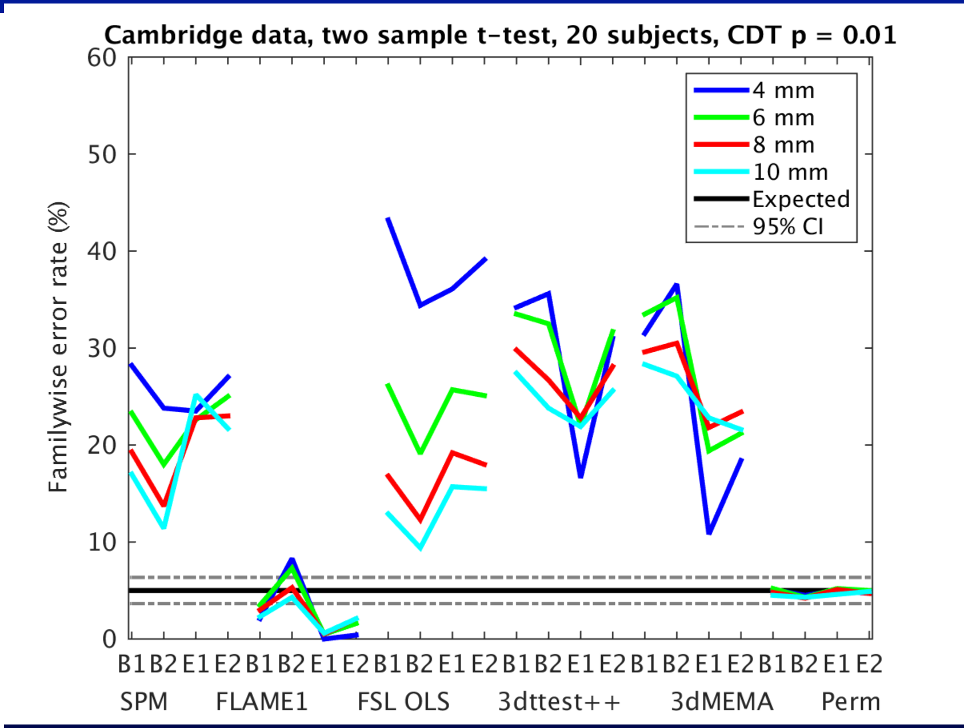




Massive Group fMRI Evaluation Cluster-wise CFT p=0.01

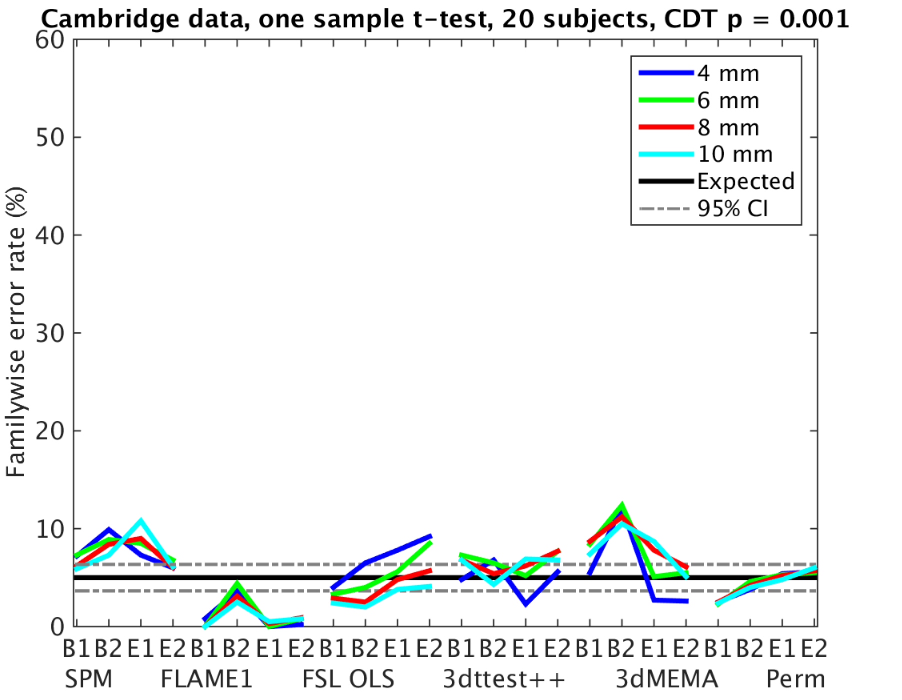
- Cluster-wise a catastrophe!
  - Rarely valid at cluster forming threshold (CFT) p=0.01 – default CFT in FSL

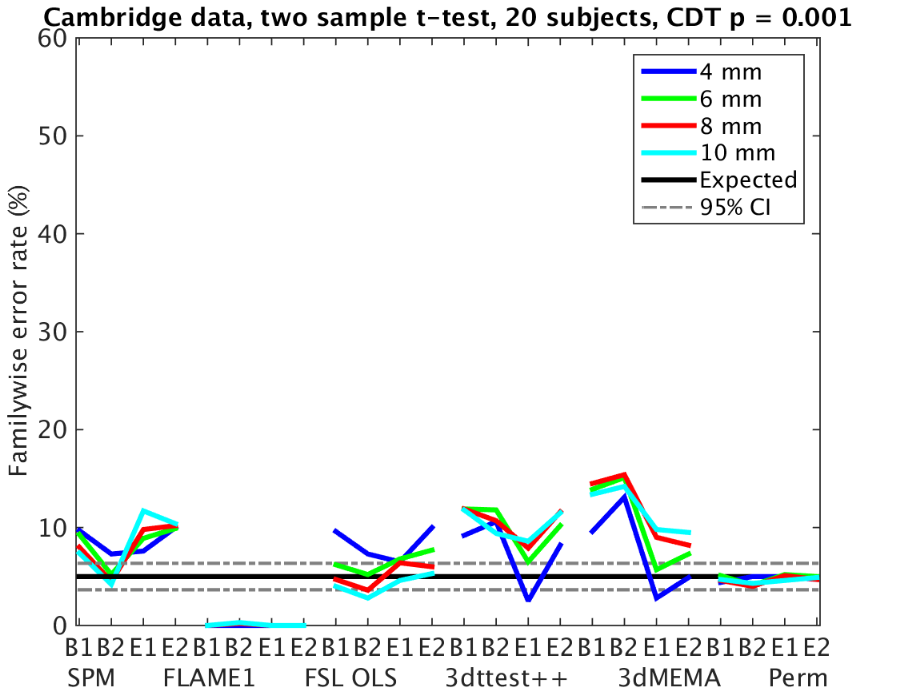




Massive Group fMRI Evaluation
Cluster-wise CFT p=0.001

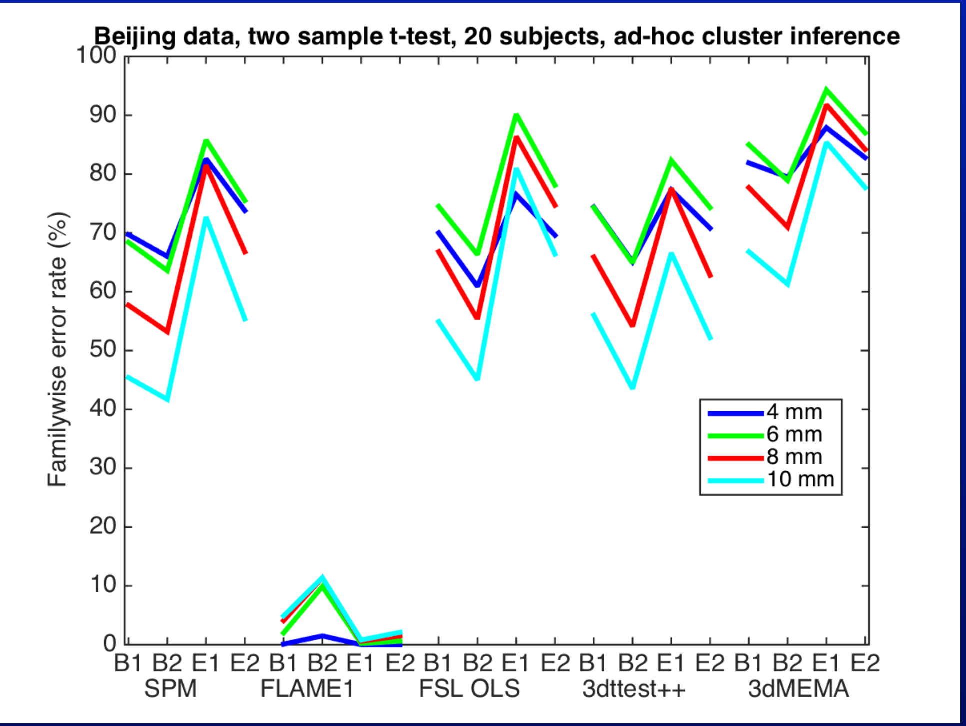
- Cluster-wise CFT p=0.001 better
  - Valid ≈ 50% time, depending on design





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

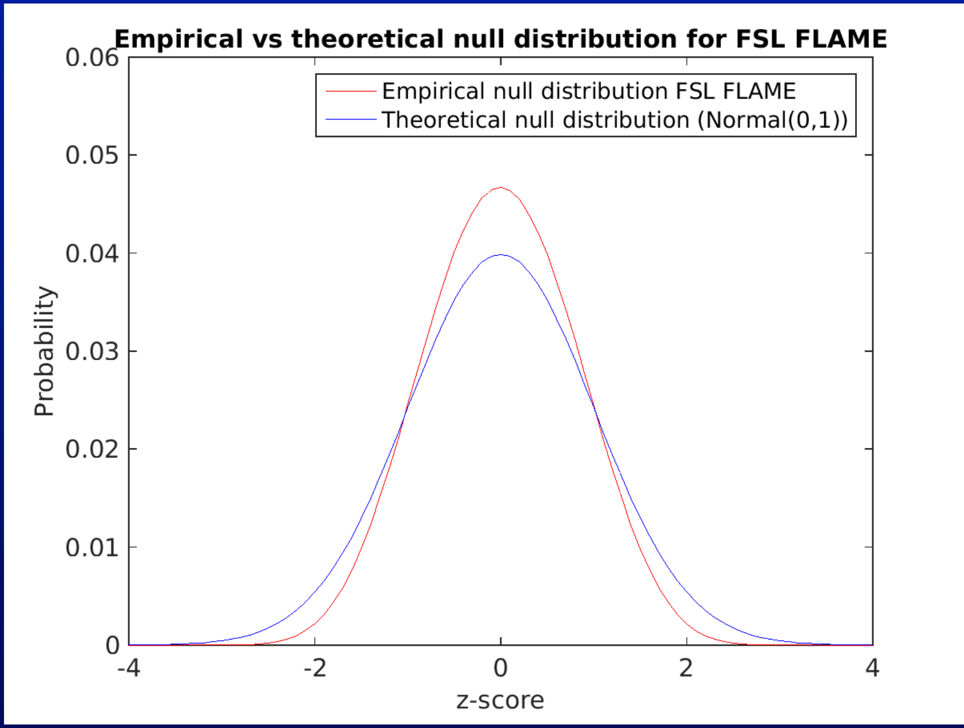


Massive Group fMRI Eval: What's going wrong?

- 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

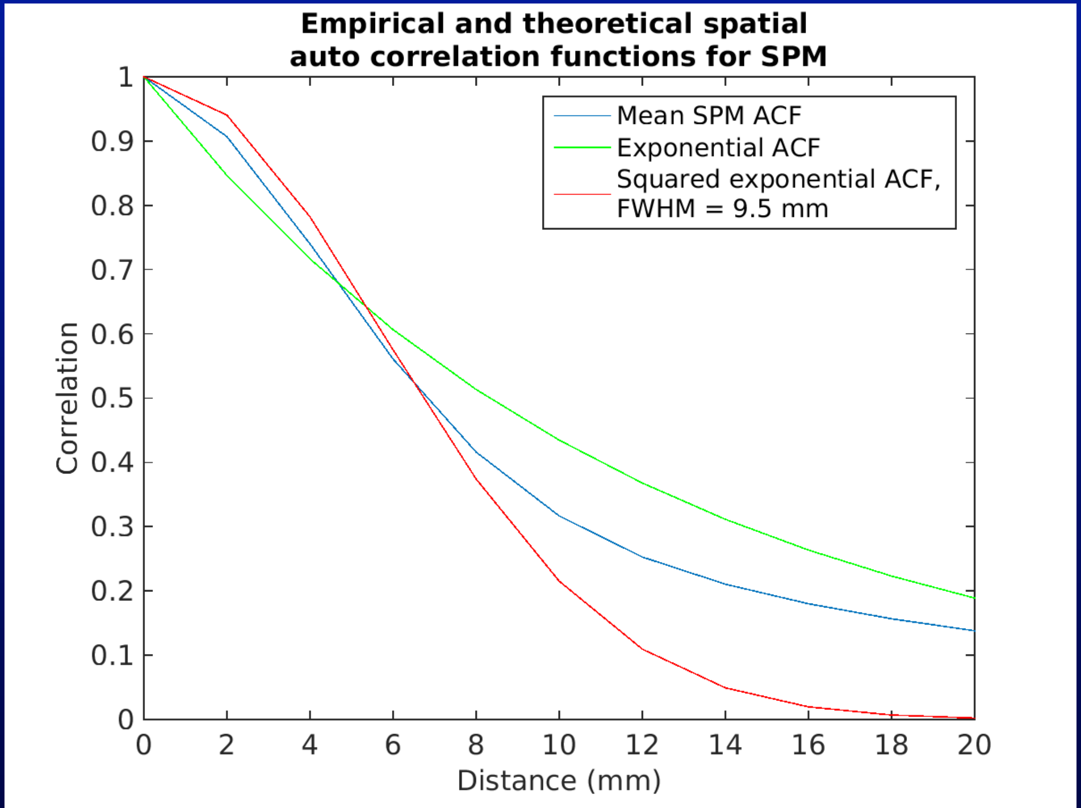
- Univariate P-values are conservative
  - Nothing to do with RFT!
  - Turns out to be artifact of using completely null data ($\sigma_{BTW} = 0$)
  - Using non-null ($\sigma_{BTW}>0$) data for same-vs-same 2-group comparison, resolves this.



But then results similar (not shown) to FSL OLS ☹

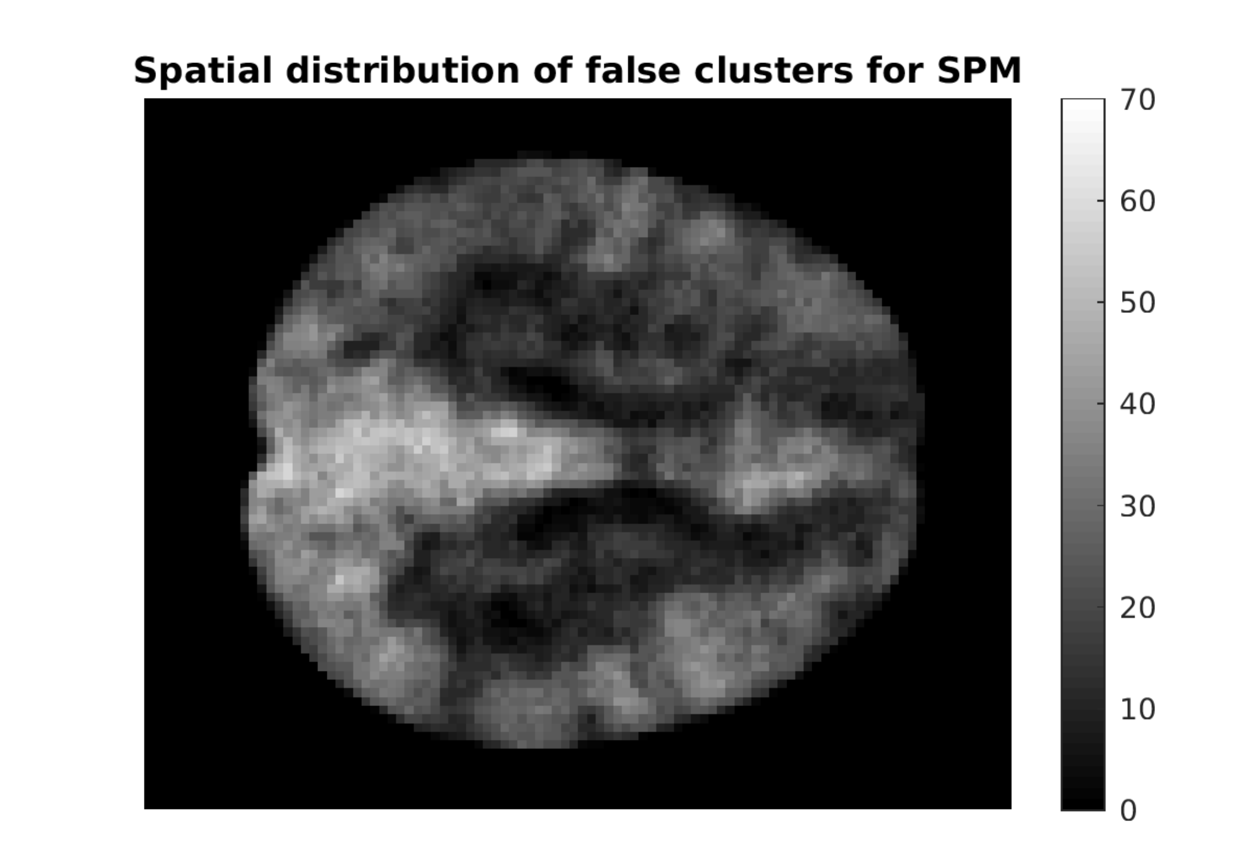
Massive Group fMRI Eval: Spatial ACF

- Much heavier tails than Gaussian pdf!



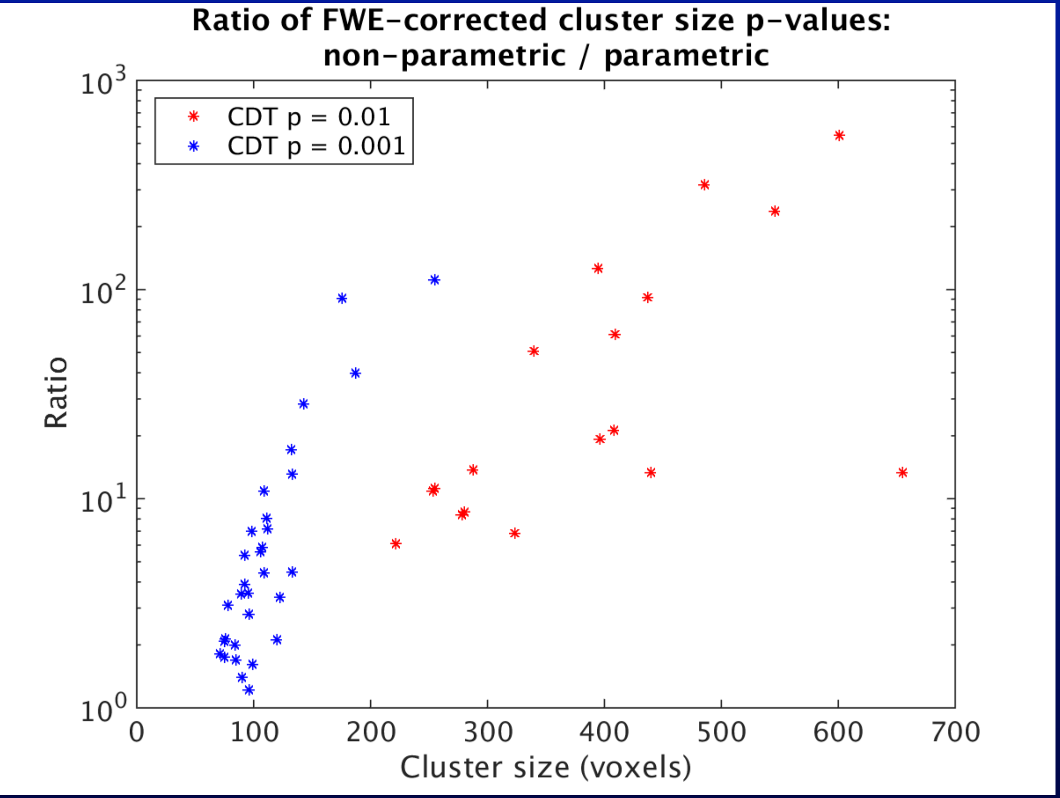
Massive Group fMRI Eval: Spatial Dist^n of False Clusters

- 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 signficance"

Conclusions

- Gaussian Monte Carlo results show…
  - *t* random field results conservative for
    - Low df & smoothness
    - 9 df & ≤12 FWHM;   19 df & < 10 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 avaiable!
  - FSL: randomise   SPM: SnPM toolbox

References

- TE Nichols and AP Holmes.
  Nonparametric Permutation Tests for Functional Neuroimaging: A Primer with Examples.
  Human Brain Mapping, 15:1-25, 2002.
- Eklund, A., Nichols, T., & Knutsson, H. (2015). Can parametric statistical methods be trusted for fMRI based group studies? http://arxiv.org/abs/1511.01863