Decision Thresholds in Functional MR Image Analysis

Michelle Liou[1], Hong-Ren Su[1,2], Arthur C. Tsai[1],

[1] Institute of Statistical Science, Academia Sinica,
128, Academia Rd. Sec.2, Taipei 115, Taiwan
[2] Dept. of Computer Science, National Tsing Hua University, Hsinchu, Taiwan
101, Sec.2, Kuang-Fu Rd., Hsinchu, 300 Taiwan

email: {mliou, stevensu, arthur}@stat.sinica.edu.tw

Abstract. In functional magnetic resonance imaging (fMRI) studies, statistical parametric maps (SPMs) plots in an anatomical background those voxels exceeding a *p*-value threshold. In this study, we explore some limitations of a sole use of *p*-values including the false discovering rate (FDR) control over experimental-wise error rates. As a comparison, we also include the receiver-operator characteristic (ROC) curve approach for finding the optimal decision threshold. In a real data example, we apply different methods to analyzing the same data set. The results show that both p-value and ROC approaches suggest similar findings when functional images are less contaminated by noise. The FDR control of experimental-wise errors is sensitive to the proportion of voxels being classified as active. We finally discuss the use of different approaches for analyzing fMRI data.

Keywords: FDR**;** Reproducibility analysis; ROC; SPM .

1. Introduction

Research findings in functional magnetic resonance imaging (fMRI) studies are normally summarized using statistical parametric maps (SPMs) which highlight in an anatomical background those voxels exceeding a *p*-value threshold (e.g., $p < 0.05$). Brain voxels with small *p*-values are not just more responsive to experimental stimuli as compared with a control condition, they are also much greater in response magnitude. Empirical studies have shown that there are functional regions consistently engaged in experimental tasks with smaller response amplitude, however. It is important to acquire other thresholding methods which are sensitive to activity in those regions, but less dependent on the response magnitude. In this study we formally contrast and compare between *p*-value approach and the receiver-operator characteristic (ROC) curve method for finding the optimal decision threshold for classifying image voxels into the active/inactive status. In the empirical example, we consider a data set which was collected for studying the representation of objects in the human occipital and temporal regions via an on-and-off paradigm (Ishai et al., 2000). The error rates of different thresholding methods will be computed based on the empirical data set. In the next section, we will discuss the background behind the *p*-value thresholds including approaches for controlling the family-wise error rate (FWER). The ROC thresholding method will also be introduced. Based on the aforementioned data set, we will show that different thresholding methods will give convergent results when image data are less contaminated by noise. For noisy images, however, the *p*-value approach will bypass many important activities. The ROC approach along with a reproducibility criterion, on the other hand, offers enough findings that go beyond those obtained by the SPM approach. We finally discuss the use of different methods for analyzing fMRI data.

2. Method

fMRI experiments are usually performed over a period of time and are divided into smaller experimental runs to allow subjects taking a rest between runs. In this section, we will briefly review methods for computing decision statistics in fMRI data analyses, and compare the different thresholding methods for assigning voxels to the active/inactive status.

978-0-7695-3118-2/08 $25.00 © 2008 IEEE
DOI 10.1109/BMEI.2008.297

IEEE computer society

2.1 Decision statistics

Assume that the image data are pre-whitened by removing autocorrelations and other artifacts. In the SPM generalized linear model, the fMRI responses in the ith run can be expressed as

$$y_i = X_i \beta_i + e_i ,$$

where y_i is the vector of pre-whitened image intensity in a particular voxel; X_i is the design matrix decided by the stimulus presentation, and β_i is the vector containing the unknown regression parameters. With a random effect model, the regression parameters β_i are additionally assumed to be random from a multivariate Gaussian distribution with common mean μ and variance Ω. The empirical Bayes estimate of β_i in the model will shrink all estimates toward the mean μ, with greater shrinkage at noisy runs. By analogy to the generalized linear model, *t*-values of a particular contrast within runs can be computed by normalizing the estimated $\hat{\beta}_i$ using the standard errors of e_i. For each design contrast, there are *M* such *t*-values, and *M* is the total number of runs. The overall *T*-values can also be computed for parameters in μ using the corresponding standard errors in Ω.

2.2 Thresholding methods

If the true active/inactive status is known, and a decision threshold k^* is available, the voxel-wise *T* values can be grouped into a 2x2 table (see Table 1). In the table, λ denotes the proportion of truly active voxels. The proportion of correct classification is $P_o = a + d$, and its expected value is $P_c = (a+c)\lambda + (b+d)(1-\lambda)$. In the literature, the Kappa index (Cohen 1960), false discovering rate (FDR), FWER, and Type-I error (or false alarm) are defined respectively as Kappa = $(P_o - P_c)/(1-P_c)$, FDR = *c/(a+c)*, FWER = *c*, and Type-I error = $c/(1-\lambda)$. Given an experimental contrast, the exceeding probability (*p*-value) of a *T*-value can be computed using a student *t*-distribution with specified degrees of freedom (i.e., *M* minus the number of unknowns associated with μ). The *p*-values of in-brain voxels can be ordered from the most to least significance. The SPM threshold is a cut-off point on the ordered *p*-values such that anything below the point is classified as active. The easiest way would be classifying any voxel having $p \le 0.05$ as active. In the fMRI literature, statistical issues associated with thresholding methods mainly concern with the control of the FWER. As *p*-values are weakly correlated with each other, the Bonferroni correction classifies $p \le 0.05/V$ as active for controlling the FWER at $\alpha = 0.05$, where *V* is the total number of voxels considered. The correction becomes too stringent as V increases, and the sensitivity of statistical tests deteriorates.

There are several sharper Bonferroni procedures proposed, such as the FDR control proposed by Benjamini and Hochberg (1995). Let $p^{(1)} < p^{(2)} < \cdots < p^{(V)}$ be the ordered sequence of *p*-values from the most to least significance. In the proposed FDR control, a *j*-th voxel with $p^{(j)}$ in the sequence is classified as active if $p^{(j)} \le j\frac{\alpha}{V}$. This procedure was originally introduced by Simes (1986) for a weak control of the FWER at the α level (e.g., $\alpha = 0.05$), and theoretically proven by Benjamini and Hochberg of its equivalence to controlling the FDR at the α level. The procedure can also be viewed as a maximization procedure to enlarge *a+c* in Table 1 as much as possible, and at the same time, to constraint *c/(a+c)* within α. It is true that the FDR is identical to FWER when λ is zero, and becomes more powerful as λ increases to one. In a sense, the FDR control is more sensitive to active voxels than the conventional Bonferroni correction when the proportion of truly active voxels is greater compared with the proportion of truly inactive voxels. In most of the fMRI studies, however, this is unlikely to occur because λ is often time a smaller proportion (e.g., less than 0.2 in our empirical study). Therefore, the FDR control might not perform better as desired (Nichols and Hayasaka 2003). In the empirical study, we will show that controlling the family-wise error rate with either the Bonferroni or

FDR procedures enlarges the Type-II error rate (b/λ in Table 1), and makes the *p*-value approach less sensitive to active voxels

The true status of each voxel is unknown, but can be estimated using the *t*-values within runs derived from the random effect model. If we select *K* distinct thresholds in increasing order of magnitude, these *t*-values (in absolute value) of size *M* can be classified into *K+1* groups (e.g., *K*=10 in our empirical study). The count of *t*-values in each group is shown in Table 2. In the table, P_{A_k} denotes the conditional probability of *t* values assigned to the *k*-th group given the truly active status, and P_{I_k} carries the same definition given the truly inactive status. Both are unknown parameters in the table with observed counts γ_k. If a threshold k^* is selected, sensitivity and the false alarm rate are defined respectively as $P_A \equiv \sum_{k=k^*}^{K} P_{A_k}$ and $P_I \equiv \sum_{k=k^*}^{K} P_{I_k}$ (Refer to Table 1). The ROC curve is a bivariate plot of P_A and P_I across all possible thresholds. By considering all in-brain voxels, the unknown parameters can be estimated by maximizing the likelihood of observing the patterns of γ_k in a mixed multinomial model (the two conditional distributions given the true status in Table 2). When maximizing the likelihood, we also assume a prior distribution for the mixing proportion λ. After estimating *K* pairs of (P_A, P_I), the ROC curve can be interpolated via a smoothed function. In this study, we consider the decision threshold k^* which maximizes the Kappa value on the ROC curve. Unlike the *p*-value approach relying on the overall *T*-values, the ROC method classifies a within run *t*-value as active if its value is greater than or equal to k^*. In the empirical example, we will show that a voxel is classified as active if the *t*-values of *M* runs suggest that the decision is strongly reproducible across runs.

3. Empirical Example

In the empirical example, we consider a data set with six subjects involved, each performing twelve runs of a delayed match-to-sample task with either photographs or line drawings (Ishai et al., 2000). In the task, a target stimulus (houses, faces or chairs) was followed, after a 0.5 sec delay, by a pair of choice stimuli presented at a rate of 2 sec. Subjects indicated which choice stimuli matched the target by pressing a button with the right or left thumb. All runs involved phased, scrambled pictures presented at the same rate as the control stimuli. In this study, we inserted three orthogonal contrasts in the design matrix of the generalized linear model -- namely, meaningful objects (i.e., faces, houses and chairs) versus the control condition (i.e., phased, scrambled pictures), faces versus houses/chairs, and houses versus chairs.

In the data analysis, the effects due to different contrasts for each run along with the average effect across runs were computed using the random effect model for each individual subject. Table 3 gives the maximum Kappa value, and the corresponding estimated FDR and Type-I error rate for each subject and each design contrast. In this data example, the Kappa values are higher when comparing meaningful objects with the control condition. Also, Type-I errors range from 0.03 to 0.04, and FDRs range from 0.23 to 0.37 for this contrast. When comparing between objects (e.g., faces versus houses/chairs), the Kappa values are reduced to a range of 0.26 to 0.32, and the FDRs increase to as high as 0.58. As was indicated, the proportion of truly inactive voxels (1 – λ) is generally much greater than that of truly active voxel λ. In real applications, the total number of voxels that can be classified as active (i.e., a+c) must be extremely small in order to control the FDR at the .05 or 0.10 level.

By maximizing Kappa for selecting the decision threshold, the size of Type-I errors can still be controlled within a reasonable range (i.e., 0.03-0.10). Distributions of the averaged *T*-values across subjects are plotted in Figure 1 for comparing meaningful objects with the control condition. Here we define a voxel to be strongly reproducible if its active status remains the same in at least 90% of runs, and moderately reproducible in 70-90% of the runs. The averaged *T*-values are plotted separately for strongly and moderately reproducible voxels. As a comparison, we also plot the average values for those voxels consistently classified as inactive across runs.

It is interesting to note that strongly and moderately reproducible voxels have a sizable overlap in their *T*-values. Neuroimaging research has recently suggested that the precuneus is tonically active in a resting state and deactivated when subjects are engaged in a wide variety of cognitive tasks (Raichle et al. 2001). The *T*-values of voxels in the precuneus showing increased and decreased responses in the matching task are also plotted in Figure 1. On average, the magnitude of *T*-values does not directly imply reproducibility. The activation status using a single cutoff point (*p*-value) on the T-values could bypass many strongly reproducible findings. It is also interesting to note that the Type-I errors for each design contrast differ only within a range of 0.01 to 0.02 even though images of individual subjects are noisy to a great and less degree.

The activation maps for comparing objects with the control condition are plotted in Figure 2 for Subjects 2 and 3. Functional images of Subject 3 are less contaminated by noise, and give the highest Kappa value relative to other subjects in the same experiment. The activation maps of this subject using the strongly reproducible criterion are clearly visible. For this subject, moderately reproducible voxels are mainly distributed in the neighborhood of strongly reproducible voxels. Subject 2 has the lowest Kappa value, although moderately reproducible voxels are still spatially closer to the strongly reproducible voxels except for a few regions in the cerebellum. The ensuing SPMs given a T-value threshold closely resemble those strongly reproducible maps when subjects are less contaminated by noise. In the figure, there is essentially no activation region in the SPMs when controlling the FDR at .05 for subjects with either the high or low degree of noise contamination.

4. Discussion

The ROC method along with the strongly reproducible criterion have been designed to maximize the between run reproducibility via the random effect model. Although the threshold selected by maximizing the Kappa value can control the empirical Type-I error within a reasonable range, the FDR suggested that there are a sizable false positive hits among those voxels being classified as active. By counting on the strongly reproducible criterion, the method still preserves enough true positive voxels and bypasses those false positive. This has been the strength of using the ROC approach and the strongly reproducible criterion. The FDR is sensitive to the ratio between λ and (1- λ). In applications, a control of FDR at the 0.05 or 0.10 level should be too stringent to find any active voxels. Although there have been many new suggestions for improving the original FDR proposed by Benjamini and Hochberg, the results in Figure 2 suggest that both SPMs without controlling the FWER and reproducibility maps closely resemble each other when functional images are less contaminated by noise.

References

1. Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, B 57, 289-300.
2. Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20, 37-46.
3. Ishai, A., L. G. Ungerleider, A. Martin, and J. V. Haxby (2000). The representation of objects in the human occipital and temporal cortex. Journal of Cognitive Neuroscience 12:S2, 35-51.
4. Nichols, T. and S. Hayasaka (2003). Controlling the family-wise error rate in functional neuroimaging: a comparative review. Statistical Methods in Medical Research 12, 419-446.
5. Raichle, M. E., A. M. MacLeod, A. Z. Snyder, W. J. Powers, D. A. Gusnard, and G. L. Shulman (2001). A default mode of brain function. Proceedings of the National Academy of Sciences, USA 98, 676-682.
6. Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. Biometrika 73, 751-754.

Acknowledgments. The authors are indebted to the fMRIDC at Dartmouth College for supporting the datasets analyzed in this study. This research was supported by the grant NSC 94-2413-H-001-001 from the National Science Council (Taiwan).

Table 1: Cross-classification of voxels given a decision threshold.

Decision

		Inactive	Active	
True Status	Active	b = λ(1 - P_A)	a = λP_A	λ
	Inactive	d = (1 – λ)(1 – P_I)	c = (1 – λ)P_I	1 - λ
		b+d	a+c	a+b+c+d=1

Table 2: Observed counts and conditional probabilities in the ROC analysis.

Decision

		← Inactive			Active →		
		0	1	2	…..	K	
True Status	Active	P_{A_0}	P_{A_1}	P_{A_2}	…..	P_{A_k}	$\sum_{k=0}^{K} P_{A_k} = 1$
	Inactive	P_{I_0}	P_{I_1}	P_{I_2}	…..	P_{I_k}	$\sum_{k=0}^{K} P_{I_k} = 1$
	Counts	γ_0	γ_1	γ_2	…..	γ_k	$\sum_{k=0}^{K} \gamma_k = M$

Table 3: Kappa and error rates associated with different design contrasts.

	Objects vs. Control			Faces vs. Houses/Chairs			Houses vs. Chairs		
Subj.	Kappa	FDR	Type-I	Kappa	FDR	Type-I	Kappa	FDR	Type-I
1	0.54	0.25	0.04	0.26	0.58	0.10	0.26	0.57	0.10
2	0.43	0.37	0.06	0.27	0.55	0.09	0.26	0.58	0.10
3	0.63	0.18	0.03	0.32	0.50	0.08	0.29	0.52	0.09
4	0.59	0.23	0.04	0.28	0.55	0.10	0.29	0.52	0.09
5	0.54	0.25	0.04	0.27	0.55	0.10	0.27	0.54	0.10
6	0.58	0.25	0.03	0.29	0.53	0.08	0.30	0.52	0.09

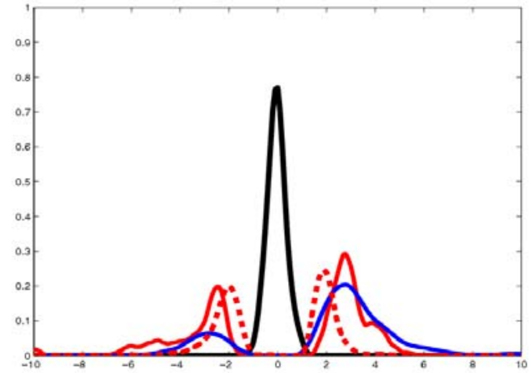

Figure 1: The density distributions of average T values across the 6 subjects for comparing the meaningful objects with the control condition. The area under different distributions is normalized to have the same value of one. The average values for voxels consistently classified as inactive in the 12 runs are plotted in black; those consistently classified as active in the 12 runs are plotted in red; voxels classified as moderately reproducible in the 12 runs (i.e., 8 to 10 runs) are plotted as dotted line in the figures. Those strongly reproducible voxels located in the precuneus with either positive or negative T values are plotted in blue.

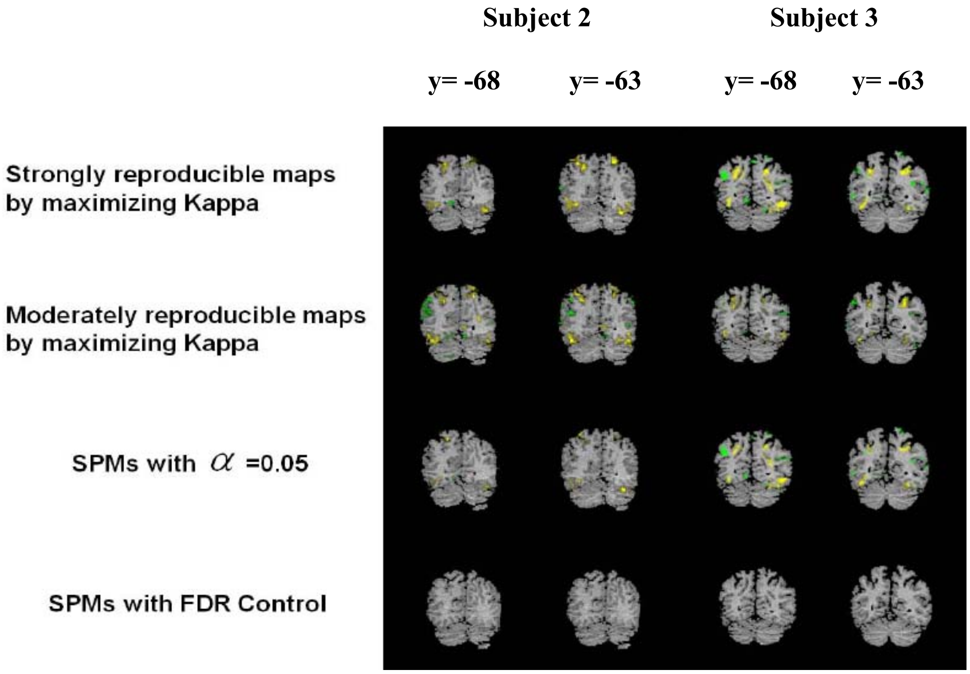
Subject 2
Subject 3
y= -68
y= -63
y= -68
y= -63
Strongly reproducible maps by maximizing Kappa
Moderately reproducible maps by maximizing Kappa
SPMs with α =0.05
SPMs with FDR Control

Figure 2: The activation maps for comparing meaningful objects with phased scrambled photographs for Subjects 2 and 3 in the Ishai et al. study. Colored voxels in yellow have positive T values and those in green have negative T values.