Power and Specificity of FDR Under Smoothness

Wei Xie and Thomas E Nichols

(weixie@umich.*edu*) (*nichols@umich.edu*)

Department of Biostatistics, the University of Michigan, Ann Arbor

NIH Neuroinformatics
Human Brain Project

Abstract & Introduction

False Discovery Rate (FDR) [1] is a measure of false positives that is more lenient than the traditional Family wise error (FWE) rate [2] and hence is more powerful. Despite its growing use in neuroimaging, there is relatively little characterization of the power and specificity of FDR, especially under the high-smoothness conditions typical in neuroimaging (though see [3]).

In this work we evaluate the sensitivity and specificity of the Benjamini & Hochberg FDR method over a range of smoothness and signal magnitudes and spatial extents. We find that the power of FDR grows with signal magnitude *and extent*, and that main impact of smoothing is a slight conservativeness, with actual FDR rates being around 0.02 instead q=0.05 for very smooth data.

Methods

FDR Method

For a given threshold u, the False Discovery Proportion is the fraction of false positives

$$FDP = V / R \, I_{\{R>0\}},$$

where R is the total number of suprathreshold voxels and V (≤ R) is the number of false detections among those voxels; $I_{\{R>0\}}$ is 1 if R>0 and 0 otherwise. Since we do not know which voxels are false detections, we control the expected FDP: FDR is defined as

$$FDR = E(FDP).$$

To find a threshold that controls FDR at level q, we use the Benjamini & Hochberg (BH) method [4] which only requires the ordered P-values and is valid under positive spatial dependence.

Control of FDX

A common mistake with FDR is to infer the number of false detections. If a q = 0.05 FDR procedure is used and R = 100 voxels are found, it is tempting to state "there were 0.05 × 100 = 5 false positives." This is incorrect, as expectation is not linear:

$$E\,(V / R \, I_{\{R>0\}}) = q \;\;\not\Rightarrow\;\; E(V) = q \times E(R).$$

Instead, methods with control of *False Discovery Exceedance* (FDX) are required [5]. Instead of an expectation, the FDX is the probability that FDP exceeds q. A valid FDX method has confidence 1-α that FDP < q, and the user can conclude that the number of false positives is no more than q × R with confidence 1-α , because

$$P\,(V / R \, I_{\{R>0\}} < q) \leq 1 - \alpha \;\;\Rightarrow\;\; P\,(V < q \times R) \leq 1 - \alpha$$

In the simulations below we measure the BH-FDR method's control of FDX, though we stress that the BH method was not designed to control FDX.

Simulated Data

We applied BH-FDR, q = 0.05, to both Gaussian images and T statistics images.

Gaussian Images **10,000 simulated datasets consisted of 32x32x32 volume images of standard Gaussian noise, smoothed with Gaussian kernels of size 0, 1.5, 3, 6 and 12 voxel FWHMs. We added spherical signals of radius 0, 1, 2, 4 and 8 voxels, comprising 0%, 0.02%, 0.10%, 0.85% and 6.64% of the volume. We considered signal magnitudes of 0.5, 1, 2, 4 & 8.**

T Images **We also generated one-sample t statistics (d.f. = 6) images from the 32x32x32 volume images of smoothed standard Gaussian noise. We again added spherical signals of radius 0, 1, 2, 4 and 8 voxels and considered signal magnitudes of 0.5, 1, 2, 4 & 8.**

Specificity Metrics

Specificity metrics describe aspects of false positive control. We used FDR and FDX; if FDR exceeds q = 0.05 the method is invalid, and if less than q = 0.05 it is not exact. To be more comparable with FDR, we plot 1 − FDX or P (V / R $I_{\{R>0\}}$ > q); FDX is controlled if 1 − FDX is less than α = 0.05.

Sensitivity/Power Metrics

In the multiple testing setting, there is no one unique measure of power. We considered three:

Average Power

The probability of detecting a given voxel with a (non-null) signal. Equivalently, the chance, averaged over all signal voxels, of a detection.

Familywise Power

The probability of detecting a one or more signal voxels.

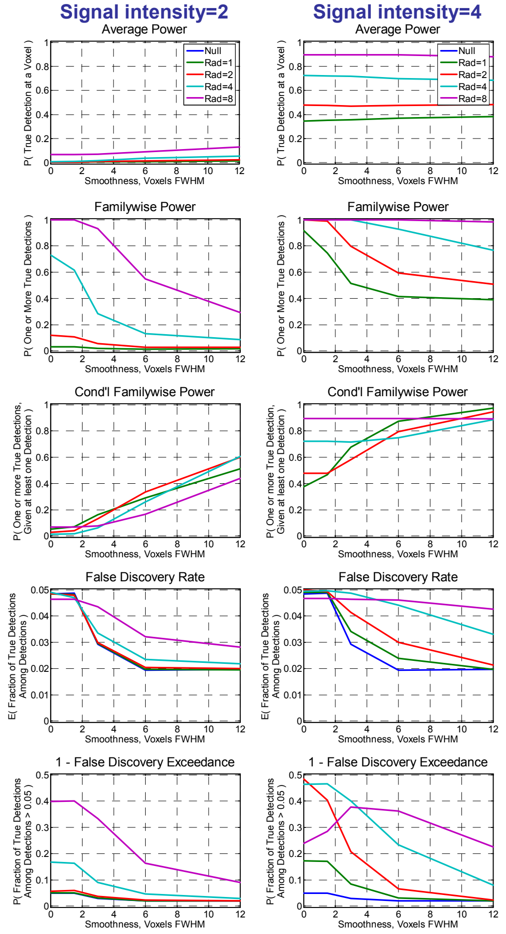
Conditional Familywise Power

Conditional on detecting at least one voxel at all, the probability of detecting one or more signal voxels.

While Average Power is the typical definition, users may feel that detection of *one or more* true signal voxels is more important; Familywise Power measures this. Another concern is when there are no detections at all (R=0); assuming we discard all such data, we might ask what the Familywise Power is when only considering statistic images with one or more detections; this is the Conditional Familywise Power.

Results

Plots below show the BH-FDR's performance in 3 power measures and the 2 specificity measures. For reasons of space we only show results for signal magnitude 2 & 4 (Radius=0 show the null results). Gaussian and T results were qualitatively similar and so have omitted plots for T.



Results Summary

		Sensitivity (Power) Metrics			Specificity Metrics	
		Avg. Pw.	Fw. Power	Cnd'l Fw. Pw.	FDR	1-FDX
↑	Signal Magnitude	▲	▲	▲	—	—
↑	Signal Extent	▲	▲	▴	▴	▴
↑	Smoothness	—	▾	▲	▼	▴ ▾

Conclusion

Our evaluations revealed some loss in specificity (FDR) with increasing smoothness, though it was not dramatic (FDR as low as 0.02 when q=0.05). Surprisingly, in low-intensity, small-extent signal settings, the BH-FDR method controlled FDX, though in many settings it exceeded 0.05. This points to the danger of using q × R to estimate the number of false positives V, and indicates that control of expected FDP (i.e. FDR) doesn't rule out results with high proportions of false discoveries.

As is to be expected, power always increases with signal magnitude. But unique to FDR, Average Power also increases with signal extent: the more voxels with signal, the greater the chance of any one signal voxel being detected.

We find that Familywise Power *falls* with increased smoothness while Average Power *rises*. This is a seeming contradiction, but can be explained by noting that Conditional Familywise Power rises with smoothness: As smoothness increases, you have less chance of any detections, but when you do detect any you'll find a lot of true positives.

In summary we find that, while BH-FDR becomes slightly conservative with increasing smoothness, this conservativeness is overcome by increases in Average Power. Thus it is a valid and powerful method suitable method for smooth neuroimaging data exhibiting positive dependence.

References & Acknowledgements

[1]Genovese, Lazar, Nichols, 2002, NI 15:870-78. [2]Nichols & Hayasaka, 2004, StatMethMedRes 12:419-46. [3]Logan & Rowe, 2004, NI 22:95-108. [4]Benjamini & Hochberg, 1995, JRSS-B 57:289-300. [5]Genovese & Wasserman, (2002). JRSS-B, 64: 499-518.

Supported by NIH's Human Brain Project. Thanks to Xiaobi Huang for help with poster