UNIVERSITEIT GENT

FACULTY OF **PSYCHOLOGY AND EDUCATIONAL SCIENCES**

Stability Based Testing for the Analysis of fMRI Data

Joke Durnez & Beatrijs Moerkerke

Department of Data Analysis, Ghent University, Belgium

7^{th} International Conference on Multiple Comparison Procedures
September 1, 2011

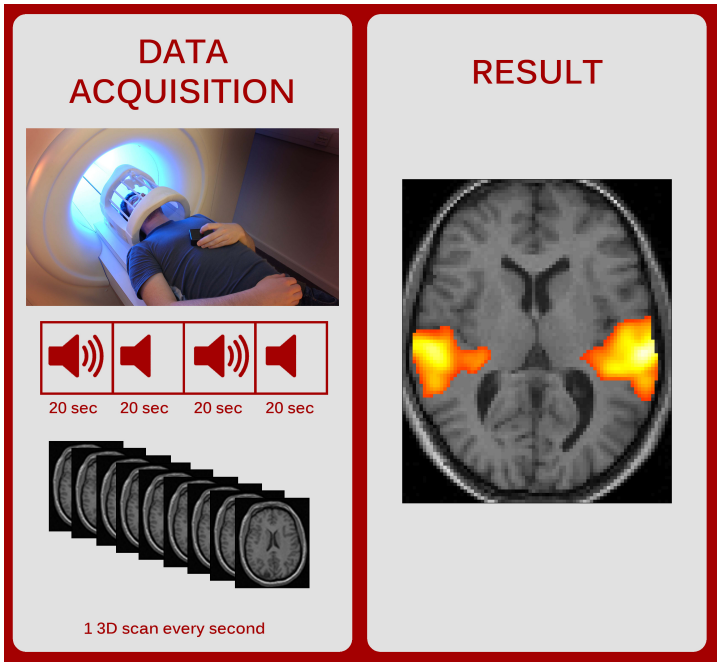
UNIVERSITEIT GENT

Table of Contents

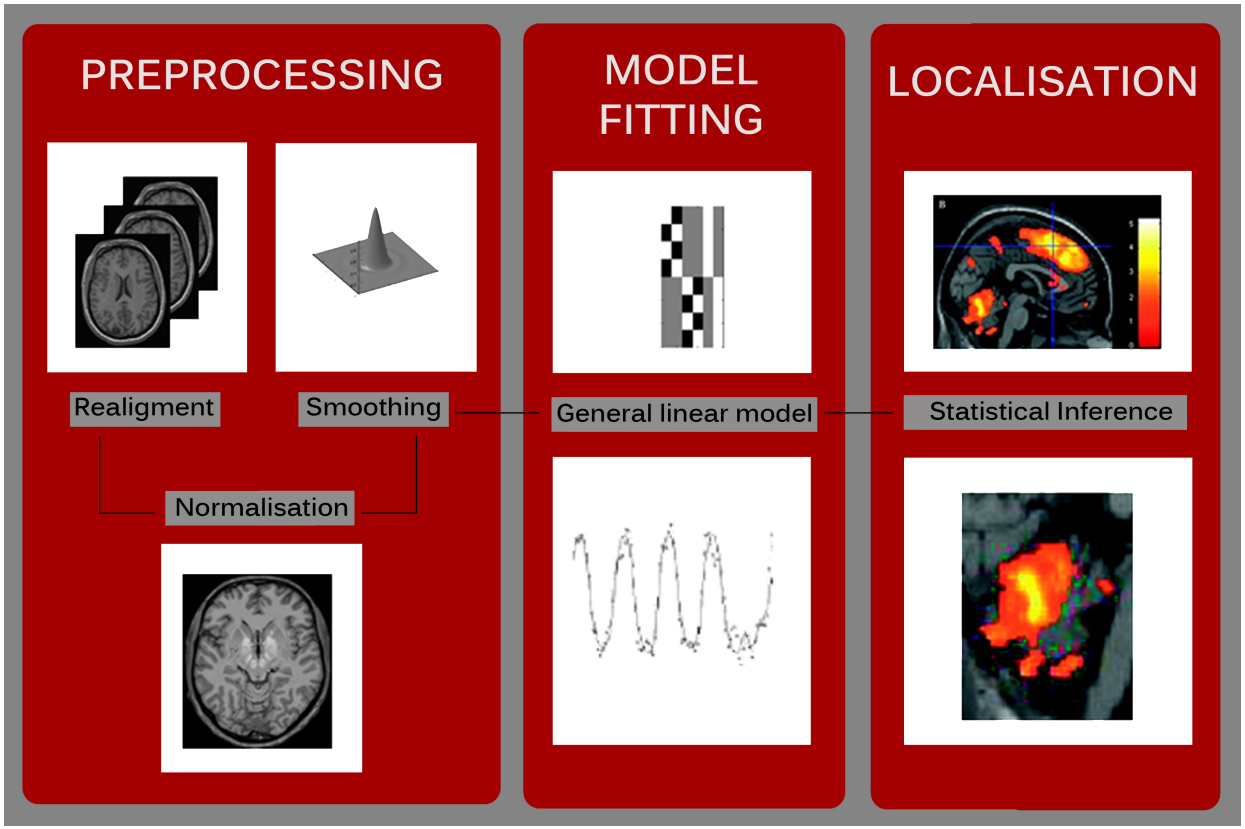
UNIVERSITEIT GENT

functional Magnetic Resonance Imaging

- MRI: detects neurological structures/fluids/...
- functional: detects brain functions, eg. auditive



functional Magnetic Resonance Imaging



- Inference: testing 200 000 voxels simultaneously
- Huge multiple testing problem

UNIVERSITEIT GENT

functional Magnetic Resonance Imaging

	Active (H_1)	Non active (H_0)
Significant		False positive (type I error)
Non significant	False negative (type II error)	

- Inference: testing 200 000 voxels simultaneously
- Huge multiple testing problem

UNIVERSITEIT GENT

Multiple Testing Corrections for fMRI

Control of the family-wise error rate (FWER)

- **Bonferroni**
  - Loss of power
  - Spatial correlation
- **Random Field Theory**
  - Estimates the number of independent tests = resels
  - FWER on the number of clusters: $P(c > C|H_0) < .05$
  - Less strict than Bonferroni
  - Still very conservative
- General literature: different procedures accounting for correlation structures
- BUT: FWER conservative as a measure

UNIVERSITEIT GENT

Multiple Testing Corrections

Control of the false discovery rate (FDR)

- allow more type I errors
- FDR: the proportion of type I errors among all significant voxels
- **Benjamini-Hochberg**
  - Less conservative→ more power

HOWEVER

- Power is only gained by increasing p-value threshold
- Ranking of voxels remains the same

UNIVERSITEIT GENT

Need for new procedures

- Statistical significance $\neq$ biological importance
  - $\rightarrow$ balance type I and type II errors
- What about reproducibility of results?

UNIVERSITEIT GENT

Need for new procedures

The relationship between type I and type II errors

- P(type I error) $\downarrow \Rightarrow$ P(type II error) $\uparrow$
- Consequences type I error:
  - Further research to false activation
  - False theories
- Consequences of type II error:
  - True activation is not detected
- Need for a better balance between type I and type II errors (Moerkerke & Goetghebeur, 2006; Lieberman & Cunningham, 2009)

UNIVERSITEIT GENT

Need for new procedures

Validation of procedures

- Average performance with respect to error measures
- But also important: stability of test results
  - Stability is related to reproducibility of results
  - Stability eg. can be measured as the variance on the number of selected voxels
  - Largely unexplored
  - Choice of multiple testing method influences the stability
  - statistical genetics: FDR controllingprocedures tend to be less stable than FWER controlling procedures. (Qiu, Xiao, Gordon, & Yakovlev, 2006)

UNIVERSITEIT GENT

Need for new procedures

Goal of the current research

- Develop a new selection mechanism that allows to weigh P(type I error) and P(type II error) in the selection mechanism.
- Take into account the stability of the voxels.

Gorden, Chen, Glazko and Yakovlev (2009)

- Resample gene arrays
- Generate 'new' datasets and apply selection criterion (eg. BH)
- New criterion: select only the *genes* that are selected in h% of the resamples
- h: selection percentage
  - h = c1/(c1+c2)
    - c1 = weight to type I errors
    - c2 = weight to type II errors
- Our goal: adapt this method to fMRI data

UNIVERSITEIT GENT

New method: stability based testing

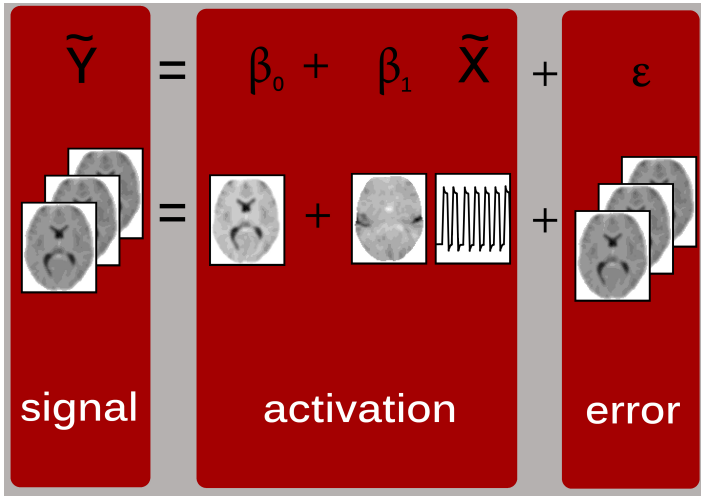
Step 1: Whitening the data

- Resampling assumes temporal independence
- fMRI data is autoregressive
- Whitening the data = removing temporal correlation
- Asuming AR(1) structure:
  $W \equiv V^{-1/2}$
  $\tilde{X} = WX$
  $(X = \textit{designmatrix})$
  $\tilde{Y} = WY$
  $(Y = \textit{signal (after preprocessing)})$

UNIVERSITEIT GENT

New method: stability based testing

Step 2: GLM



Step 3: Generate new datasets

- Take a bootstrap sample of the errors
- Add them to the estimated activation
- $\Rightarrow$ **new dataset**

UNIVERSITEIT GENT

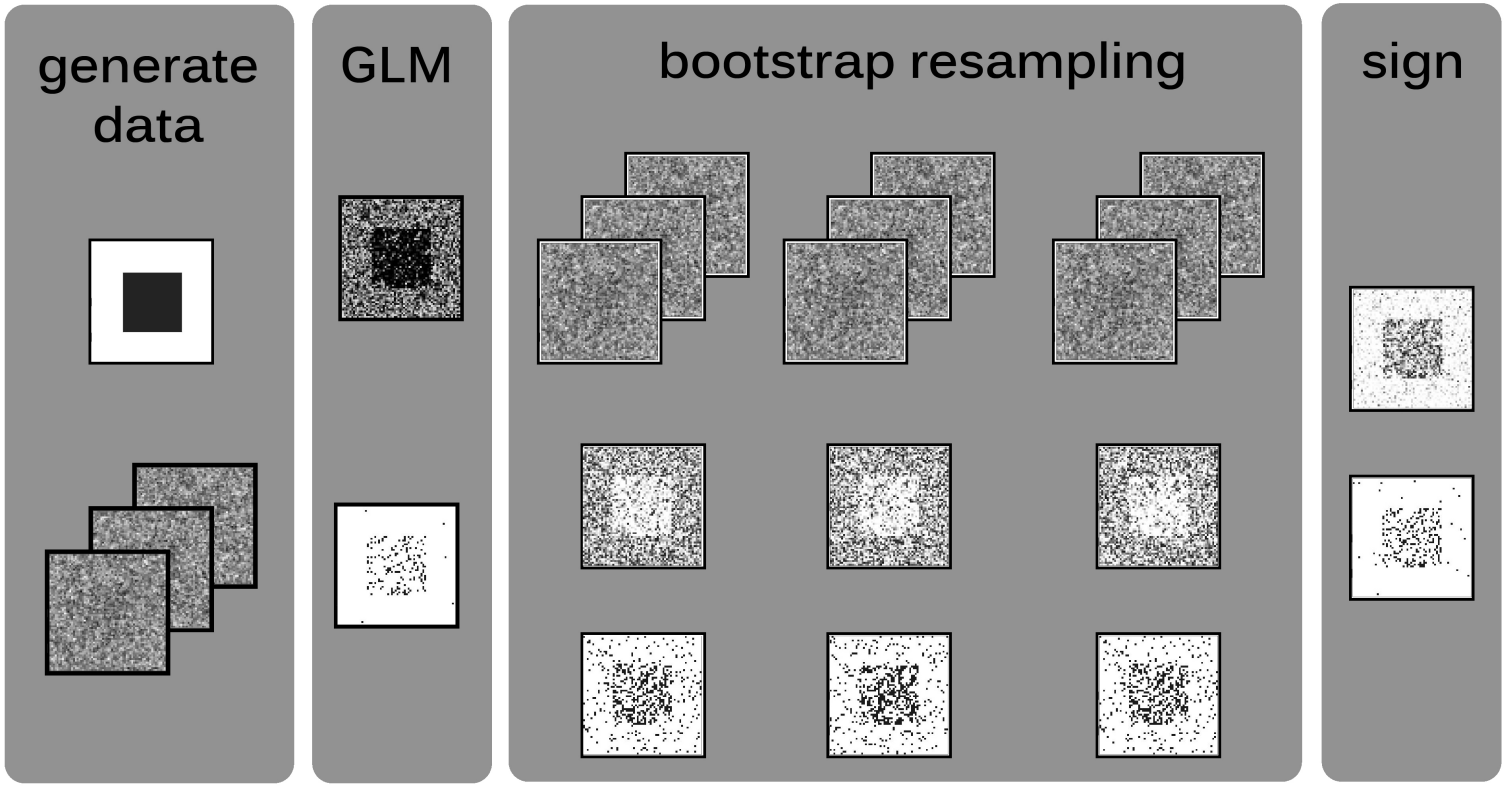
New method: stability based testing

Step 4: Activation detection

- Create 100 new datasets
- Test for activation in each dataset: significant or not?
- For each voxel: frequency of occurence H
- Threshold = predefined percentage h

UNIVERSITEIT GENT

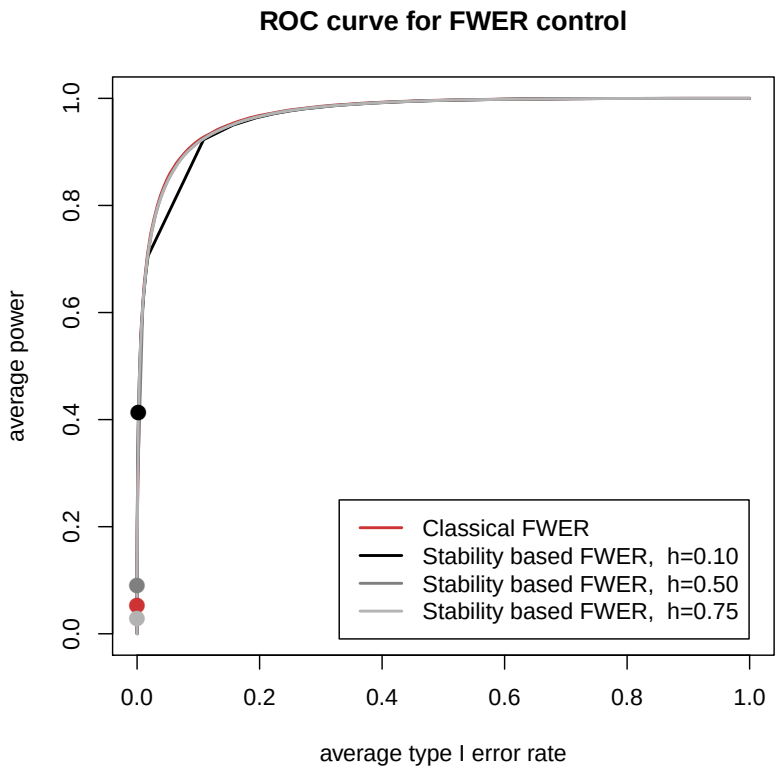
Simulated data

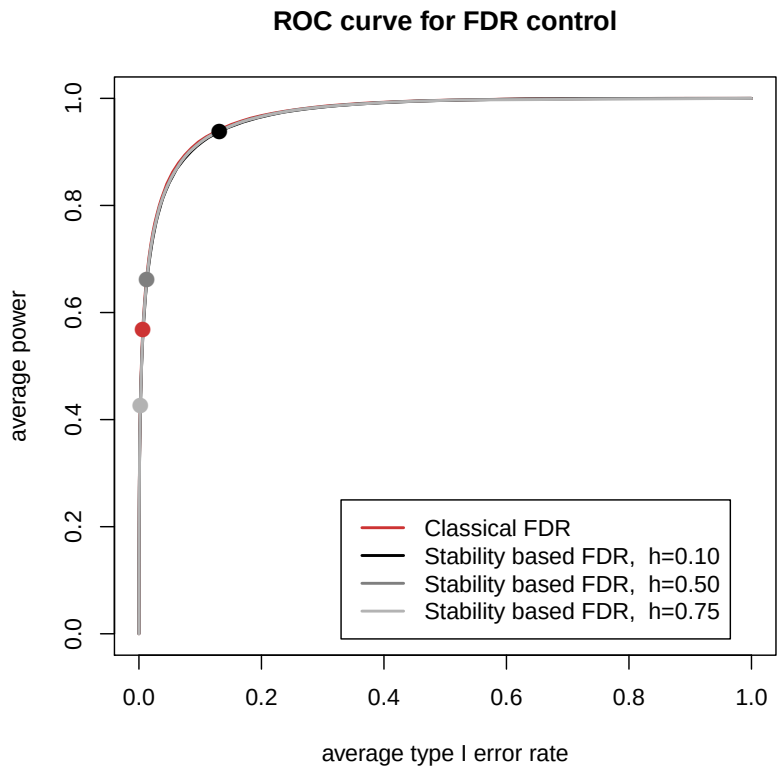


UNIVERSITEIT GENT

Simulated data: error rates

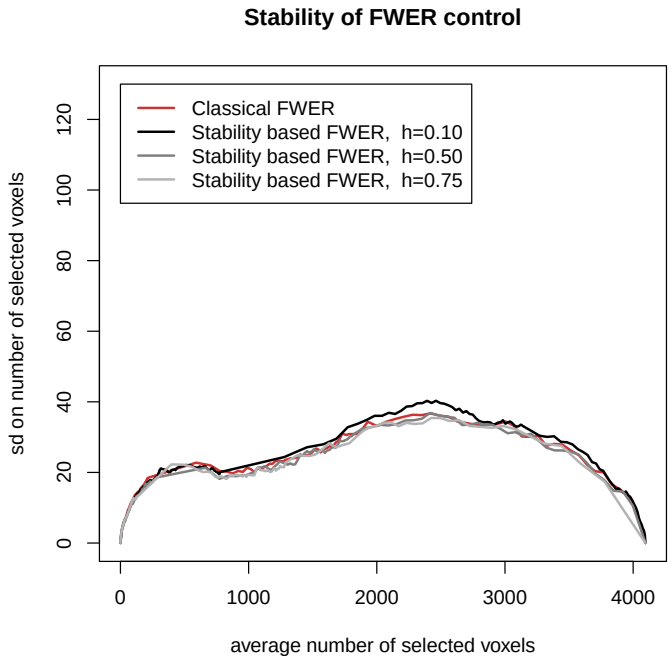
based on 100 bootstraps

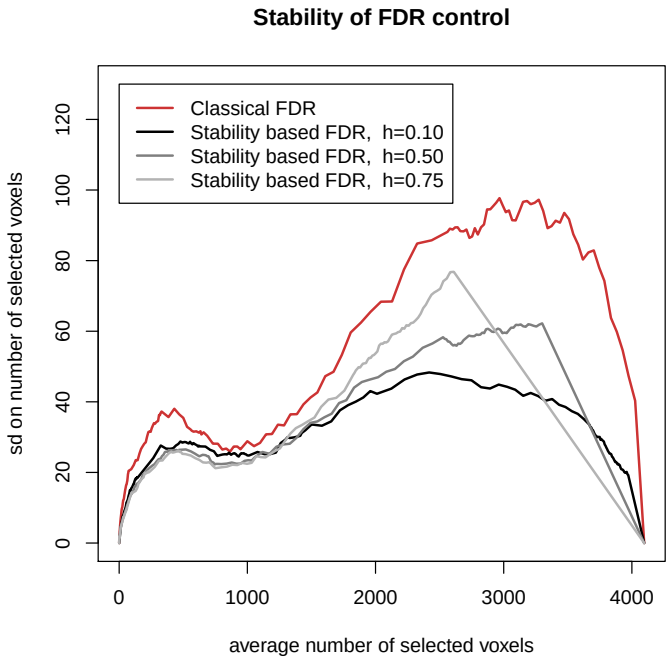




UNIVERSITEIT GENT

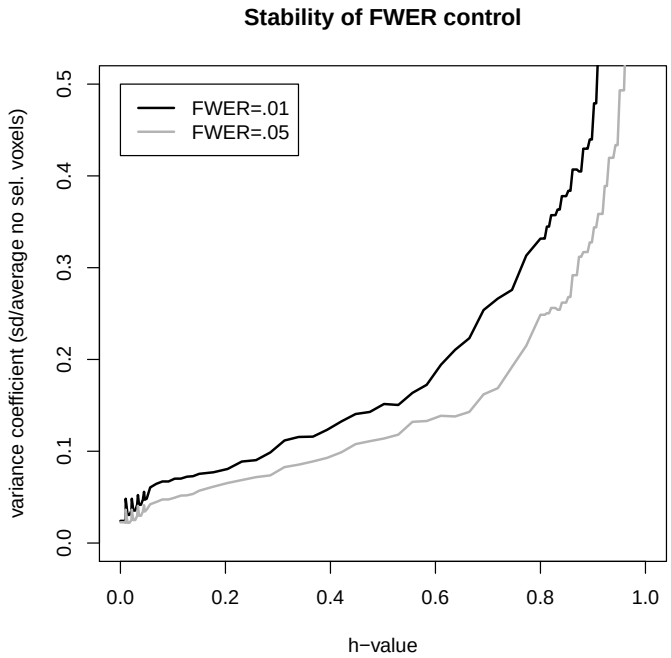
Simulated data:stability

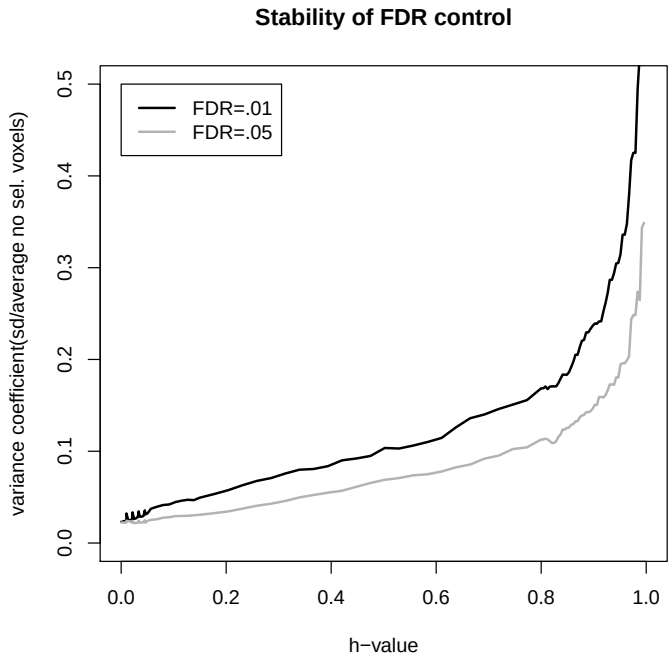




UNIVERSITEIT GENT

Simulated data:stability





UNIVERSITEIT GENT

Simulated data

Results

- ROC: Balance between sensitivity and specificity remains the same
- BH becomes more stable when the selection percentage is taken into account
- Role of h-value

UNIVERSITEIT GENT

Real data

Meaning of stability on real data

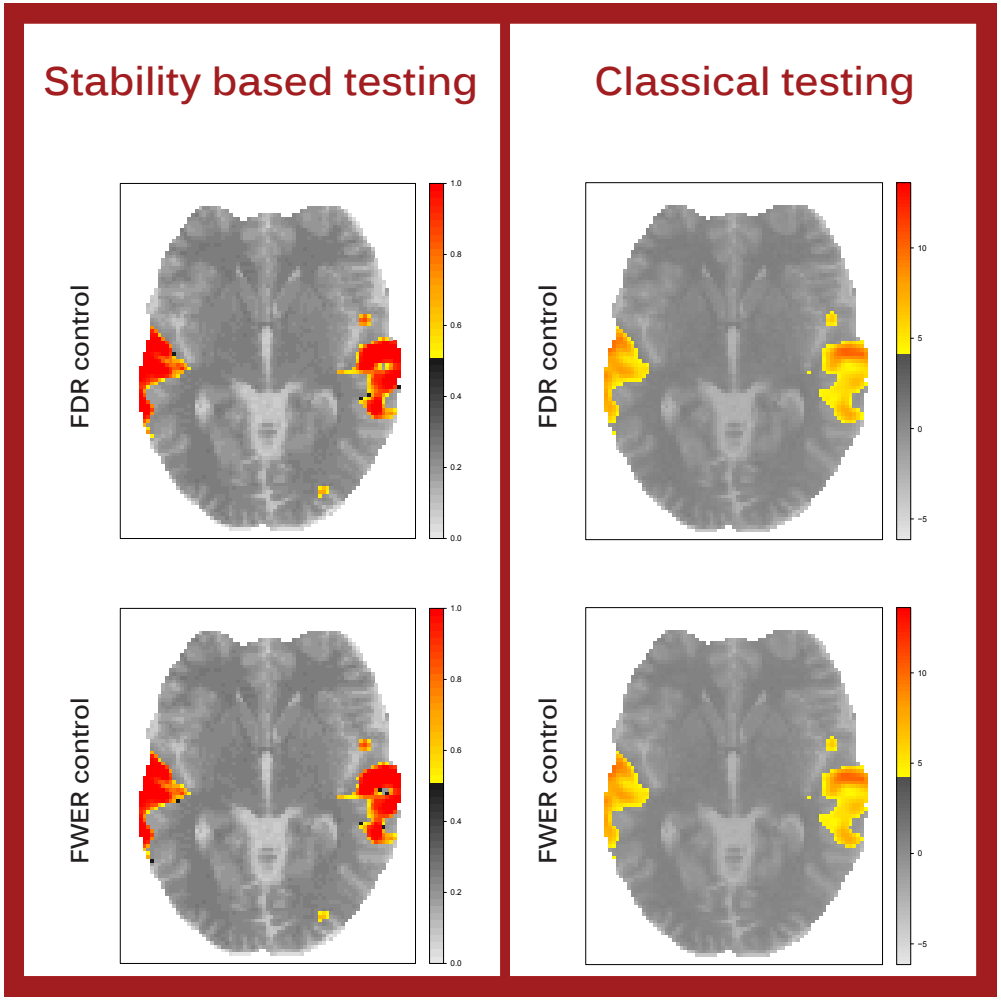
- Stability based procedure involves bootstrapping
- Evaluation on real data: simulation $\rightarrow$ bootstrapping
- $\Rightarrow$ 2 levels of bootstrapping: computationally too heavy
- For now:
  - Bootstrap used to give results with stability based testing

Auditory dataset

- (Friston, 2007)
- Single subject blocked design: rest-auditory stimulation
- 64 x 64 x 64 voxels

UNIVERSITEIT GENT

Real data


Stability based testing
Classical testing
FDR control
FWER control
FDR control
FWER control

Concluding remarks

- FDR control was less stable than FWER control, but stability was improved using stability based testing
- No change in balance between error rates
- h-threshold: relative costs of type I and type II errors
- Cluster-based methods: stability of clusters
- Bootstrapping procedure
  - Alternative procedures
  - Role of smoothing

UNIVERSITEIT GENT

References

Friston, K. J. (2007). Topological inference. In K. J. Friston, J. T. Ashburner, S. J. Kiebel, T. E. Nichols, & W. D. Penny (Eds.), *Statistical parametric mapping: The analysis of functional brain images* (pp. 237–245). Academic Press.

Gordon, A., Chen, L., Glazko, G., & Yakovlev, A. (2009). Balancing type one and two errors in multiple testing for differential expression of genes. *Computational Statistics and Data Analysis*, *53*, 1622–1629.

Lieberman, M. D., & Cunningham, W. A. (2009). Type i and type ii error concerns in fmri research; re-balancing the scale. *Social cognitive and affective neuroscience*, *4*, 423–428.

Moerkerke, B., & Goetghebeur, E. (2006). Selecting 'significant' differentially expressed genes from the combined perspective of the null and the alternative. *Journal of computational biology*, *13*, 1513–1531.

Qiu, X., Xiao, Y., Gordon, A., & Yakovlev, A. (2006). Assessing stability of gene selection in microarray data analysis. *bioinformatics*, *7*(50).

UNIVERSITEIT GENT