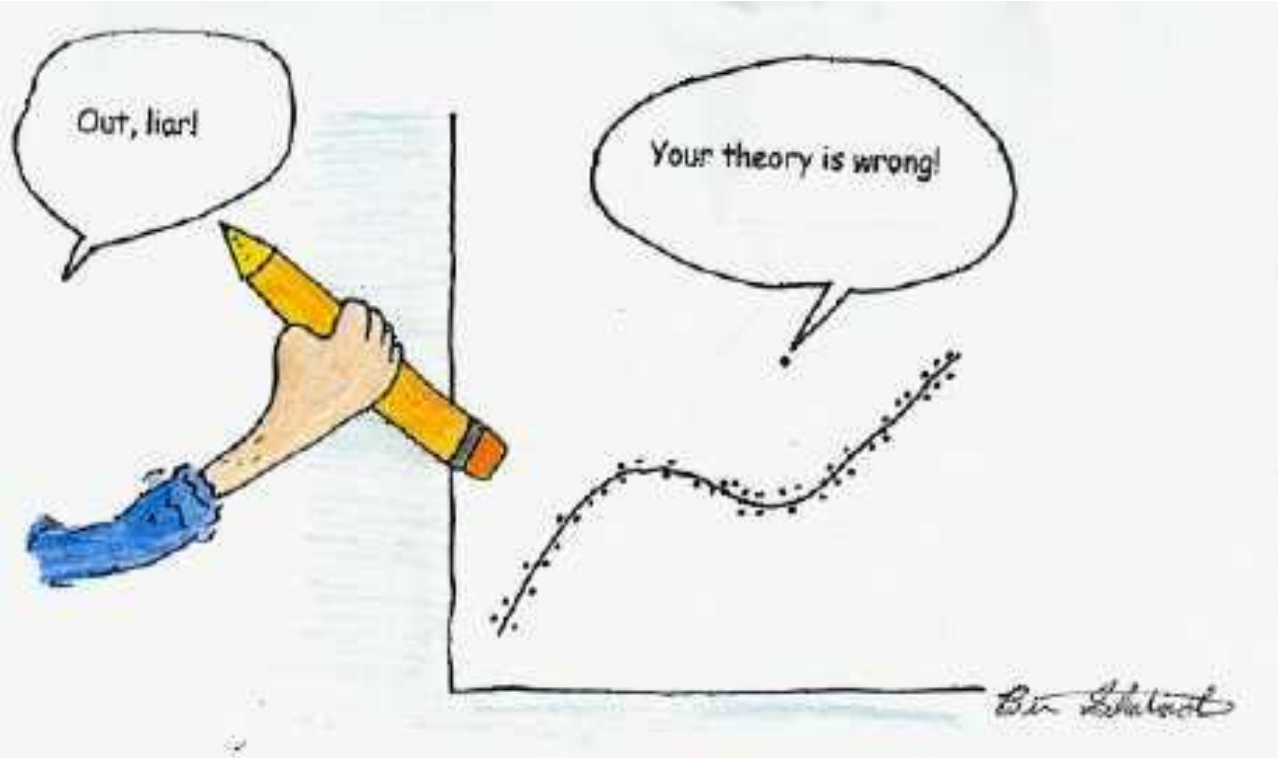
6th of April, 2016



Are your analyses too parametric?

That's not Normal!

Martin M Monti, Ph.D.
UCLA Department of Psychology
http://montilab.psych.ucla.edu
monti@psych.ucla.edu

Why IS your analysis parametric?

i. **Optimal power** (defined as the probability to detect a real difference) – when assumptions are met. Particularly important in neuroimaging:
   - Low SNR
   - Low df (data acquisition is expensive and time intensive)
   - Standard massive-univariate approach requires correction for multiple comparison, reducing sensitivity further

Why IS your analysis parametric?

i. **Optimal power** (defined as the probability to detect a real difference) – when assumptions are met. Particularly important in neuroimaging:
ii. **Computationally simple** – very important considering it is computed over more than 100,000 voxels
iii. **Flexible framework** – allows looking at multiple factors simultaneously and/or factoring out influence of variables of non-interest (think of the GLM approach)
iv. **Graceful failure (for 1 sample t-tests)** – when assumptions are not met it becomes more conservative

Why your analysis should NOT be parametric …

In parametric analyses we are making many assumptions concerning the distribution of the data which are not always met.

Violations

- Identically distributed:
  - Outliers can influence data in unexpected ways, even for large samples.
- Independence:
  - p-values too liberal; false positives; nominal degrees of freedom is overestimate.
- Normality:
  - p-values are wrong, no simple rule for determining in what way.
- Equal variance:
  - p-values too liberal; false positives; nominal degrees of freedom is overestimate.

Slide from M Lindquist

PART I:
IS YOUR ROI ANALYSIS TOO PARAMETRIC?

In parametric analyses we are making many assumptions concerning the distribution of the data which are not always met.

Violations

- Identically distributed:
  - Outliers can influence data in unexpected ways, even for large samples.
- Independence:
  - p-values too liberal; false positives; nominal degrees of freedom is overestimate.
- Normality:
  - p-values are wrong, no simple rule for determining in what way.
- Equal variance:
  - p-values too liberal; false positives; nominal degrees of freedom is overestimate.

Slide from M Lindquist

PART I:
IS YOUR ROI ANALYSIS TOO PARAMETRIC?

frontiers in
HUMAN NEUROSCIENCE

PERSPECTIVE ARTICLE
published: 03 May 2012
doi: 10.3389/fnhum.2012.00119

Improving standards in brain-behavior correlation analyses

Guillaume A. Rousselet[1]* and Cyril R. Pernet[2]

[1] *Centre for Cognitive Neuroimaging (CCNi), Institute of Neuroscience and Psychology, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, UK*
[2] *Brain Research Imaging Center, Division of Clinical Neurosciences, University of Edinburgh, Western General Hospital, Edinburgh, UK*

Edited by:
Russell A. Poldrack, University of Texas, USA

Reviewed by:
Martin M. Monti, University of California, Los Angeles, USA
Tal Yarkoni, University of Colorado at Boulder, USA

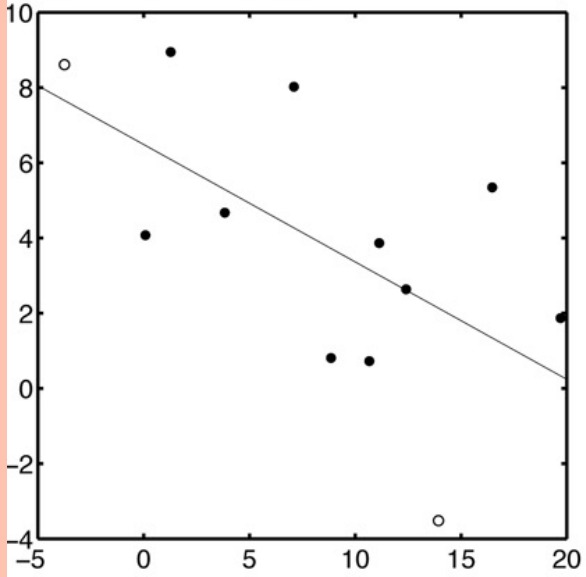
Associations between two variables, for instance between brain and behavioral measurements, are often studied using correlations, and in particular Pearson correlation. However, Pearson correlation is not robust: outliers can introduce false correlations or mask existing ones. These problems are exacerbated in brain imaging by a widespread lack of control for multiple comparisons, and several issues with data interpretations. We illustrate these important problems associated with brain-behavior correlations, drawing examples from published articles. We make several propositions to alleviate these problems.

Keywords: Pearson correlation, Spearman correlation, skipped correlation, outliers, robust statistics, multiple comparisons, multivariate statistics, confidence intervals

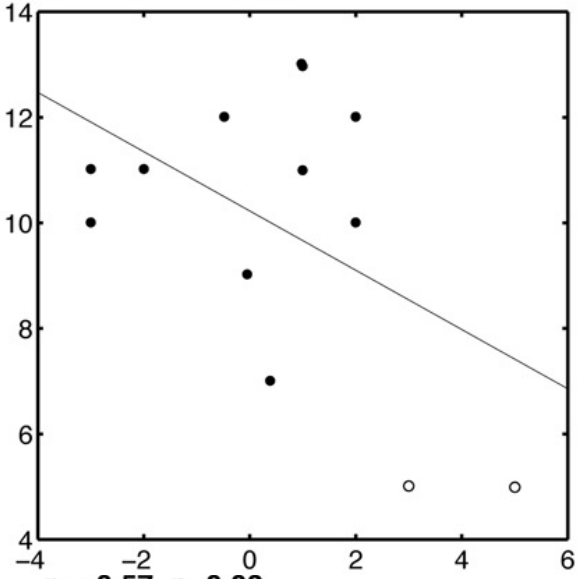
Brain-behavior correlations

- ***Pearson correlation***:
  - Most widely used
  - Non-robust estimator, particularly sensitive to outliers (and magnitude of the slope around which points are clustered, magnitude of the residuals, heteroscedasticity).
  - Outliers can affect correlations both ways:
    - *False positive problem*: create the impression of an association greater than zero where there is, in fact, none
    - *Power problem*: mask the presence of a significant effect
- Alternatives:
  - ***Spearman*** – calculates the Pearson correlation on the rank of the data; less sensitive to marginal (univariate) outliers
  - ***(Wilcox) Skipped correlations*** – calculates the Spearman correlation ***after*** having performed multivariate robust outlier detection (and removal)

r=-0.614, p=0.034
r=-0.60, p=0.040
r_s=-0.57, p=0.055
r_p=-0.42, t=1.46, t_{crit}=2.90

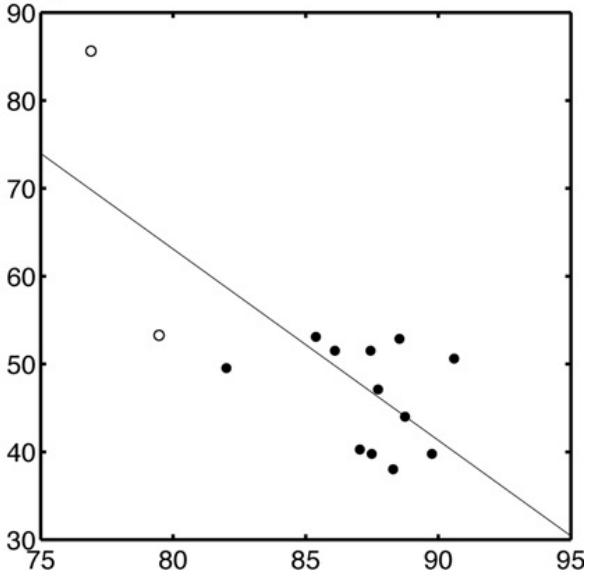


r=0.46, p=0.09
r=–0.48, p=0.10
r_s=–0.29, p=0.33
r_p=0.18, t=0.60, t_{crit}=2.85

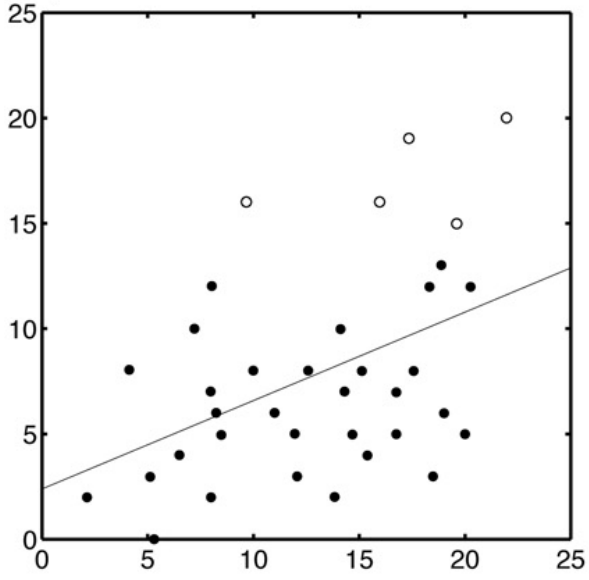


Published papers: Outlier driven correlations

r_s=-0.57, p=0.03
r=-0.74, p=0.003;
r_s=-0.58, p=0.03
r_p=-0.33, t=1.20, t_{crit}=2.82

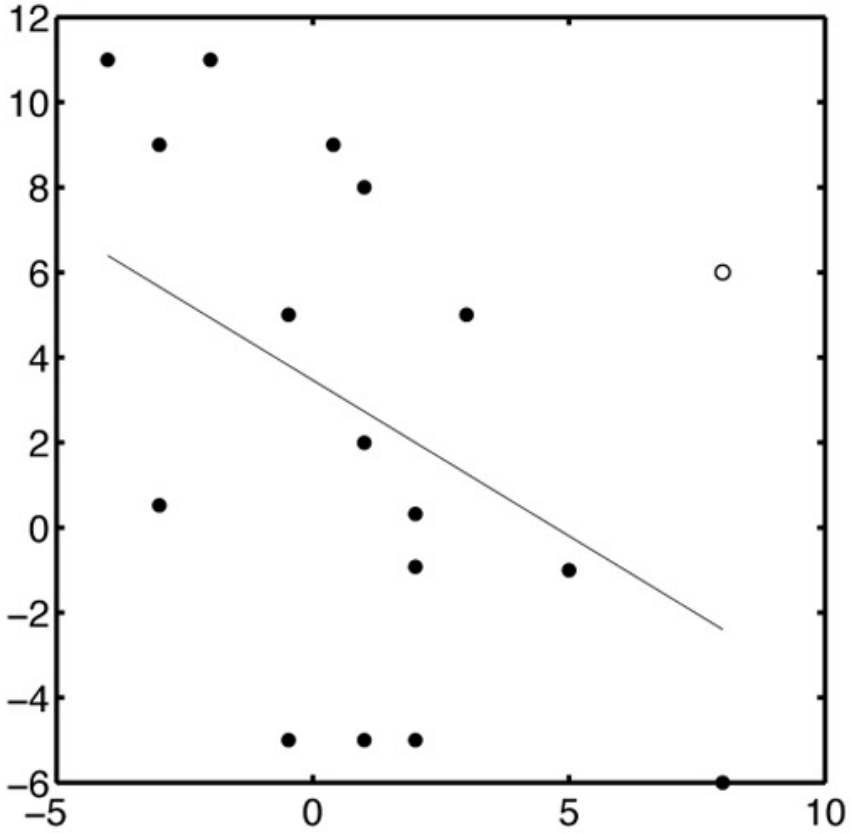


r=0.418, p=0.011
r=0.449, p=0.006;
r_s=0.38, p=0.022
r_p=0.27, t=1.63, t_{crit}=2.51

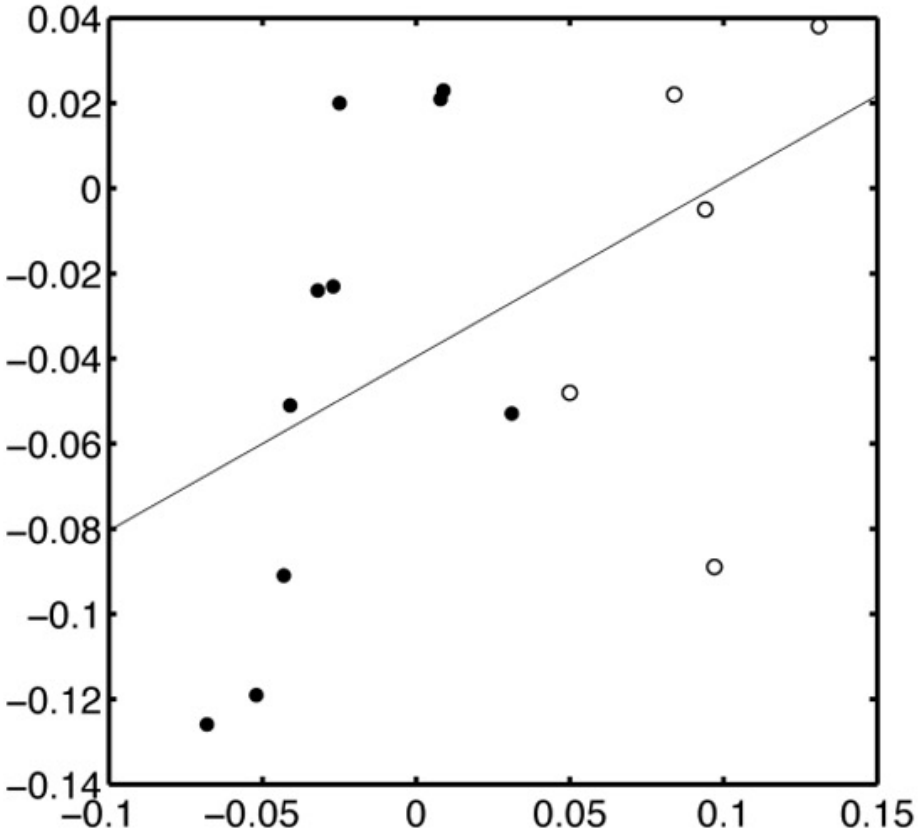


Published papers: Partial masking

r=-0.40, p<0.05
r=-0.43, p=0.086
r_s=-0.52, p=0.03
r_p=-0.63, t=3.10, t_{crit}=2.73



r=0.48, p<0.05 (one-tailed)
r=0.48, p=0.07
r_s=0.59, p=0.023
r_p=0.75, t=4.03, t_{crit}=2.78



PART II:
IS YOUR GROUP ANALYSIS TOO PARAMETRIC?

In parametric analyses we are making many assumptions concerning the distribution of the data which are not always met.

Violations

- Identically distributed:
  - Outliers can influence data in unexpected ways, even for large samples.
- Independence:
  - p-values too liberal; false positives; nominal degrees of freedom is overestimate.
- Normality:
  - p-values are wrong, no simple rule for determining in what way
- Equal variance:
  - p-values too liberal; false positives; nominal degrees of freedom is overestimate.

Slide from M Lindquist

Moving parts (decision points)

- Group level model (e.g., FFX, RFX, MFX)
- Outlier management
- Thresholding method & correction for multiple comparisons (e.g., cluster threshold, voxel, parametric, non-parametric)

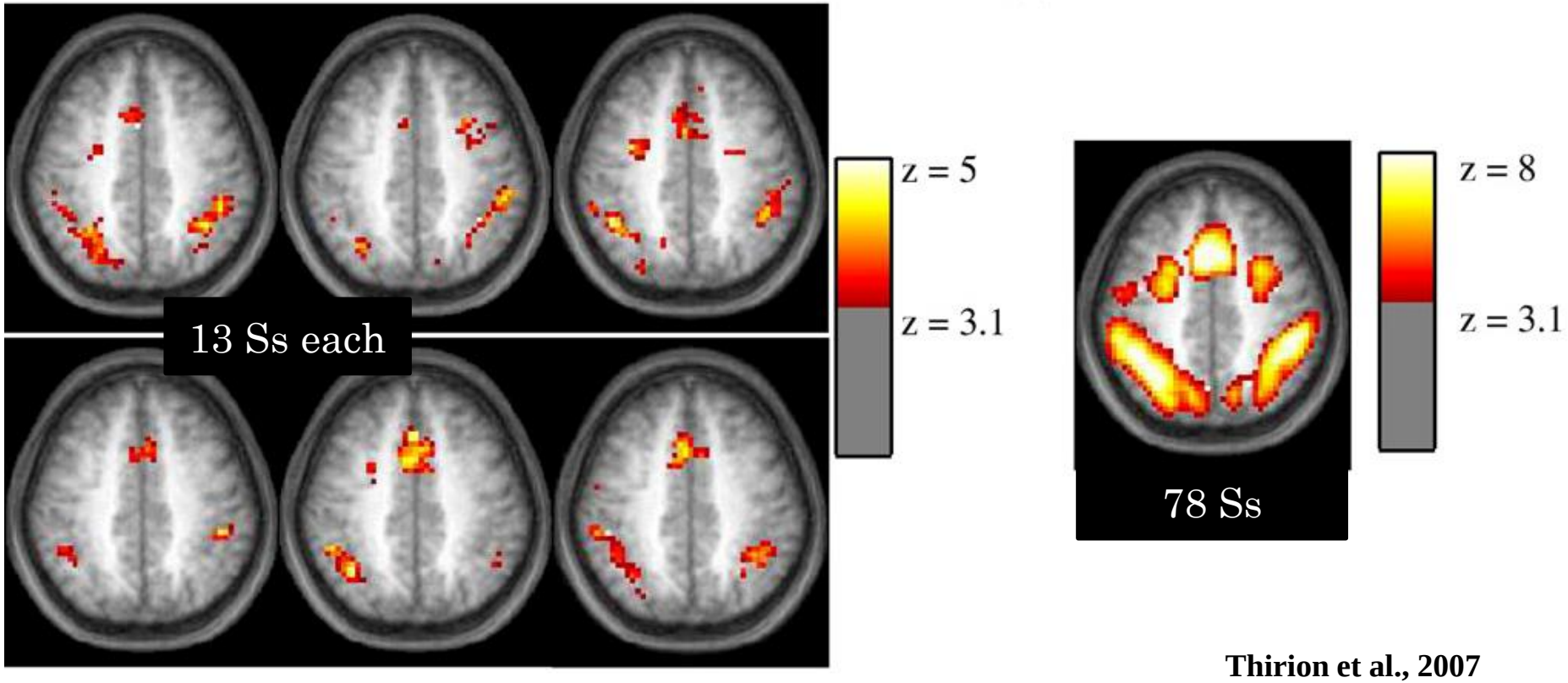
Several possible sources of heteroschedastic variance

- In fMRI, there is sizeable inter-subject variance because of several factors:
  i. **Spatial mismatch** between subjects' cortical structures (can be as large as 1cm!), which can yield a structured but variable pattern of noise
  ii. **Activation magnitude differences** (both across subjects and from session to session): physiological fluctuations, motion, baseline, instruction misunderstanding, …
  iii. Differences in elicitation of brain networks across subjects, due to **genetic/epigenetic differences** or different **cognitive strategies**
- All these factors end up being modeled as the variance term in group analysis (i.e., t-test denominator).

Thirion et al., 2007

THE PROBLEM IS:

(a)



Thirion et al., 2007

THE PROBLEM IS:

"We observed that […] the analysis of 6 different groups of 13 subjects would lead to different reports of the set of activated regions for the same experimental condition and standard threshold."

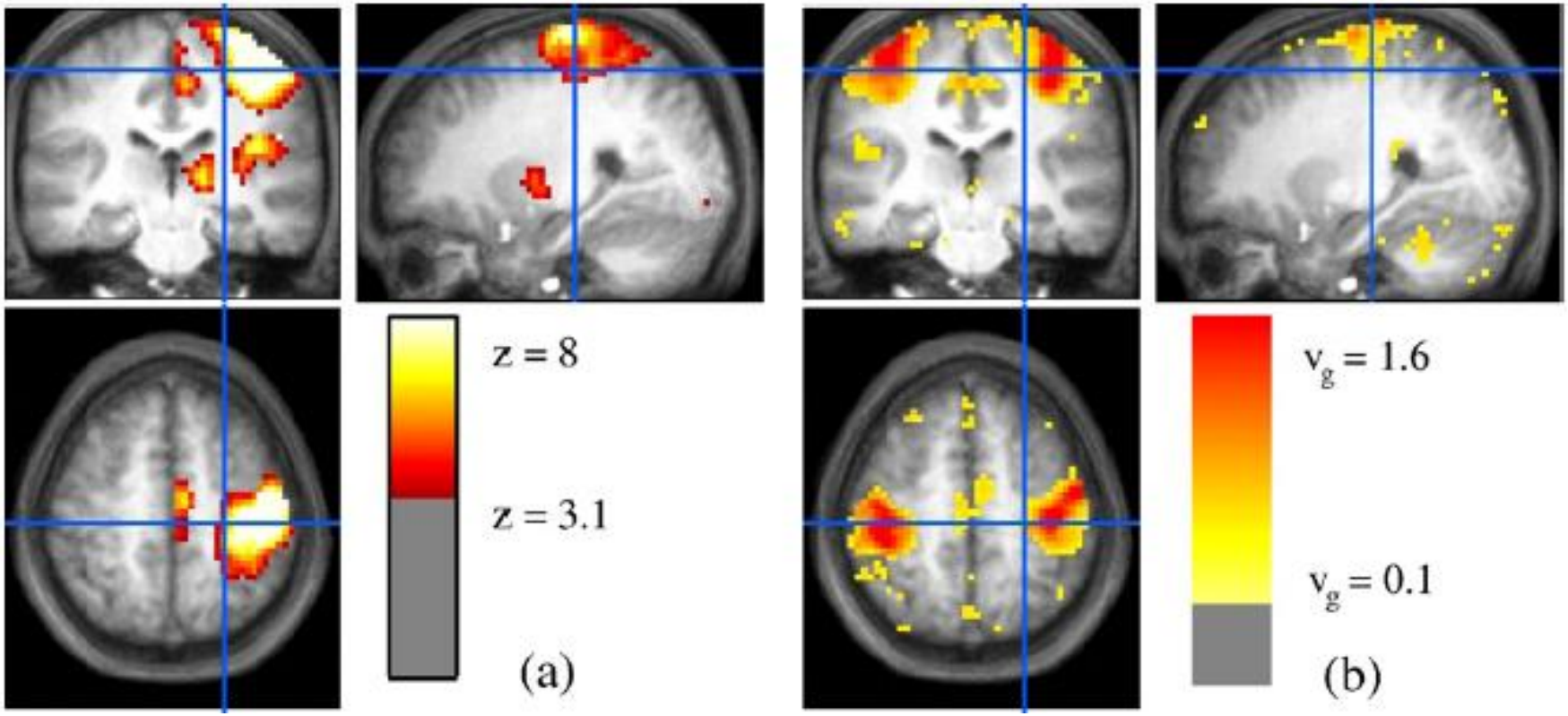
Thirion et al., 2007

N=81

Areas of high variance coincide with areas with significant effect size

Group-level activation map (p<0.001)

Group-level variance

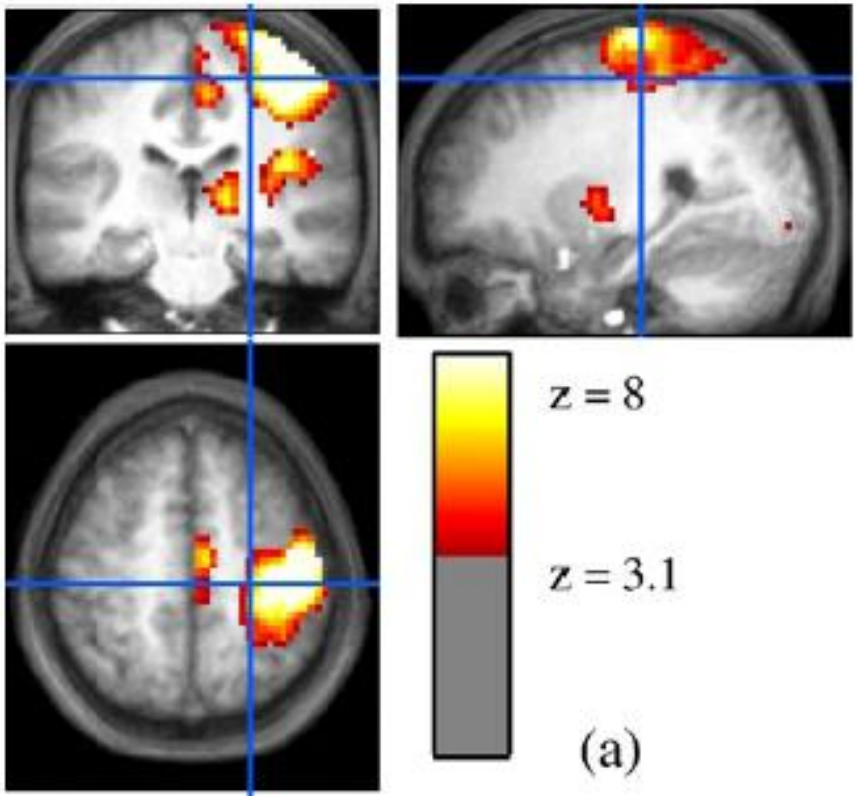


- The group effect ($\bar{\beta}(v)$) is not independent of the variance ($v_g(v)$), penalizing the statistic/sensitivity

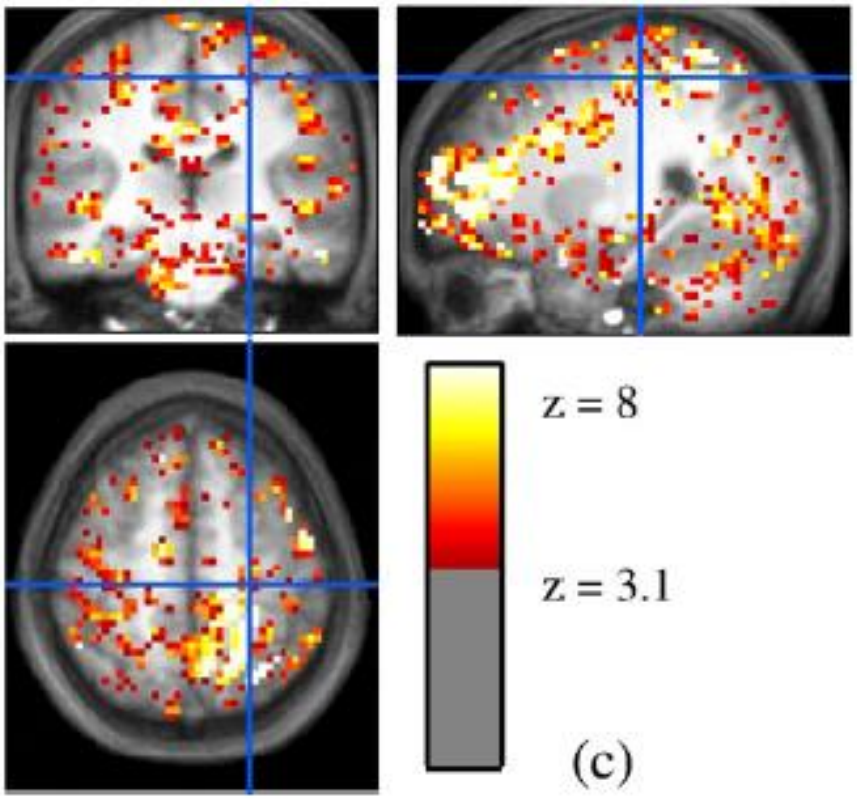
N=81

Large areas of non-normality of $\hat{\beta}$

Group-level activation map (p<0.001)



D'Agostino-Pearson normality test

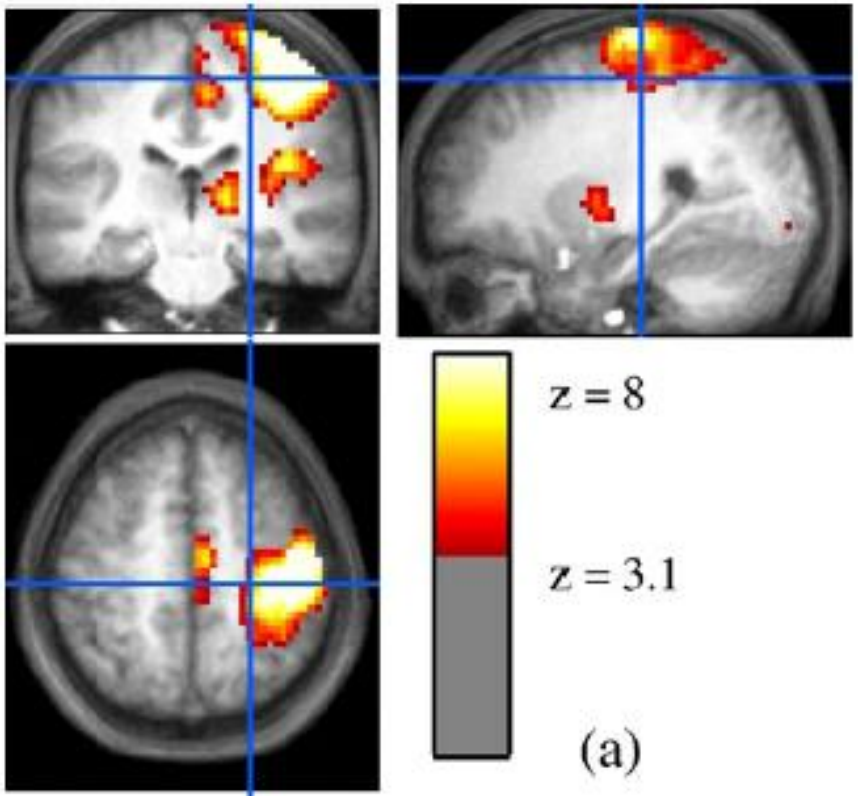


- Up to 30% of brain voxels fail the D'a-P test of normality for the effect $\hat{\beta}$

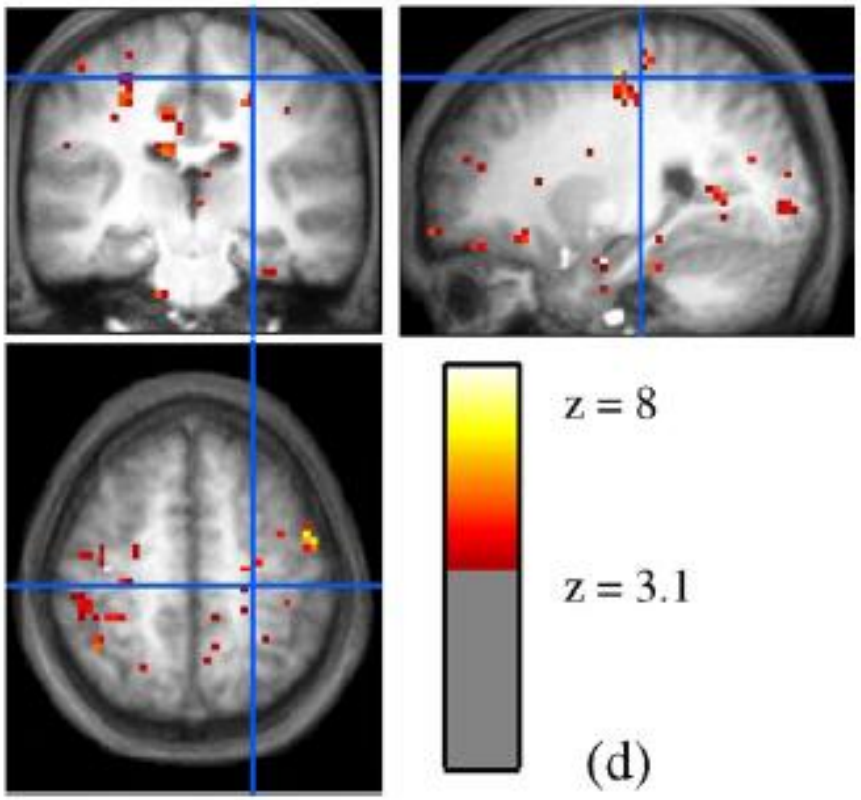
N=81

SMALLER AREAS OF NON-NORMALITY OF $\tau = \frac{\hat{\beta}}{\hat{\sigma}}$

Group-level activation map (p<0.001)



D'Agostino-Pearson normality test

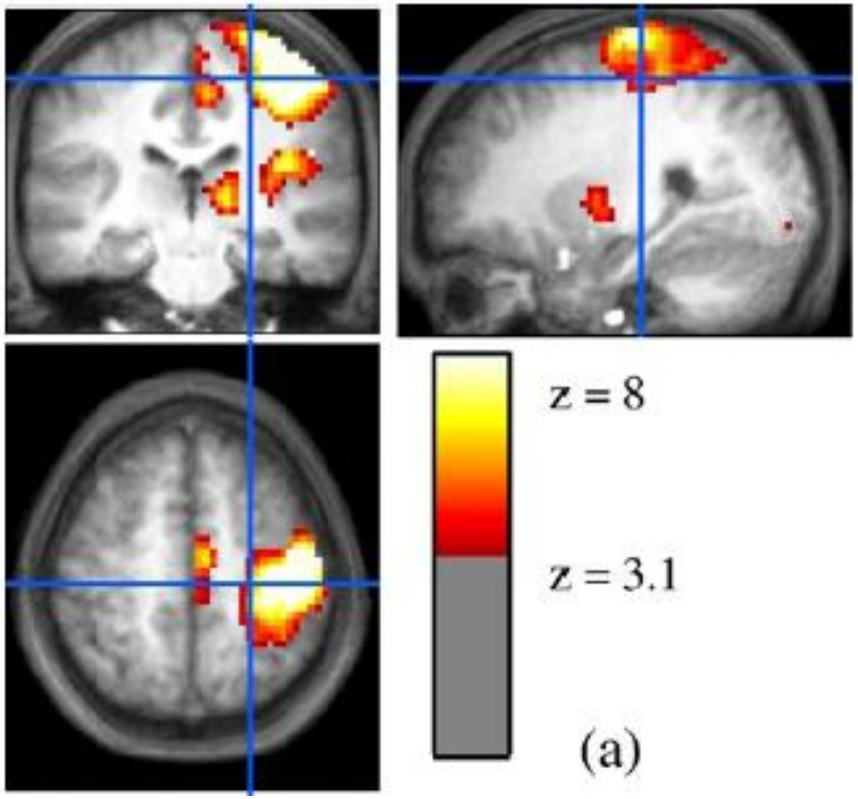


- Up to 10% of brain voxels fail the D'a-P test of normality for the normalized effect $\tau = \frac{\hat{\beta}}{\hat{\sigma}}$

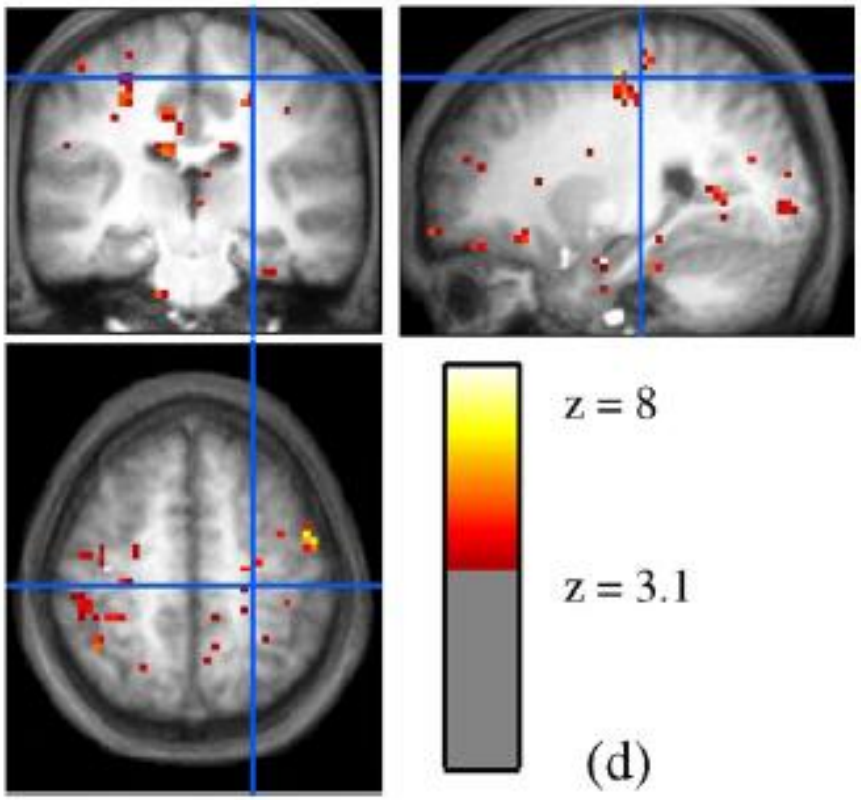
N=81

SMALLER AREAS OF NON-NORMALITY OF $\tau = \frac{\acute{\beta}}{\acute{\sigma}}$

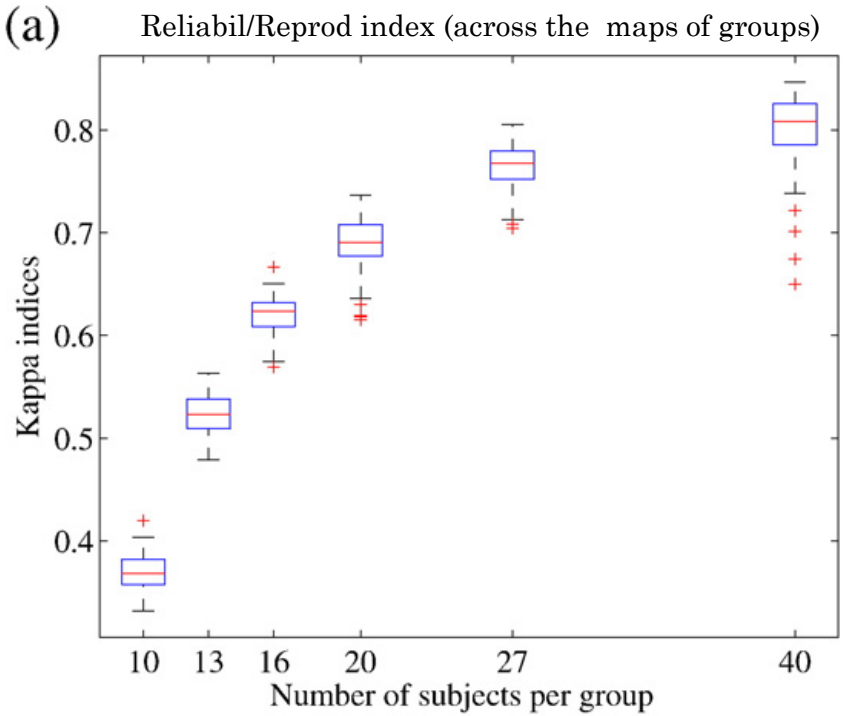
Group-level activation map (p<0.001)

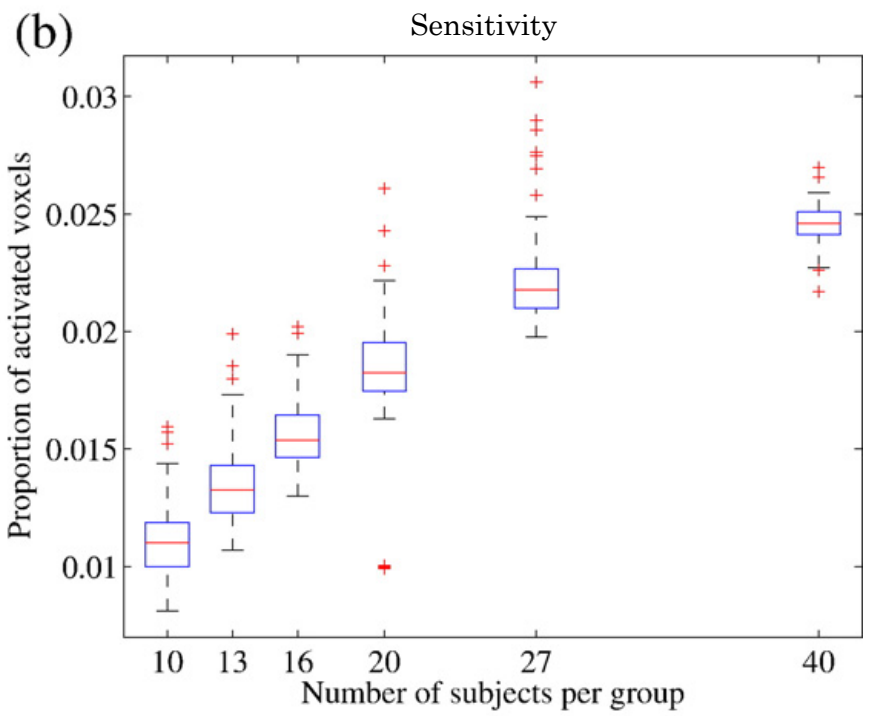


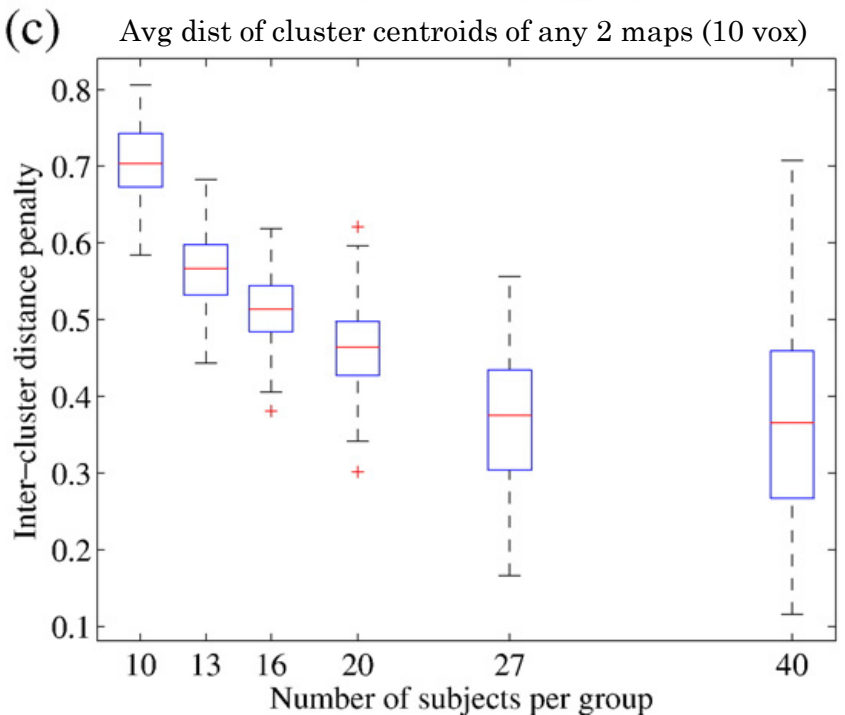
D'Agostino-Pearson normality test

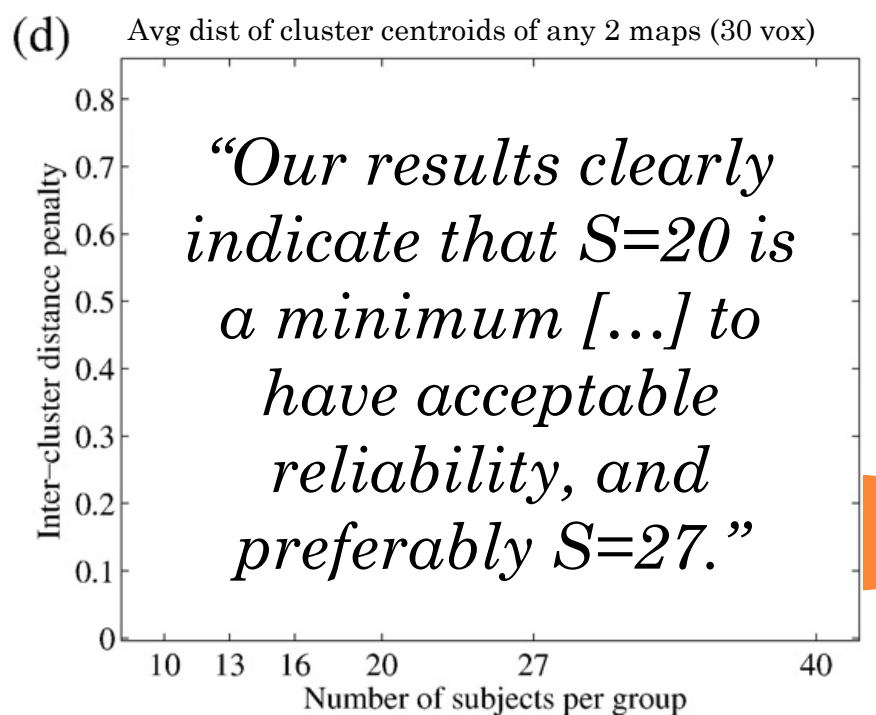


- Non-normality does not appear to co-localize with areas of activation



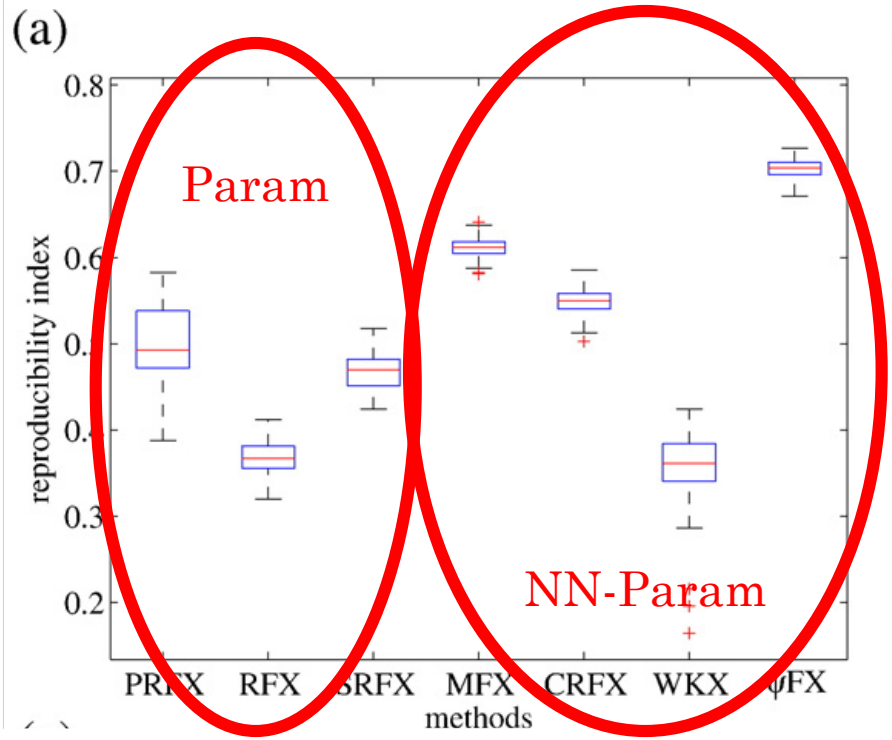


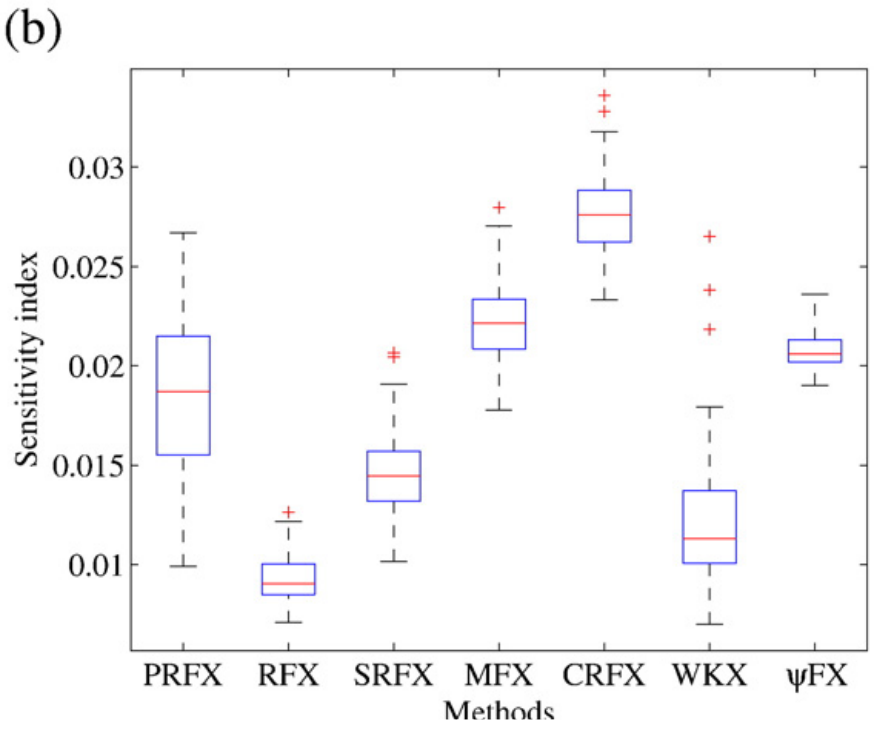




"Our results clearly indicate that S=20 is a minimum […] to have acceptable reliability, and preferably S=27."

How many subjects?





PRFX: Parcel (N=500) RFX
RFX: Random effects t-test (5mm FWHM), ign
SRFX: Random effects t-test (12mm FWHM), i
MFX: Mixed effects, with permutation testing
CRFX: Cluster-based RFX, with permutation t
WKX: Wilcoxon signed rank test
ψFX: Pseudo MFX (weighted average of the si

"In general, it is advisable to use non-parametric assessment to obtain reliable thresholds."

What group model?

TESTING OUR TOOLS

CAN PARAMETRIC STATISTICAL METHODS BE TRUSTED FOR FMRI BASED GROUP STUDIES?

Anders Eklund [a,b,c], *Thomas Nichols* [d], *Hans Knutsson* [a,c]

[a]Division of Medical Informatics, Department of Biomedical Engineering,
Linköping University, Linköping, Sweden
[b]Division of Statistics and Machine Learning, Department of Computer and Information Science,
Linköping University, Linköping, Sweden
[c]Center for Medical Image Science and Visualization (CMIV),
Linköping University, Linköping, Sweden
[d]Department of Statistics, University of Warwick, Coventry, United Kingdom

ABSTRACT

The most widely used task fMRI analyses use parametric methods that depend on a variety of assumptions. While individual aspects of these fMRI models have been evaluated, they have not been evaluated in a comprehensive manner with empirical data. In this work, a total of 2 million random task fMRI group analyses have been performed using

title or abstract). The first fMRI experiments consisted of simple motor tasks, while more recent examples involve resting state fMRI to study (dynamic) brain connectivity [3, 4]. Despite the popularity of fMRI as a tool for studying brain function, the statistical methods used have rarely been validated using real data, likely due to the high cost of fMRI data collection. Validations have instead mainly been performed

TESTING OUR TOOLS

CAN PARAMETRIC STATISTICAL METHODS BE TRUSTED FOR FMRI BASED GROUP STUDIES?

Anders Eklund [a,b,c], *Thomas Nichols* [d], *Hans Knutsson* [a,c]

[a]Division of Medical Informatics, Department of Biomedical Engineering,
Linköping University, Linköping, Sweden

[b]Division of Statistic

[c]Ce

[d]Departme

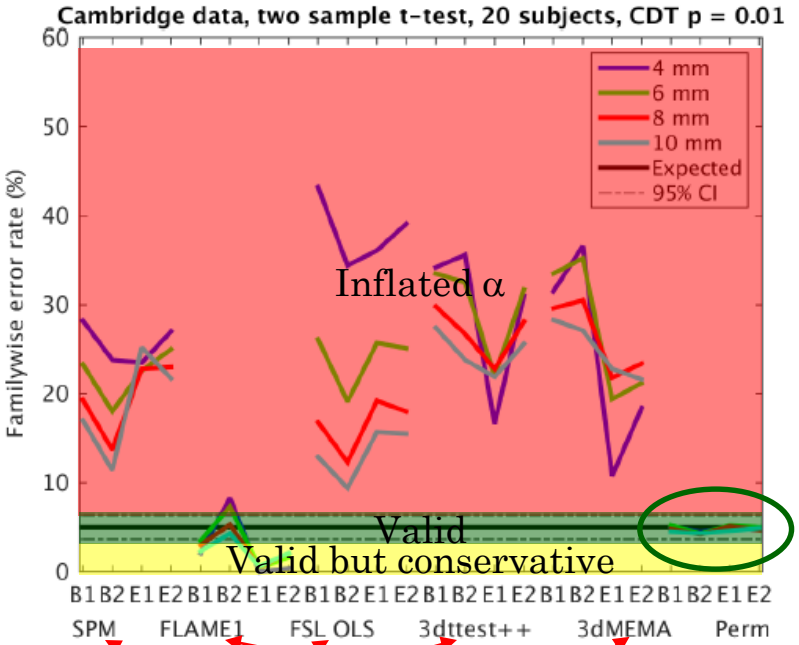
5 Nov 2015

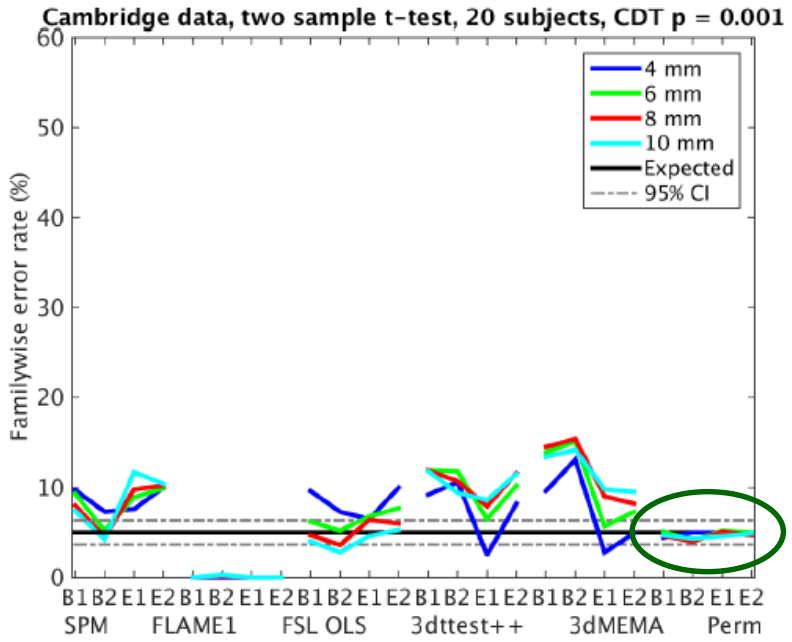
Analysis performed with:

1. **RFX:** SPM, FSL(OLS), AFNI(3dttest++)
2. **MFX:** FSL(FLAME1), AFNI(3dMEMA)
3. **NN-PARAM (perm):** BROCCOLI [like FSL-randomize but much much faster!]

Parameter	Values used
fMRI data	Beijing (198 subjects), Cambridge (198 subjects)
Activity paradigm	Block (B1, B2), event (E1, E2)
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)

2-SAMPLE T-TEST + CLUSTER FEW CORR





RFX

MFX

Z=2.3

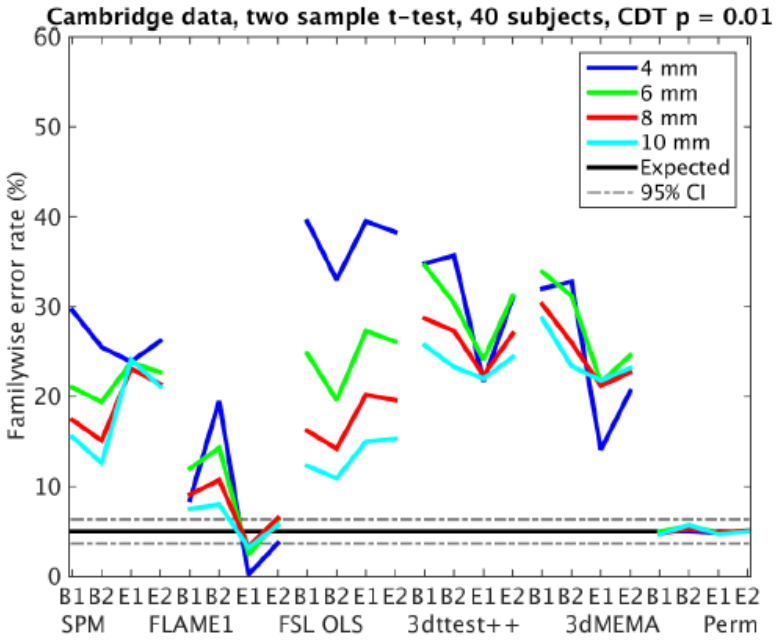
Z=3.1

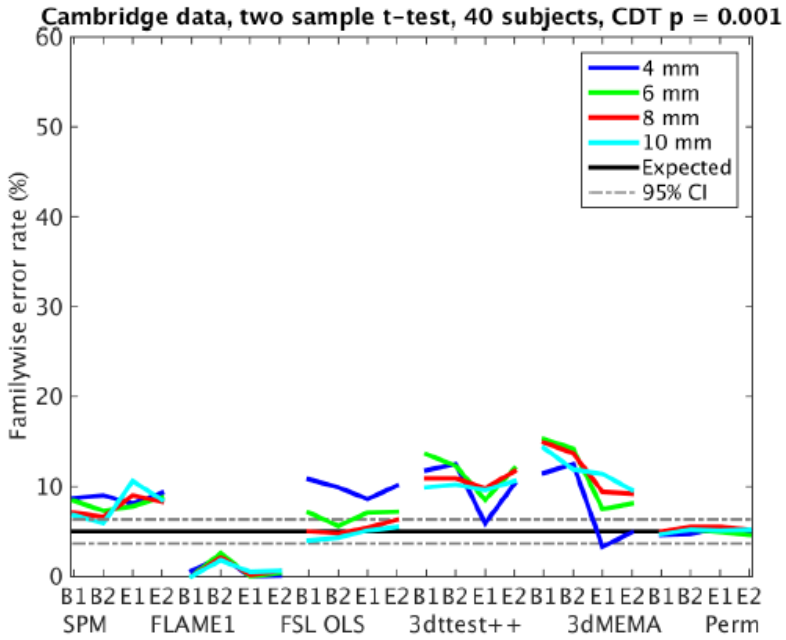
Parametric

Non-Paramm

N=20

2-SAMPLE T-TEST + CLUSTER FEW CORR



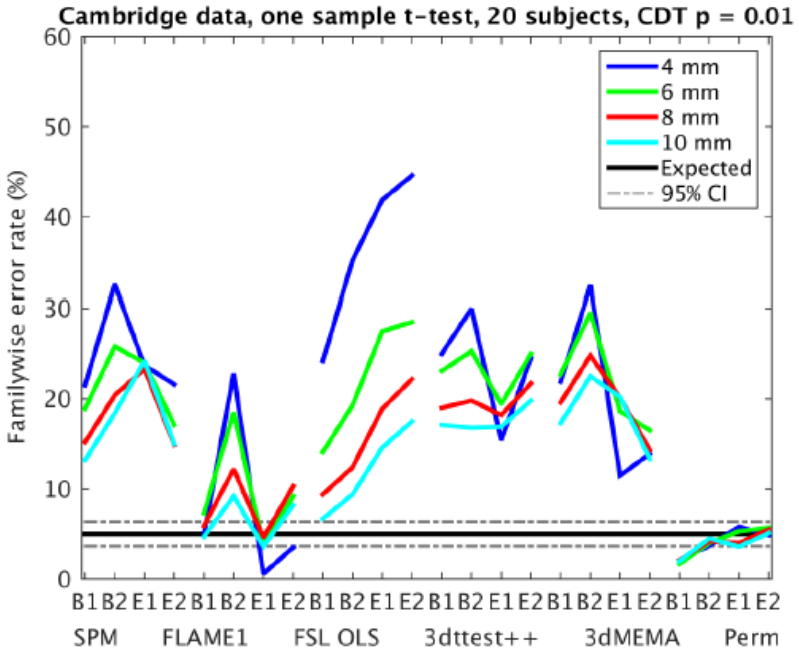


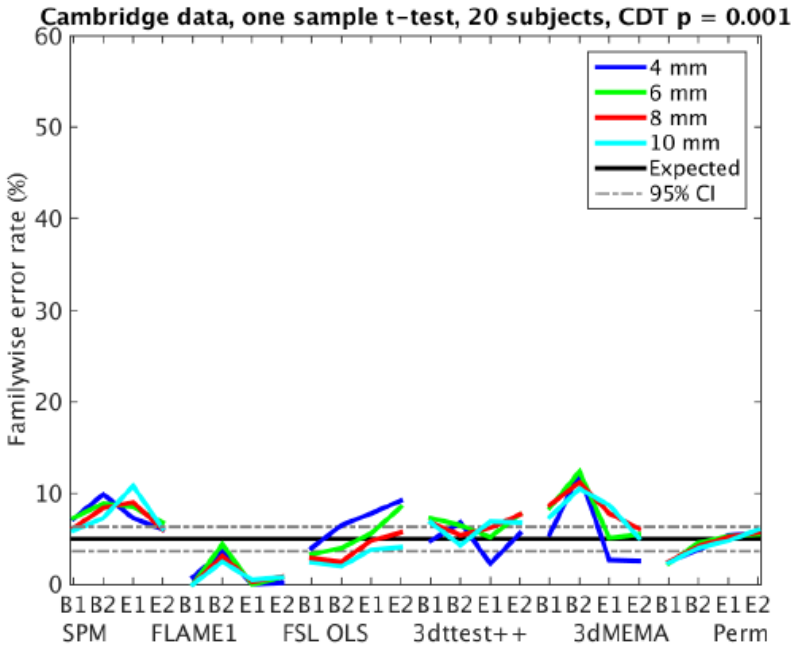
Z=2.3

Z=3.1

N=40

1-SAMPLE T-TEST + CLUSTER FEW CORR

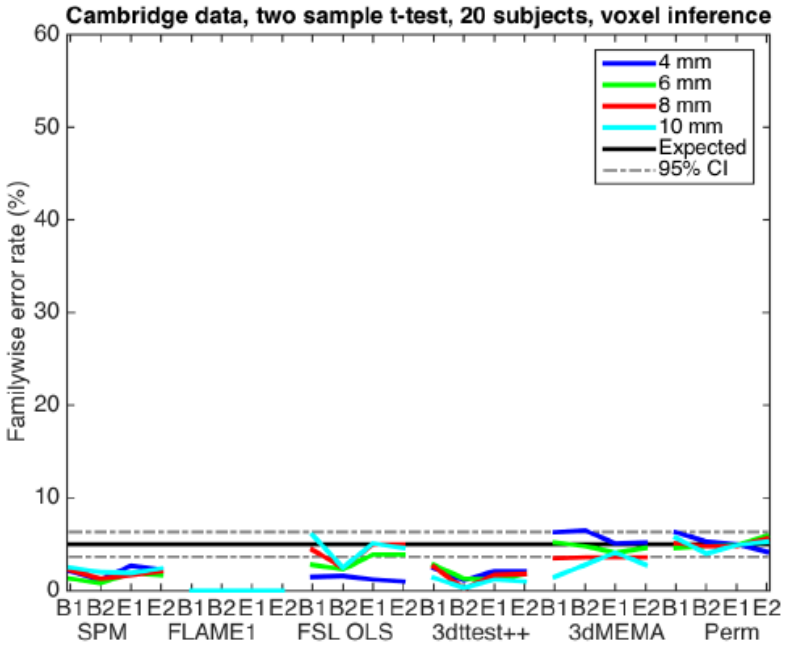


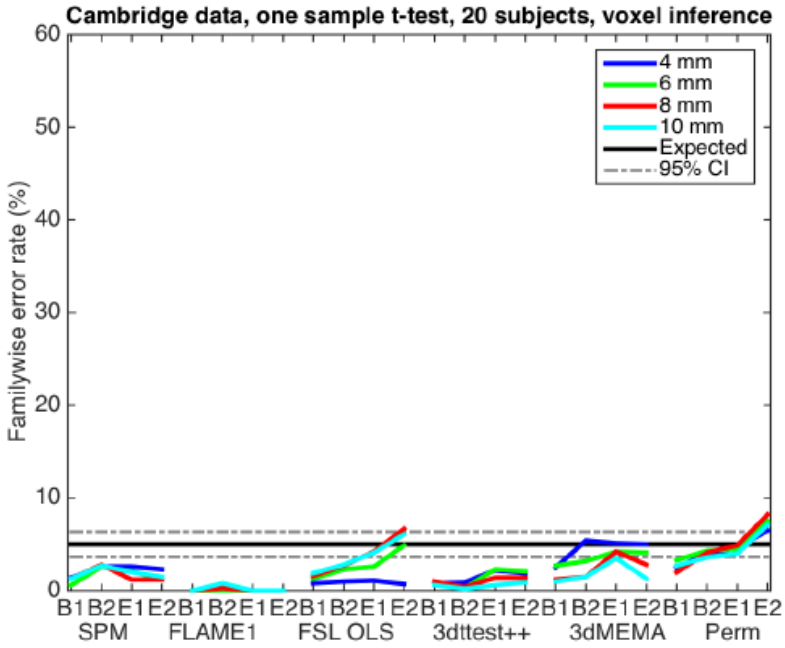


Z=2.3

Z=3.1

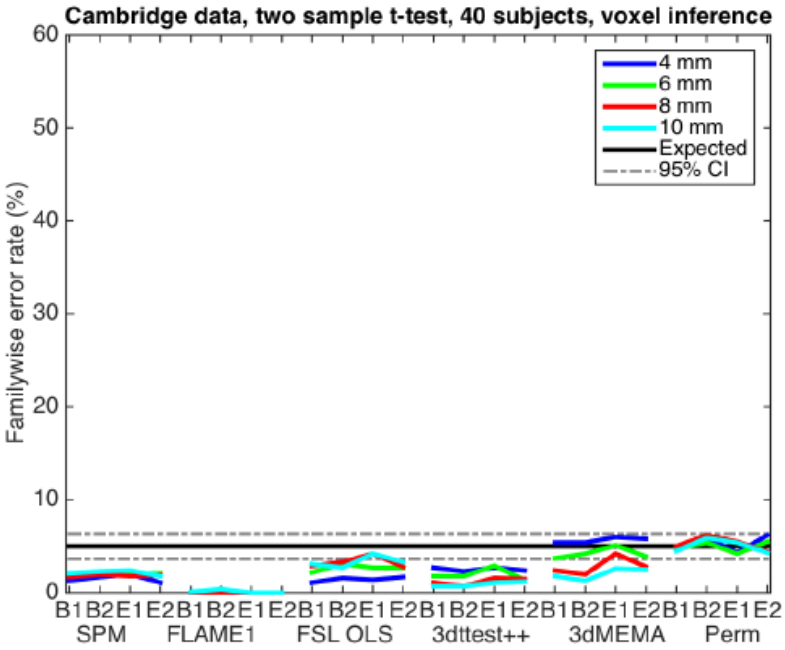
1&2-SAMPLE T-TEST + VOXEL FEW CORR

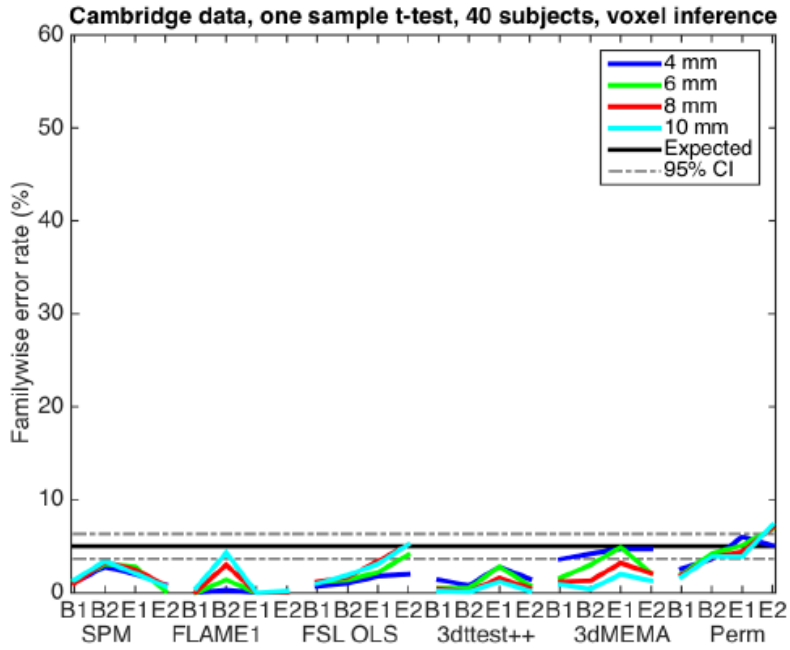




N=20

1&2-SAMPLE T-TEST + VOXEL FEW CORR



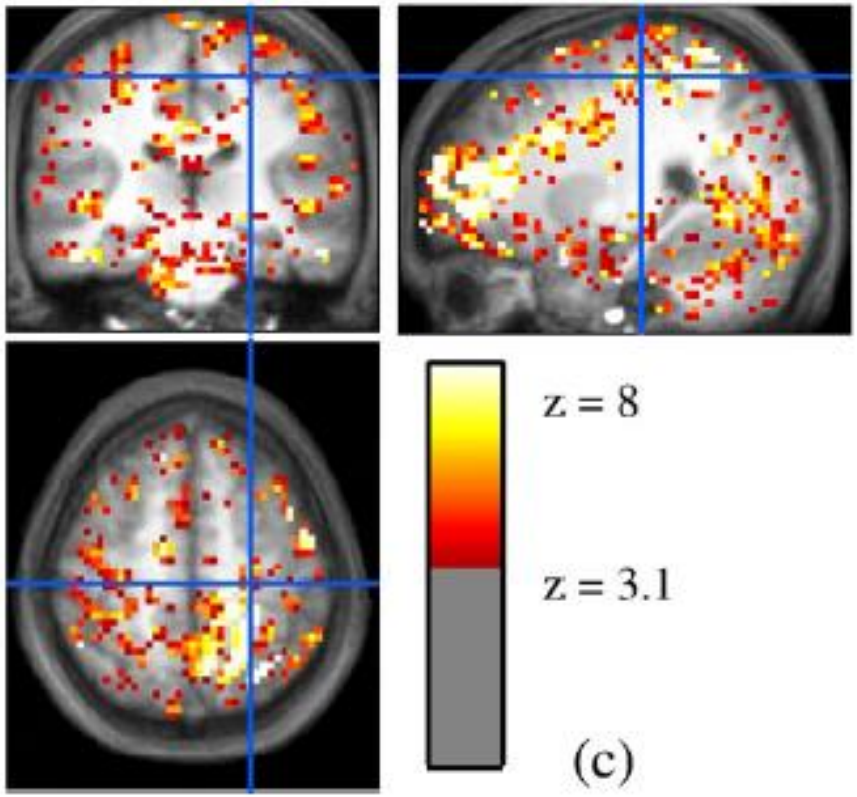


N=40

2-SAMPLE T-TEST + ADHOC: P<0.001 & 10VOX

What are the problems?

I. Remember Thirion et al (i.e., βs are not normal)?

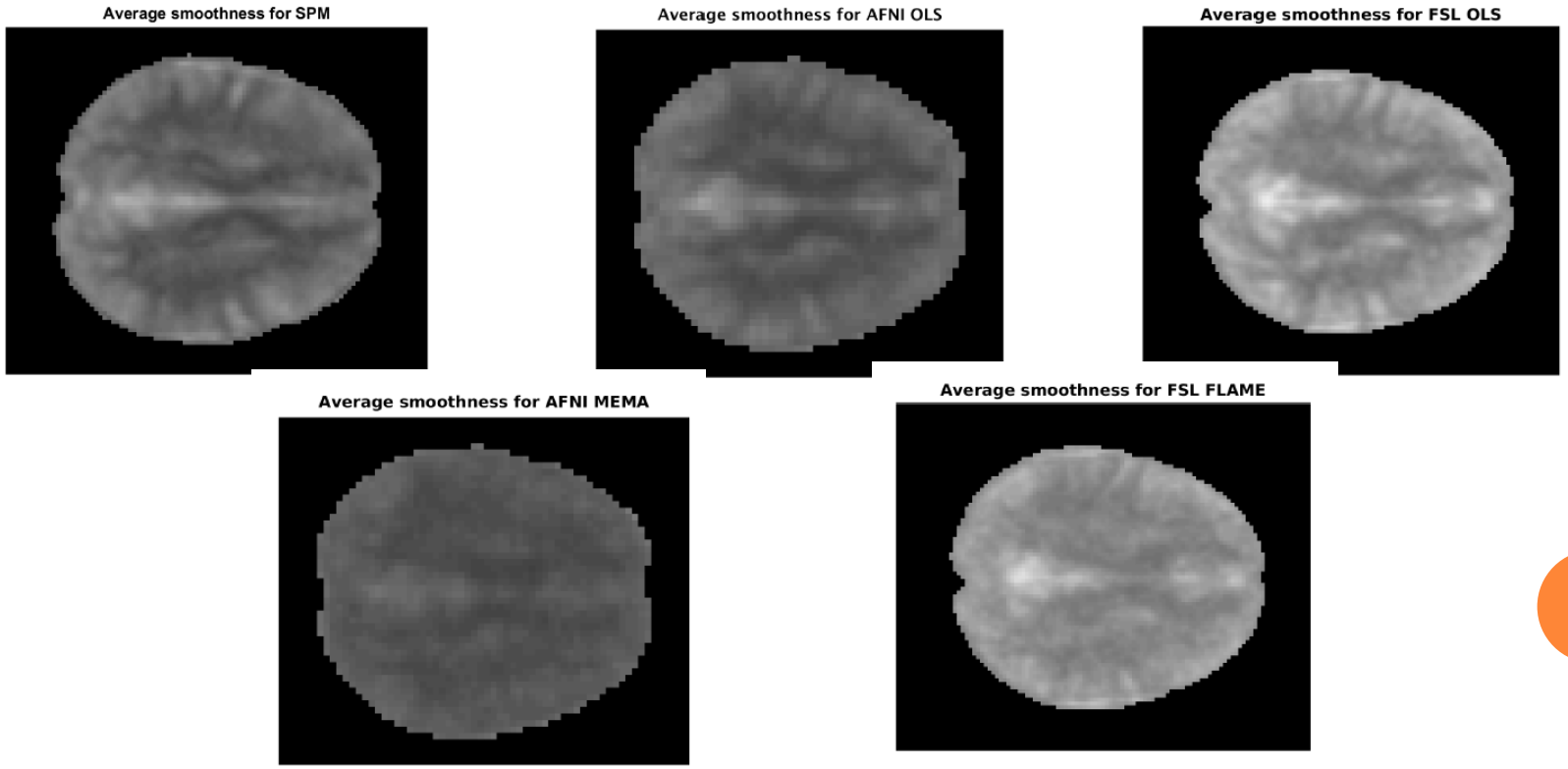


What are the problems?

I. Remember Thirion et al (i.e., βs are not normal)?

II. Gaussian RFT assumptions for cluster-wise FWE:

X Stationary spatial smoothness:

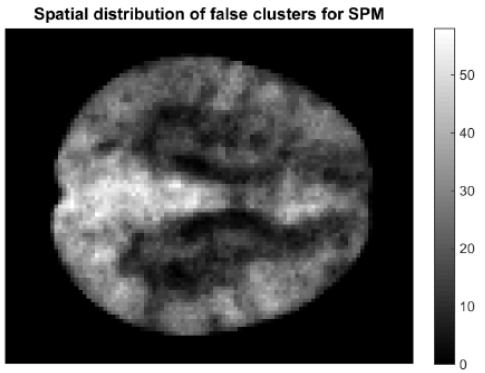


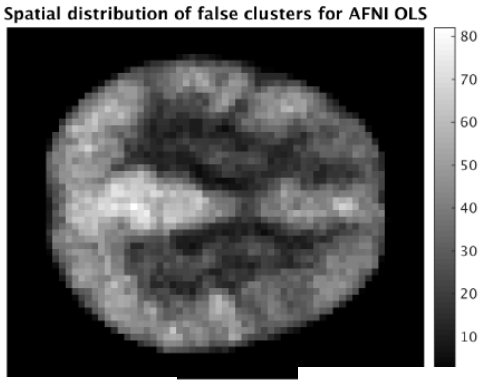
What are the problems?

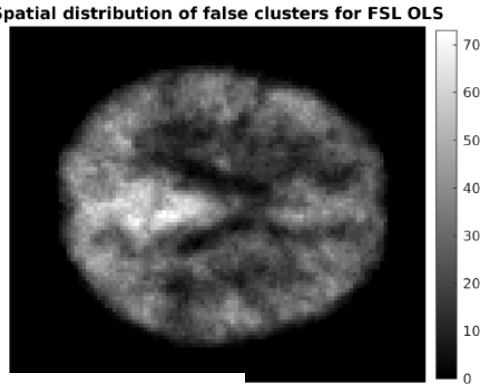
I. Remember Thirion et al (i.e., βs are not normal)?

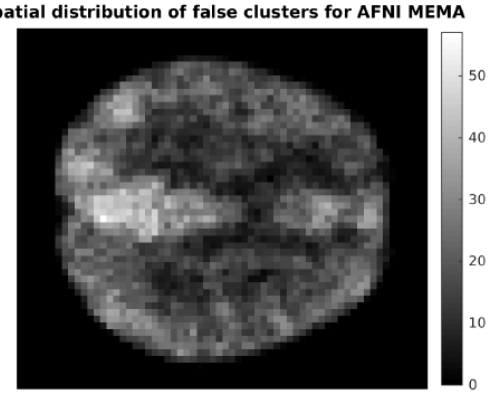
II. Gaussian RFT assumptions for cluster-wise FWE:

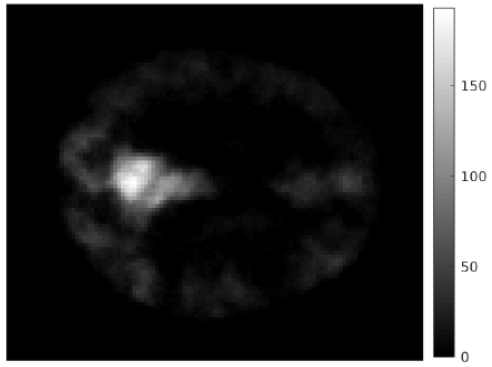
   i. Non-stationarity co-localizes with false activations:











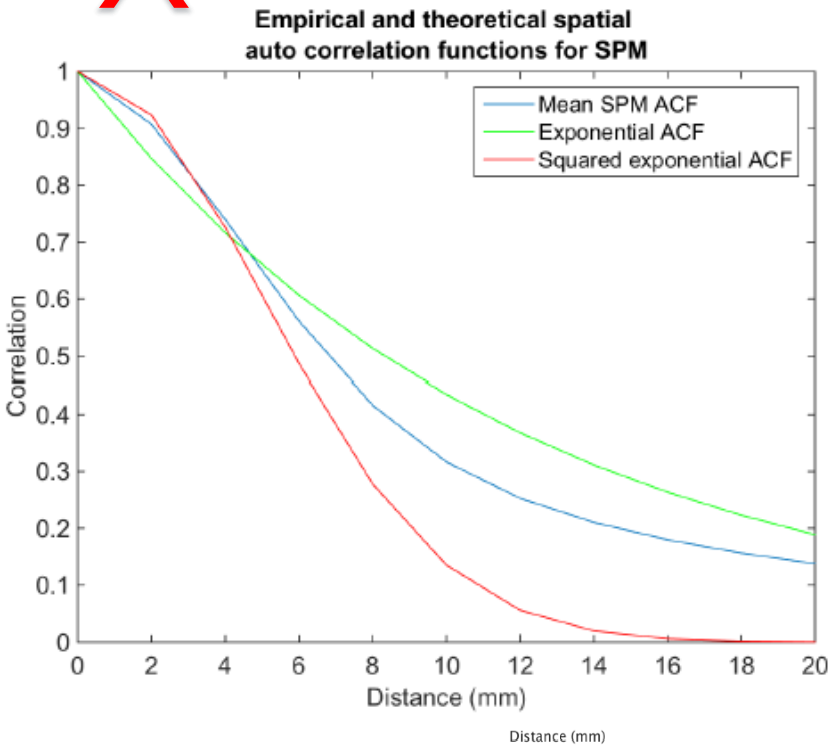
What are the problems?

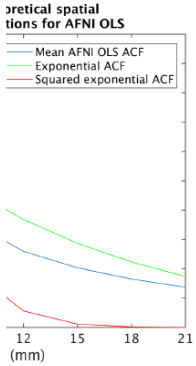
I. Remember Thirion et al (i.e., βs are not normal)?

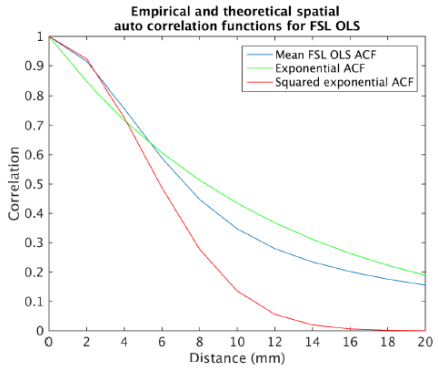
II. Gaussian RFT assumptions for cluster-wise FWE:

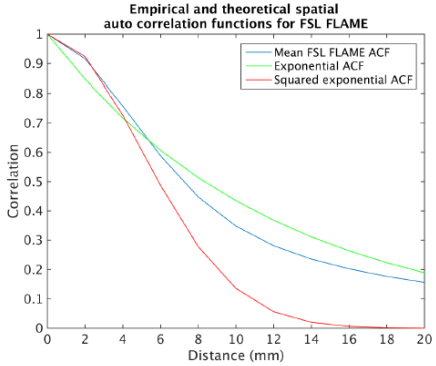
~~1.~~ Stationary spatial smoothness

2. Spatial autocorrelation function ~ squared exponential







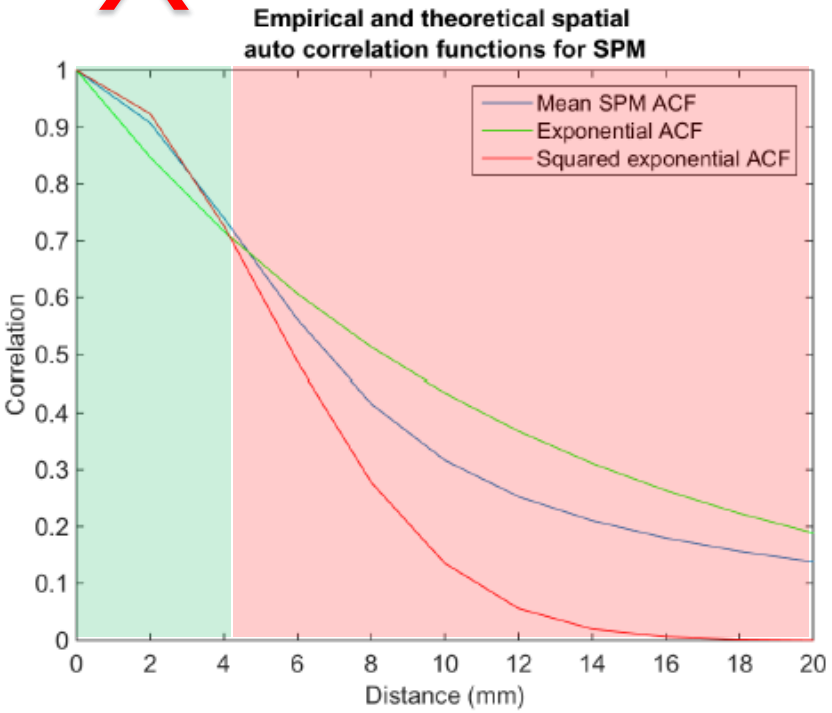


What are the problems?

I. Remember Thirion et al (i.e., βs are not normal)?

II. Gaussian RFT assumptions for cluster-wise FWE:

   1. ~~Stationary spatial smoothness~~
   2. ~~Spatial autocorrelation function ~ squared exponential~~

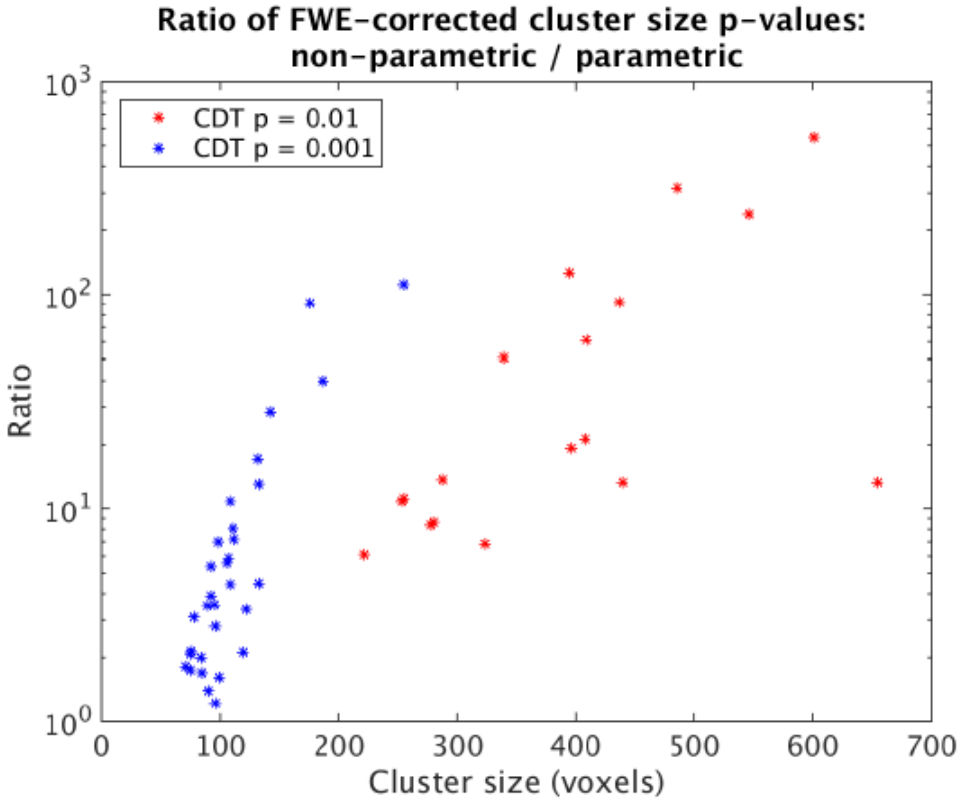


For short distances the approximation holds, it's for long distances that it does not. This might explain why, with high cluster-forming thresholds (Z=3.1), parametric tests' α were less inflated*

What are the problems?

I. Remember Thirion et al (i.e., βs are not normal)?
II. Gaussian RFT assumptions for cluster-wise FWE:
III. Gaussian RFT assumptions for voxel-wise FWE only:
    I. Activity map has to be sufficiently smooth (e.g., 3 vox FWHM)
    II. Spatial autocorrelation function must be twice differentiable

How about task data?

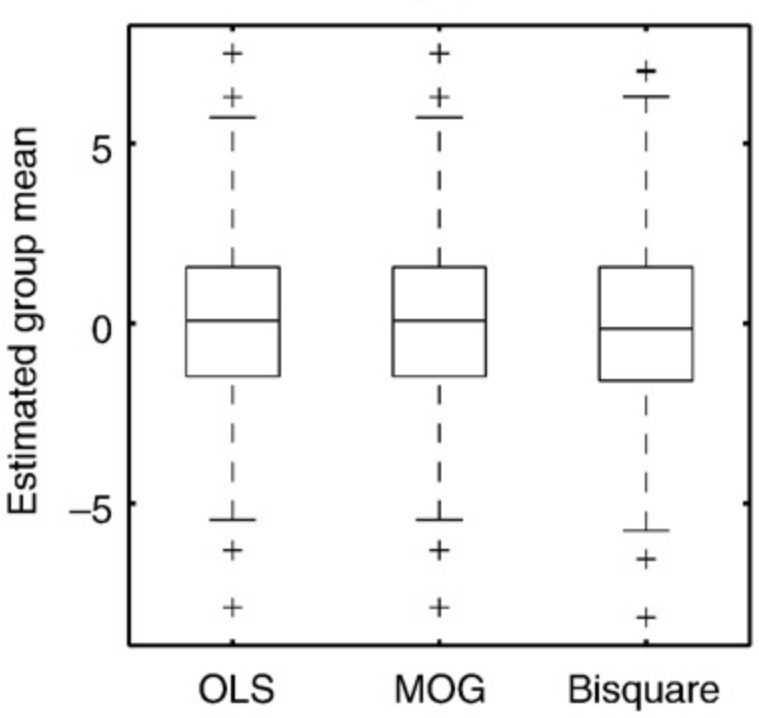


As compared to non-parametric approaches, parametric (cluster FWE corr) p-values are inflated by a factor of 2-3 (for Z=2.3) and 1-2 (for Z=3.1) orders of magnitude.

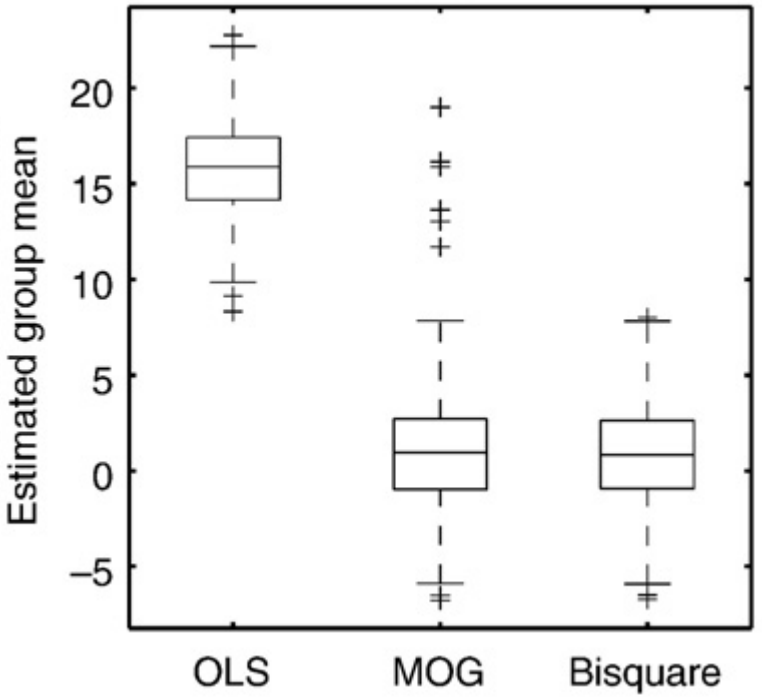
Outliers

- Woolrich 2008

OLS – ordinary least squares
MOG – mixture of gaussians
Bisquare – outlier de-weighting *via* iterative reweighted least squares (IRLS)
Randomise – permutation testing



M=0 & No outliers

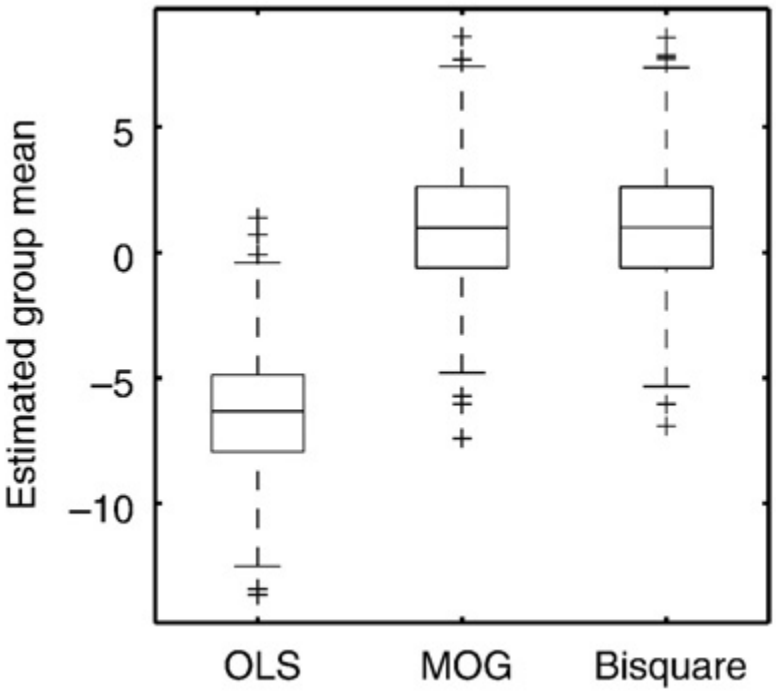


M=1 & 2 positive outlier

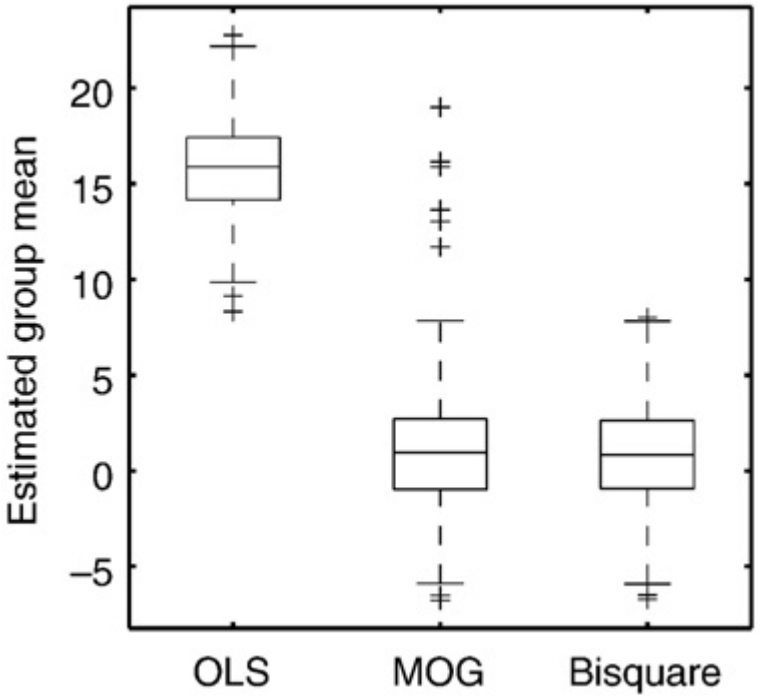
Outliers

- Woolrich 2008

OLS – ordinary least squares
MOG – mixture of gaussians
Bisquare – outlier de-weighting *via* iterative reweighted least squares (IRLS)
Randomise – permutation testing



M=1 & 1 negative outlier



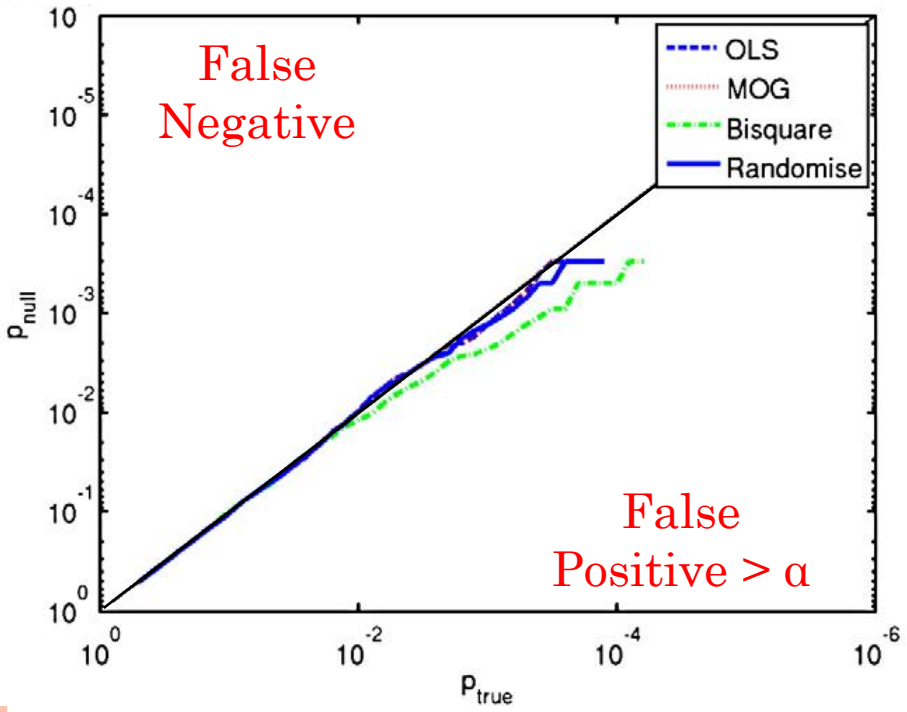
M=1 & 2 positive outlier

Outliers

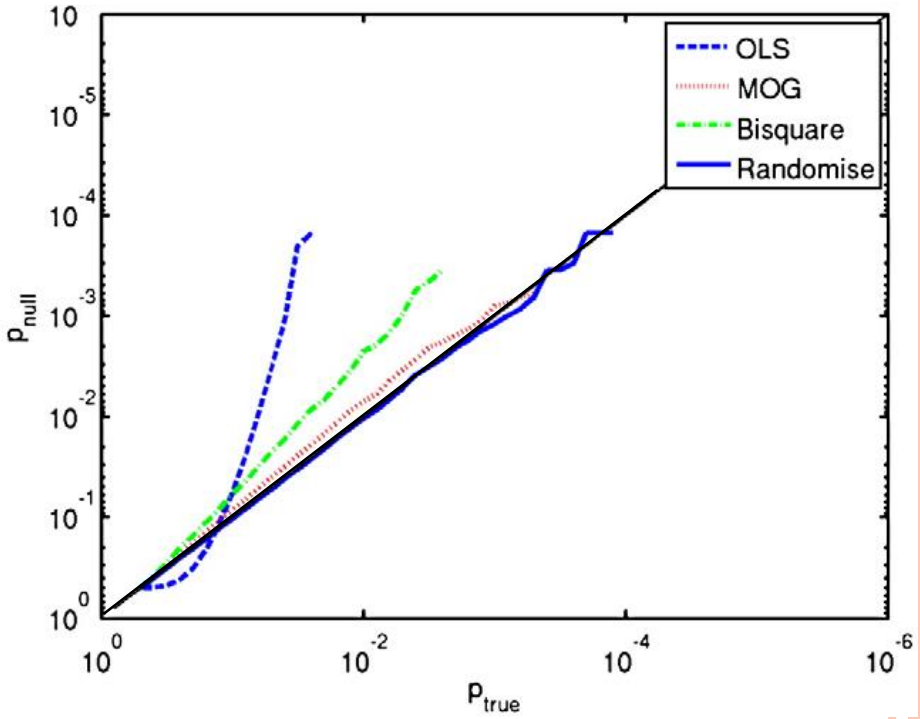
- Woolrich 2008

OLS – ordinary least squares
MOG – mixture of gaussians
Bisquare – outlier de-weighting *via* iterative reweighted least squares (IRLS)
Randomise – permutation testing

Simulation



No outliers



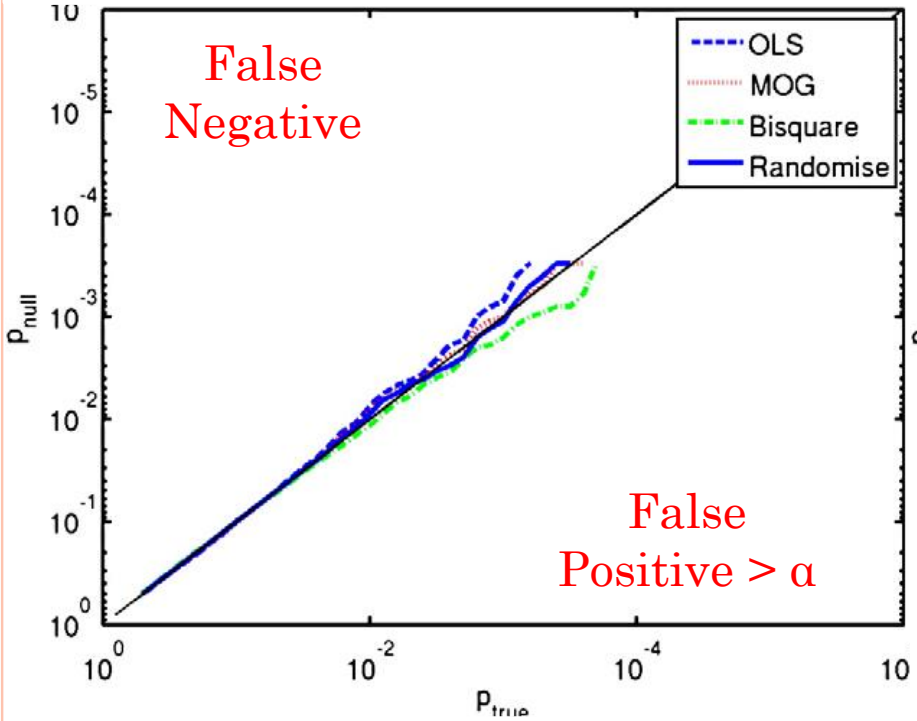
With 1 positive outlier

Outliers

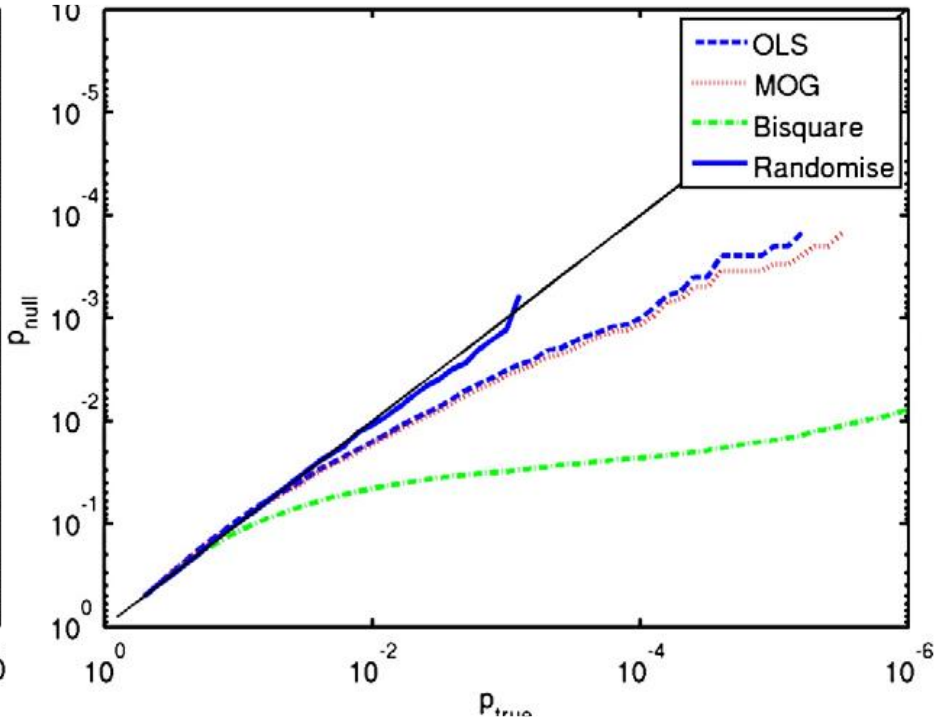
- Woolrich 2008

OLS – ordinary least squares
MOG – mixture of gaussians
Bisquare – outlier de-weighting *via* iterative reweighted least squares (IRLS)
Randomise – permutation testing

Real data



No outliers



With 1 covariate outlier

What can YOU do about it?

i. Ignore it (i.e., use an OLS [stand. SPM, AFNI 3dttest++, FSL-OLS]; more common than you'd think…)
ii. MFX, de-weight outliers/robust regression (i.e., use a WLS/GLS – e.g., FSL-FLAME)
iii. Use non-parametric (permutation) tests and forget **all** of the problems we discussed above:
    i. Does not depend on paradigm, smoothing, inference level (voxel v cluster), cluster thresholding
    ii. Only assumption: **exchangeability**
    iii. Available software: SnPM, FSL randomize*, BROCCOLI, [***extra perks:** (i) TFCE, (ii) it does permutation on $\hat{\beta}/\hat{\sigma}^2$]

ROIs:

- Rousselet GA & Pernet CR (2012) Improving standards in brain-behavior correlation analyses, *Frontiers in Human Neruoscience*, doi: 10.3389/fnhum.2012.00119

Group Analyses:

- Nichols TE & Holmes AP (2001) Nonparametric permutation tests for functional neuroimaging: a primer with examples, *Human Brain Mapping* 15: 1-25.
- Thirion *et al* (2007) Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses, *NeuroImage* 35: 105-120.
- Woolrich M (2008) Robust group analysis using outlier inference, *NeuroImage* 41: 286-301.
- Eklund A *et al* (2016) Can parametric statistical methods be trusted for fMRI based group studies? *arXiv preprint arXiv:*1511.01863

Question: If I am a reviewer, should I demand a non-parametric re-analysis?

- Well, ***theoretically yes***, since we now have data clearly showing that most tools have much higher error rates for a nominal 5% (perhaps with the exception of FLAME under specific parameter choices) and *you want this field to be better!*
- In practice, it depends on ***you***. However, in my opinion, if the paper uses FSL and they did a standard FSL group analysis, then ***there is no excuse not to run*** randomise which, if you've already done a group analysis, takes 1 line and a little (computer) time.