A COMPARISON OF THRESHOLDING METHODS FOR STATISTICAL PARAMETRIC MAPS

ANDERSON MARCELO WINKLER* HUMBERTO REMIGIO GAMBA* JOÃO RICARDO SATO ARNOLFO DE CARVALHO NETO*****
*** Imaging and Electronic Instrumentation Laboratory (LABIEM), Federal Technological University of Paraná (UTFPR), Curitiba, Brazil**
**** Institute of Mathematics and Statistics, University of São Paulo (USP), São Paulo, Brazil**
***** Discipline of Medical Radiology, Department of Internal Medicine, Federal University of Paraná (UFPR), Curitiba, Brazil**
www.labiem.cpgei.cefetpr.br - andersonwinkler@gmail.com

1. Introduction

The most common strategy to analyse fMRI data (though not necessarily optimum) is to:

- Apply transforms to the images, such as affine transformations, warpings, phase shifts and smoothing to reduce some potentially confounding effects;
- Perform a massive, voxel-wise statistical analysis of the data, fitting the obtained response to a predefined model of haemodynamic response. The estimated parameters are tested against a null hypothesis, resulting in a statistic image (sometimes known as statistical parametric map);
- Set a threshold. Voxels which statistical score is above the threshold are labeled as "active"; the remaining are labeled as "inactive".

Type I errors may happen very often if such a large amount of tests are performed. Traditional methods for dealing with this problem, like Bonferroni correction, are considered too conservative. Though this is a central problem for neuroimaging studies, the best approach is still unclear.

Two methods have emerged as the most suitable for fMRI. The *random field theory* (RFT) has become the de facto standard method for controlling the family-wise error rate (FWE), despite its complexity and restrictive assumptions. When these assumptions are met, RFT usually produces less stringent thresholds than Bonferroni. However, if the researcher is willing to accept some false-positives within the image, methods for controlling the *false discovery rate* (FDR), as the B&H procedure, can provide even more liberal thresholds, with minimal assumptions, still retaining control over an error measure.

The objective of this study is to evaluate the performance of Bonferroni correction, B&H procedure and RFT for fMRI studies.

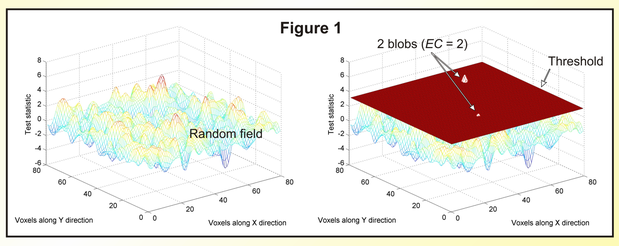
2. The random field theory (RFT)

The RFT states that the number of surviving "peaks" in a thresholded random field can be estimated if the probability density function, the smoothness of the field and the threshold itself are known (Figure 1). This measure, equivalent to the Euler characteristic (*EC*), can be well approximated to the probability of obtaining any false-positive in a given statistical parametric map (i.e. the family-wise error rate - FWE). To calculate the threshold, the general equation is:

$$p_{\mathrm{FWE}} = EC \approx \sum_{d=0}^{D} r_d(S)\rho_d(t)$$

where $r_d(S)$ is the resel count (a measure of smoothness) for the search region S, and $\rho_d(t)$ is the Euler characteristic density, which depends on the probability density function of the field and on the threshold t itself. Usually, the EC is set to desired FWE level (usually $p_{\mathrm{FWE}} = 0.05$), and the threshold t is then estimated.

Figure 1



3. The false discovery rate (FDR)

FDR is the proportion of incorrect rejections of the null hypotesis (false positives) among those tests where the null hypothesis was rejected. This proportion is the maximum amount of false positives that the researcher is willing to accept. A procedure to control the FDR at level q consists in calculate the uncorrected p-value for each voxel and order them so that the ordered p-values are $p_1 \leq p_2 \leq \ldots \leq p_V$. Then, find the largest i so that $p_i \leq (i.q)/(V.c(V))$. The voxels $v_1, \ldots, v_i$ are declared active. This is the Benjamini and Hochberg (B&H) procedure. The Figure 2 (left) shows a graphical perspective; the graph is zoomed (right) to show the threshold in greater detail.

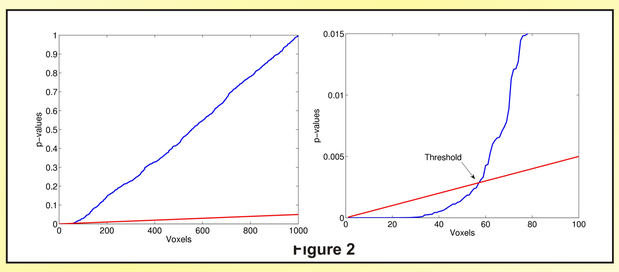


Figure 2

4. Epidemiological viewpoint

Solving for the multiple comparisons problem is rather similar to the epidemiological problem of finding the ideal cut-off value for diagnostic tests. In both cases, the results of the individual tests can be summarised in a contingency table:

	Truly active	*Truly inactive*
Declared active	Truly positive (A)	Falsely positive (B)
Declared inactive	Falsely negative (C)	Truly negative (D)

From the table, the following terms can be defined, and used to evaluate the performance of the methods:

- Sensitivity (S): A/(A+C)
- Specificity (E): D/(B+D)
- Positive predictive value (PPV): A/(A+B)
- Negative predictive value (NPV): D/(C+D)
- False discovery rate (FDR): B/A+B = 1-PPV
- Accuracy = (A+D)/(A+B+C+D)

5. Method

A real "null" fMRI dataset was acquired using a GE Signa Excite 1.5 T scanner (TR = 2000 ms, TE = 50 ms, FOV = 240 mm, matrix dimensions = 64 x 64 x 24, voxel size 3.75 x 3.75 x 5.00 mm, gap = 0 mm, interleaved acquisition, 240 volumes). The volumes were realigned and an ideal temporal high-pass filter was applied (cutoff = 128 s). The scans were randomly permuted to minimise the potential bias due to the temporal autocorrelation structure.

Patches of boxcar-like "activation" were added using the canonical haemodynamic response function, which parameters were slightly variable for each "activation" period. The amplitude of this simulated BOLD response was not the same across different patches (Figure 3, percentage of the average signal).

The general linear model was applied to both rest and with added "activation" datasets, and with and without spatial smoothing (FWHM = 2 voxels on each direction). The variance was restored after smoothing. Estimation of the smoothness was based on the residuals of the model fit.

For each of these four conditions, 1500 t-maps were generated, which were thresholded using the Bonferroni (BON), RFT and B&H approaches, as well without correction (UNC). The significance level for thresholding was $\alpha_{\mathrm{PCE}} = \alpha_{\mathrm{FWE}} = q_{\mathrm{FDR}} = 0.05$.

6. Results

In the absence of signal, for all the evaluated methods, the observed error measures were within the theory. The FWE control for Bonferroni were not as conservative as often regarded in the literature, possibly because spatial autocorrelation were not an issue in the dataset used for the simulations (Table 1A). The B&H procedure resulted in control over FWE in this settings (Table 1A). For non-smoothed images, the FWE control using RFT resulted in conservative results, while for smoothed images, the observed FWE was very close to the nominal α_{FWE} level (Tables 1A and 1B).

In the presence of signal, for non-smoothed images (Table 2A), the B&H procedure resulted in more power (60.1% vs. 42.7%) and higher accuracy (97.4% vs. 96.5%) than Bonferroni. The RFT resulted in the lowest power, possibly due to violation of the assumption of adequate lattice representation of a continuous underlying random field. The B&H procedure resulted in indirect control over the PPV, which, in many circumstances, is exactly what the researcher might want.

Smoothing the images before generation of maps resulted in blurred areas containing signal, thus preventing adequate counting of the number of voxels in the contingency table (Table 2B). Notwithstanding, the B&H was still more powerful, despite lower power, specificity and accuracy than BON and RFT. The latter provided very similar results, as can be noticed by the very close threshold values.

Table 1

(A) No signal / Non-smoothed

	Threshold	oPCE	oFWE
UNC	1.6513±0.0000	0.0514±0.0178	1.0000
BON	4.6961±0.0000	0.0000±0.0000	0.0507
B&H	4.6954±0.3277	0.0000±0.0001	0.0393
RFT	4.8714±0.0000	0.0000±0.0000	0.0320

(B) No signal / Smoothed (FWHM = 2.0 vox)

	Threshold	oPCE	oFWE
UNC	1.6513±0.0000	0.0545±0.0461	1.0000
BON	4.6961±0.0000	0.0000±0.0000	0.0400
B&H	4.2193±0.6294	0.0003±0.0041	0.0467
RFT	4.6413±0.0002	0.0000±0.0001	0.0480

Table 2

(A) Simulated signal / Non-smoothed

	Threshold	oPCE	oFDR	Sensitivity	Specificity	Accuracy	PPV	NPV
UNC	1.6513±0.0000	0.0467±0.0160	0.4908±0.0716	0.7619±0.0259	0.9503±0.0170	0.9389±0.0152	0.5092±0.0716	0.9842±0.0016
BON	4.6961±0.0000	0.0000±0.0000	0.0001±0.0005	0.4266±0.0216	1.0000±0.0000	0.9653±0.0013	0.9999±0.0005	0.9644±0.0013
B&H	2.9260±0.0239	0.0019±0.0018	0.0470±0.0357	0.6009±0.0267	0.9980±0.0019	0.9740±0.0018	0.9530±0.0357	0.9749±0.0016
RFT	4.8713±0.0000	0.0000±0.0000	0.0000±0.0003	0.4117±0.0216	1.0000±0.0000	0.9644±0.0013	1.0000±0.0003	0.9635±0.0013

(B) Simulated signal / Smoothed (FWHM = 2 voxels)

	Threshold	oPCE	oFDR	Sensitivity	Specificity	Accuracy	PPV	NPV
UNC	1.6513±0.0000	0.0834±0.0439	0.5682±0.0992	0.9274±0.0276	0.9112±0.0467	0.9122±0.0430	0.4318±0.0992	0.9949±0.0018
BON	4.6961±0.0000	0.0131±0.0018	0.2367±0.0187	0.6940±0.0326	0.9861±0.0019	0.9684±0.0013	0.7633±0.0187	0.9804±0.0020
B&H	2.6759±0.0516	0.0297±0.0131	0.3559±0.0626	0.8500±0.0366	0.9684±0.0139	0.9612±0.0119	0.6441±0.0626	0.9902±0.0023
RFT	4.6411±0.0003	0.0133±0.0018	0.2390±0.0186	0.6979±0.0326	0.9858±0.0019	0.9684±0.0013	0.7610±0.0186	0.9807±0.0020

7. Conclusions

The images must be smoothed if the RFT is to be used, otherwise the method may become too conservative.

On the other hand, smoothing should be avoided if the researcher is planning to control FDR or if there is a need to count the number of "active" voxels, as the smoothing blur the data.

For fMRI non-smoothed data, Bonferroni can provide trustworthy results, despite its lower power.

References

General

- Nichols TE and Hayasaka S. Controlling the familywise error rate in functional neuroimaging: a comparative review. *Stat. Met. Med. Res.*, 12 (2003) 419-446;
- Marchini J and Presanis A. Comparing methods of analyzing statistical parametric maps. *NeuroImage*, 22 (2004) 1203-1213.

Random field theory

- Adler RJ. *The geometry of random fields*. New York, Wiley, 1981;
- Worsley KJ, Marrett S, Neelin P, Vandal AC, Friston KJ and Evans AC. A unified statistical approach for determining significant signals in images of cerebral activation. *Hum. Brain Mapp.*, 4 (1996) 58-73.

False discovery rate

- Benjamini Y and Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, 57 (1995) 289-300;
- Genovese CR, Lazar NA and Nichols TE. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage*, 15 (2002) 870-878.