UNIVERSITEIT GENT

Improving selection stability of multiple testing procedures for fMRI

Joke Durnez and Beatrijs Moerkerke

1 Problem setting

- Simultaneously testing thousands of voxels for activation leads to an inflation of false positives (**the multiple testing problem**).
- Several (voxelwise) corrections to protect the amount of false positives exist, for example:
  - o Bonferroni (BF): protects the family-wise error rate
  - o Benjamini-Hochberg (BH): protects the false discovery rate
- Multiple testing procedures are evaluated on their average performance with respect to error rates.
- What about the variability on the results?
- The higher the **selection variability**, the lower the **selection stability**.
- Qiu et al. (2006) show that multiple testing procedures can be highly unstable.

2 Goal

1. Using bootstrap procedures, we **measure selection variability** on test results for multiple testing corrections following BF (FWER) and BH (FDR).
2. We present a **new testing strategy** which includes both **significance** and **selection variability** in the decision criterion (Gordon et al., 2009).
3. The new procedure is evaluated through Monte Carlo simulated fMRI images.

Does the new procedure improve the selection stability?

3 Methods

Creation of images

We create 2-dimensional images consisting of 64x64 voxels. Activation is placed in the center of the image (32x32 voxels). We consider a 20s ON/ 20s OFF block design (TR=1) repeated 3 times and convolved with a canonical HRF. Temporal and spatial noise is added.

GLM

Before voxelwise linear regression of the measured signal on the signal components, data are pre-whitened to account for the temporal correlation using the procedure described by Worsley et al. (2003).

voxel thresholding

p-values are thresholded according to BF and BH

New testing strategy and voxel selection

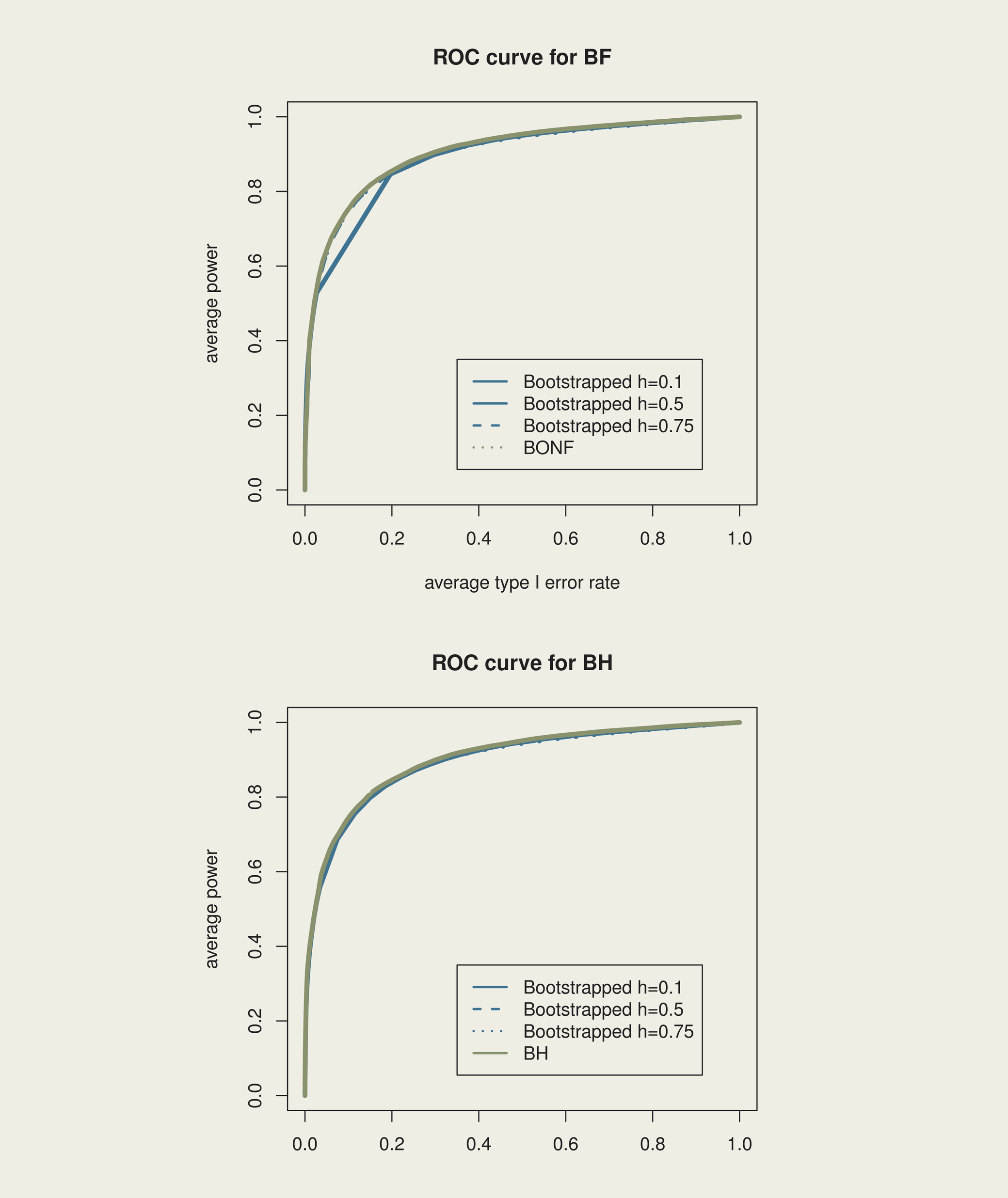
- 100 bootstrap samples of size n=120 are taken from the residuals of the voxels.
- The bootstraps are used to construct 'replicates' for the time series of images, by re-adding the temporal correlation.
- Each replicate obtained by bootstrapping is analysed as the original time series.
- 4 different criteria for voxel selection:
  (1) the original multiple testing procedure (MTP) and selected in at least
  (2) h = 10%, (3) h = 50%, (4) h = 75% of the replicates using the MTP.

simulation

- We create 500 time series of images
- The average performance over 500 simulations is investigated

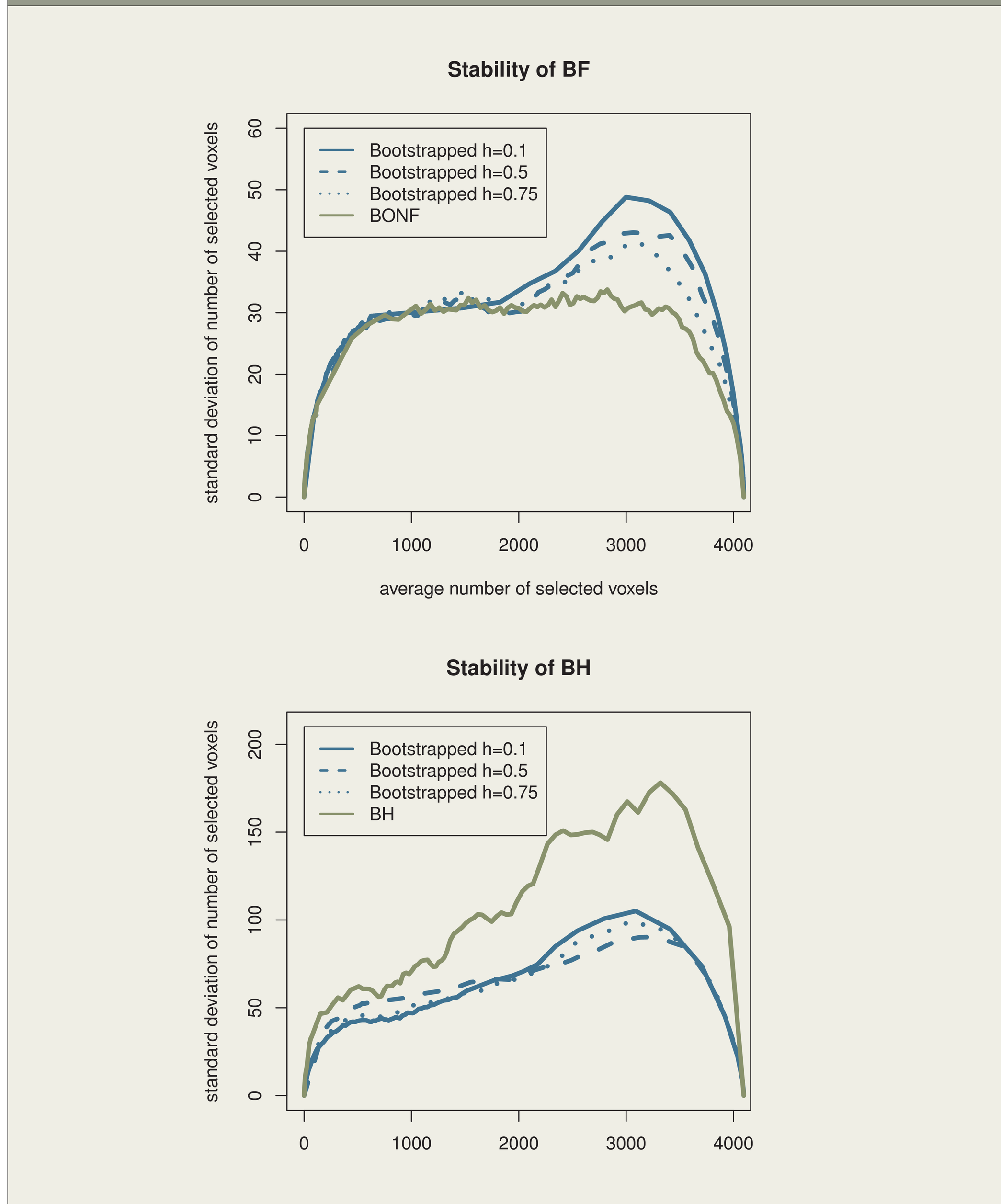
4 Results

ROC-curves



For BF, BH and their bootstrap counterparts, the relation between sensitivity and specificity is the same.

Selection variability

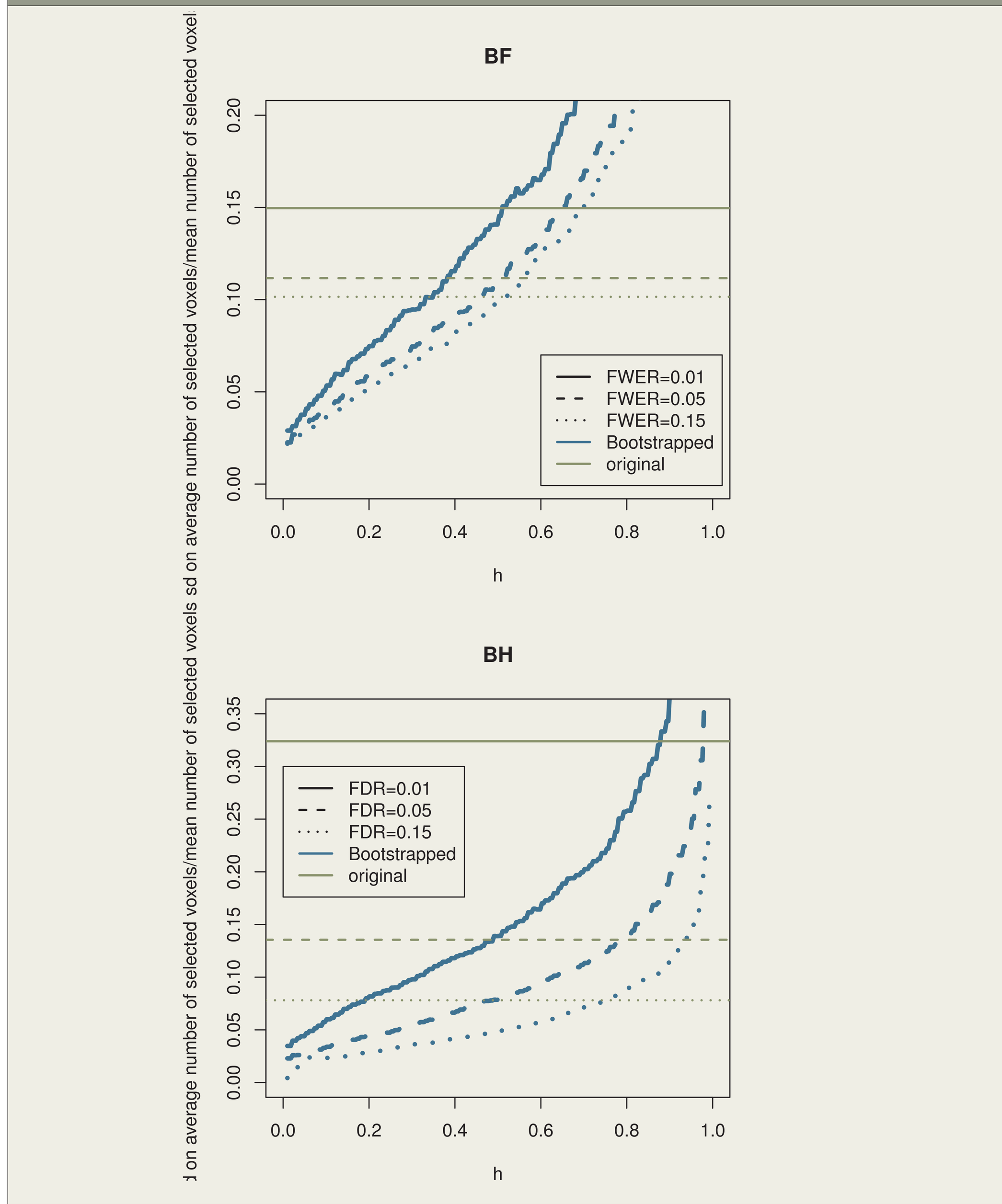


BF: Variability on test results is not reduced by including selection variability into the decision criterion.

BH:
- less stable than BF
- the stability on the test results can be strongly improved using the bootstrap procedure, up until the level of stability of BF

The role of h

h = threshold for selection frequency in bootstrap samples



The more stringent the selection criterion, the higher the variability on test results.

5 Conclusion

- We presented a bootstrap approach that can be used with any multiple testing procedure to improve selection stability.
- For BF and BH, the trade-off between power and specificity is maintained, while the stability for BH is improved
- The choice of an optimal h is a topic for further research

References

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 **52**, 1622-1629.

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

Worsley, K. J. (2003). Statistical analysis of activation images, Oxford university press, chapter 14, pp 251-270.

Corresponding author

Joke Durnez, Joke.Durnez@Ugent.be, T +32 9 264 64 34, F +32 9 264 64 87
Ghent University, Department of Data Analysis. H. Dunantlaan 1, 9000 Ghent.