Another perspective on voxel-wise multiple testing procedures in fMRI

Tom Loeys and Beatrijs Moerkerke

Department of Data Analysis, Ghent University, Belgium

Introduction

We review 3 widely used voxel-wise approaches to thresholding images of test statistics: Bonferroni (BF), Gaussian random field (GRF) and Benjamini-Hochberg (BH). While the latter controls the false discovery rate (FDR), the first two control the family-wise error rate (FWE). Comparisons of multiple testing procedures (MTP) in the neuroimaging literature have typically focused on sensitivity and specificity. However, stability (Gordon et al., 2007) is another important operating characteristic that needs to be taken into account. Here, we define stability as the variability due to the MTP in the detection of truly activated voxels.

Methods

Following Marchini and Presanis (2004), we simulated 3D Gaussian random fields using a FWHM ranging from 20 mm to 50 mm and voxel dimensions of 4 X 4 X 6 mm. To these “null” SPMs, we added positive activation by simulating an extra GRF and transforming all voxels marginally to have a Gamma(k,1) distribution (k ranging from 3 to 7). The images had dimensions 40 X 40 X 10, and of the 16000 voxels, 400 were positively activated. Each simulation setting was repeated 1000 times.

First, the performance of BF and GRF was compared at fixed theoretical levels of the FWE (ranging from 0.01 to 0.10). Next, to allow for a fair comparison of BF and GRF with BH, thresholds were determined for each procedure that result in an equal empirical FDR on average (for different levels of the FDR ranging from 0.01 to 0.10). Using these thresholds that equalized the FDR on average, we explored the number of true discoveries and its variability for each MTP.

Results

When equalizing the theoretical FWE, GRF outperforms BF in terms of mean number of true discoveries, but tends to be more variable with decreasing effect size (for the range of smoothness values considered in this simulation setting). Figures 1 and 2 show the standard deviation of the true discoveries as a function of the mean for 10 levels of the FWE (0.01 to 0.10 in steps of 0.01) for large and small effects respectively (each symbol representing a different FWE). Overall, the coefficient of variation (standard deviation divided by mean) is smaller for GRF than for BF.

When equalizing the empirical FDR, BF and GRF perform identically (as they are using exactly the same ordering of p-values). When the effect size is large (small), BH detects less (more) true activated voxels than BF=GRF, regardless of smoothness considered. In all scenarios, the variability in the number of detected voxels is larger with BH then BF=GRF. Figures 3 and 4 show the standard deviation of the true discoveries as a function of the mean for 10 levels of FDR (0.01 to 0.10 in steps of 0.01) for large and small effects respectively (each symbol representing a different FDR).Overall, the coefficient of variation is smaller for GRF=BF than BH.

Conclusions

In this simulation setting where the underlying Gaussian random field assumptions are satisfied, the stability (as measured by the coefficient of variation of the number of true discoveries) of GRF is better than BH and BF. It remains to be explored how these results are confirmed in real fMRI data.

References

Marchini, J. and Presanis, A. (2004), ‘Comparing methods of analyzing fMRI statistical parametric maps’, *NeuroImage*, vol 22, pp 1203-1213.

Gordon, A., Glazko, G., Qiu, X. and Yakovlev, A. (2007), ‘Control of the mean number of false discoveries, bonferroni and stability of multiple testing’, *The Annals of Applied Statistics*, vol 1, pp 179-190.

Figures

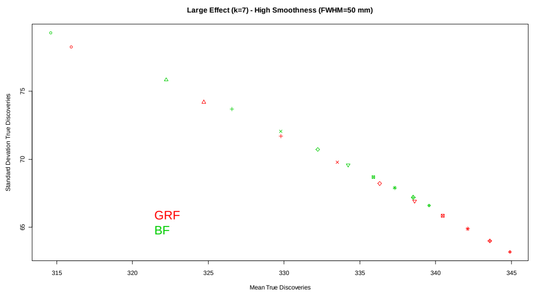

Fig 1

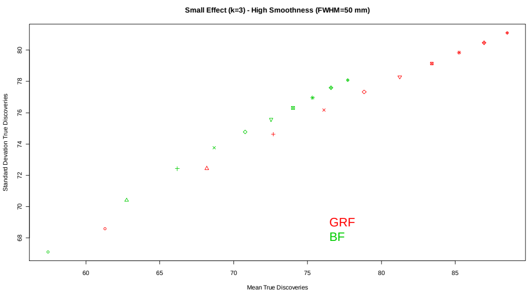

Fig 2

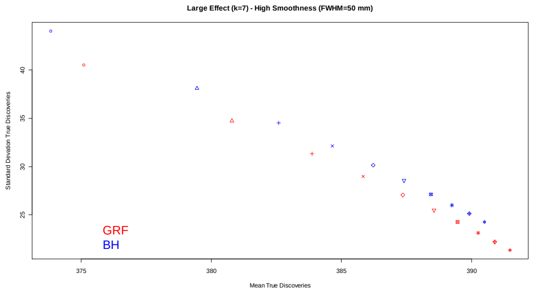

Fig 3

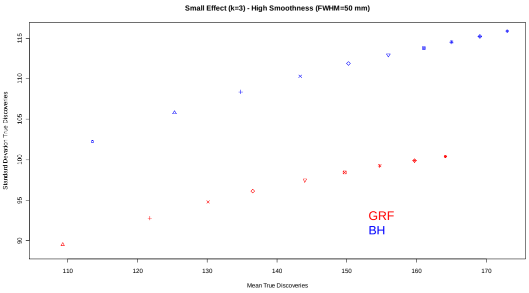

Fig 4