

# Combining Voxel Intensity and Cluster Extent with Permutation Test Framework

Satoru Hayasaka, Thomas E Nichols

*Department of Biostatistics, The University of Michigan, Ann Arbor, MI, USA*

March 26, 2004

*Running title: Intensity-Extent Combined Inference*

*Keywords: permutation test, cluster size*

Address for correspondence:

Thomas E. Nichols

Department of Biostatistics

University of Michigan

1420 Washington Height,

Ann Arbor, MI 48109

Phone: +1-734-936-1002

Fax: +1-734-763-2215

email: nichols@umich.edu

## Abstract

In a massively univariate analysis of brain image data, statistical inference is typically based on intensity or spatial extent of signals. Voxel intensity-based tests provide great sensitivity for high intensity signals, whereas cluster extent-based tests are sensitive to spatially extended signals. To benefit from the strength of both, the intensity and extent information needs to be combined. Various ways of combining voxel intensity and cluster extent are possible, and a few such combining methods have been proposed. Poline *et al.*'s (1997) minimum p-value approach is sensitive to signals whose either intensity or extent is significant. Bullmore *et al.*'s (1999) cluster mass method can detect signals whose intensity and extent are sufficiently large, even when they are not significant by intensity or extent alone. In this work, we study such combined inference methods using combining functions (Pesarin, 2001) and permutation framework (Holmes *et al.*, 1996), which allow us to examine different ways of combining voxel intensity and cluster extent information without knowing their distribution. We also attempt to calibrate combined inference by using weighted combining functions, which adjust the test according to signals of interest. Furthermore, we propose meta-combining, a combining function of combining functions, which integrates strengths of multiple combining functions into a single statistic. We found that combined tests are able to detect signals which are not detected by voxel or cluster size test alone. We also found that the weighted combining functions can calibrate the combined test according to the signals of interest, emphasizing either intensity or extent as appropriate. Though not necessarily sensitive than individual combining functions, the meta-combining function is sensitive to all types of signals, thus can be used as a single test summarizing all the combining functions.

# 1 Introduction

In a massively univariate method in brain image analysis, a linear regression model is fitted at each voxel, then a statistic image for a contrast of interest is calculated, and finally the significance of the effect of interest is assessed using various inference methods. Widely used are voxel intensity tests and cluster size tests (Petersson *et al.*, 1999). In a voxel test, statistical significance is based on intensity of a signal at each voxel, whereas in a cluster size test, the significance is based on the spatial extent or size of signals. A voxel test can be powerful for localized high intensity signals, while a cluster size test can be sensitive to spatially extended signals (Friston *et al.*, 1996; Poline *et al.*, 1997). Either test is sensitive to a specific type of signals, but if these two tests are combined, then the resulting test can be sensitive to both localized high intensity signals and spatially extended signals.

Tests have been proposed which combine voxel intensity and cluster size information. Poline *et al.* (1997) developed a combined test based on Gaussian random field theory (RFT). In their approach, the critical region is sought using the minimum p-value of an RFT peak intensity test (Adler, 1980; Worsley *et al.*, 1992) and an RFT cluster size test (Friston *et al.*, 1994), and the joint distribution of the peak intensity and the cluster size according to RFT. Their use of the minimum p-value results in a test that is sensitive to signals with either high intensity or large spatial extent. However, stringent assumptions in this approach makes this test less practical. In addition to usual RFT assumptions of smooth images and high threshold, this approach assumes that the spatial auto-correlation function is Gaussian in order to derive the joint intensity-cluster size distribution. Furthermore, this method is only applicable to Gaussian images, thus for a  $t$  image, a  $t$ -to- $Z$  transformation is required. A less stringent approach in combining intensity and cluster size was developed by Bullmore *et al.* (1999). In their approach, for each cluster, cluster mass is calculated as the integration of voxel intensities above the cluster defining threshold, and the maximum cluster mass is used as a test statistic in a permutation test in place of the maximum intensity or cluster size. Because this method uses permutations rather than a theory-based approach, it requires less

assumptions.

In the above methods of combining voxel intensity and cluster size information, they both have their own strengths and weaknesses. Poline *et al.*'s minimum p-value approach is sensitive for signals with either high intensity or large extent, not necessarily both: The method only uses information from the test producing a smaller p-value. For example, if the cluster size test produces a smaller p-value, say  $p=0.002$ , then the peak intensity p-value has no influence whatsoever on the outcome of the combined test, whether it is 0.005 or 0.5. Furthermore, this test needs to correct for two tests, an intensity test and a cluster size test, thus reducing its sensitivity. While Poline *et al.*'s test works only when the intensity or the cluster size is significantly large, Bullmore *et al.*'s cluster mass statistic can produce significant results when both intensity and extent are marginally significant, but not necessarily significant on their own. However, this cluster mass method is not consistent theoretically (Pesarin, 2001). That is, even if either voxel or cluster size test produces arbitrarily significant results, the rejection is not guaranteed, and depends on the significance of the other test. This implies that the test is sensitive to signals with high intensity AND large cluster extent at the same time, not just one of them.

Of course, depending on signals of interest, either of the combining methods above can be useful. Moreover, there are other possible ways to combine voxel intensity and cluster extent. In this paper we examine permutation tests (Holmes *et al.*, 1996; Nichols & Holmes, 2002) based on combining functions (Pesarin, 2001; Lazar *et al.*, 2002) that incorporate both voxel intensity and cluster size information. The permutation framework does not require knowledge of the exact distribution of these combining functions. We examine three combining functions: the Tippett or minimum p combining function, which is analogous to Poline *et al.*'s minimum p-value approach, the cluster mass combining function of Bullmore *et al.*, and the Fisher combining function (Pesarin, 2001; Lazar *et al.*, 2002). Furthermore we attempt to calibrate combined inference according to signals of interest, either localized high intensity signals or extended low intensity signals, by using a weight in the Tippett and Fisher combining functions. As an extension of combined inference, we also propose a meta-combining function, a combining function of combining functions, in

order to benefit from the strengths of different combining functions at once, rather than performing multiple combined tests. In this meta-combining approach, rather than selecting a combining function which produces the most significant results, an investigator can obtain a single p-value summarizing outcomes from all the combining functions.

In the following Methods section, we explain the combining functions in detail, as well as a simulation-based validation and an application of these functions to second-level fMRI analysis data sets. In the Results section, findings from the simulation and the data analyses are presented. In the Discussion section, we examine the findings.

## 2 Methods

### 2.1 Statistic Image

We assume that voxel intensities of a brain image data set have the form

$$Y(v) = X\beta(v) + \sigma(v)\varepsilon(v) \quad (1)$$

where  $v = (x, y, z) \in \mathbb{R}^3$  is an index for voxels,  $Y(v) = \{Y_1(v), Y_2(v), \dots, Y_n(v)\}'$  is an  $n \times 1$  vector of observed voxel intensities at  $v$  from  $n$  scans,  $X$  is a known  $n \times p$  design matrix,  $\beta(v)$  is a  $p \times 1$  vector of unknown parameters,  $\sigma(v)$  is a scalar of unknown standard deviation at  $v$ , and  $\varepsilon(v) = \{\varepsilon_1(v), \varepsilon_2(v), \dots, \varepsilon_n(v)\}'$  is an  $n \times 1$  vector of unknown random errors with unit variance. Images are denoted by omitting the index  $v$ , so that, for example,  $\varepsilon_i$  denotes the error image from the  $i$ th scan. Since we use permutation framework in this study, we assume that error images  $\varepsilon$  are uncorrelated across subjects or scans, as in PET and multi-subject fMRI data.

Let  $\hat{\beta}(v)$  be an unbiased estimate of  $\beta(v)$ , then the residuals can be obtained as

$$e(v) = Y(v) - X\hat{\beta}(v)$$

and an estimate of the residual variance is

$$\hat{\sigma}^2(v) = \frac{1}{\eta} e(v)' e(v)$$

where  $\eta$  is the degrees of freedom for errors. If we assume  $\varepsilon_i(v)$ 's to be independent and identically normally distributed across subjects or scans, then the statistic image  $T$  can be calculated as

$$T(v) = \frac{\mathbf{c}\hat{\beta}(v)}{\sqrt{\mathbf{c}(X'X)^{-1}\mathbf{c}'\hat{\sigma}(v)}}$$

where  $\mathbf{c}$  is a  $1 \times p$  vector expressing the contrast of interest. Based on the  $T$  image, clusters are formed as a set of contiguous voxels whose  $T(v)$  exceeding a fixed cluster defining threshold  $u_c$ . In this cluster forming scheme, voxels sharing at least one common edge are considered as neighbors (the 18 connectivity scheme for a 3D image). For example, for a  $3 \times 3 \times 3$  voxel cube in a 3D space, all the voxels except 8 corner voxels are considered to be connected to the voxel at the center.

## 2.2 Inference

Once the statistic image  $T$  is calculated, the next step is to perform a statistical inference on  $T$  to identify any signals. We perform such inference using permutation test framework (Holmes *et al.*, 1996; Nichols & Holmes, 2002), with the Tippet, Fisher, and cluster mass combining functions. Combining functions are tools for implementing multivariate testing (Pesarin, 2001), in our case voxel and cluster size tests. In a combining function, information from these tests, referred as partial tests, are used as variables, and inference is made based on the value of the resulting combining function. While the cluster mass combining function can be calculated directly from the  $T$  image, the Tippet and Fisher combining functions require p-values from the partial tests.

### 2.2.1 Partial Tests: Voxel Intensity and Cluster Size Inferences

P-values for voxel intensity and cluster size can be obtained using separate permutation tests. Each permutation test is carried out in a similar manner based on the idea of exchangeability (Nichols & Holmes, 2002). Under the null hypothesis, data labels can be permuted, or randomly reassigned, without altering the distribution of the test statistic of interest. For each such permutation, a statistic image is created based on the permuted labeling, then the test statistic is recorded. This step is repeated for a sufficient number of times (typically 1,000 to 3,000) to generate an empirical distribution of the test statistic. P-values can be calculated by comparing peak intensities or cluster

sizes to this empirical distribution. In particular, test statistics from all the permutations are ordered from the largest to the smallest, then the p-value is determined by the location where the peak intensity or cluster size falls among these ordered test statistics. For example, if the size of a particular cluster falls between the 19th and the 20th largest among the 1,000 test statistics from all the permutations, then the p-value is obtained as  $p = \frac{20}{1000} = 0.02$ .

As a test statistic in a partial test, the highest peak intensity (for voxel test)  $T_{max}$  or the largest cluster size (for cluster size test)  $S_{max}$  is used. The use of the largest value for the intensity or the cluster size is to correct for multiple comparisons among clusters. The rationale behind using  $T_{max}$  (or  $S_{max}$ ) is that the probability of observing  $T_{max}$  (or  $S_{max}$ ) larger than  $t$  (or  $s$ ) is the same as the probability of at least one or more peak intensity (or cluster size) exceeding  $t$  (or  $s$ ). Thus use of the largest value as the test statistic controls the event of a family-wise error (FWE), or false positives occurring in all the voxels or clusters collectively. P-values with this correction are known as corrected, or p-values adjusted for multiple comparisons. In practice, corrected p-values of partial tests are obtained by comparing peak intensities and cluster sizes to the empirical distribution of  $T_{max}$  and  $S_{max}$ . More on FWE control can be found in Nichols & Hayasaka (2003), with additional permutation details in Nichols & Holmes (2002).

For each cluster, the resulting corrected p-values from the peak intensity and the cluster size are used in the Tippett and Fisher combining functions discussed below. There is a practical reason for this use of corrected p-values from the partial tests. It is more efficient to record thousands of  $T_{max}$  or  $S_{max}$  in order to calculate corrected p-values, rather than to record hundreds of thousands of peak intensities or cluster sizes to calculate uncorrected p-values. Use of uncorrected p-values is explored in Discussion section.

### 2.2.2 Combining Functions

Let  $P_i^t$  be the corrected p-value for the peak intensity of cluster  $i$ , and  $P_i^s$  be the corrected p-value for the cluster size of the same cluster. Then the Tippett combining function  $W_i^T$  and the Fisher

combining function  $W_i^F$  are defined as

$$W_i^T = 1 - \min(\log P_i^t, \log P_i^s) \quad (2)$$

$$W_i^F = -2(\log P_i^t + \log P_i^s). \quad (3)$$

The Tippett combining function  $W_i^T$  is analogous to Poline *et al.* (1997)'s minimum p-value approach. In both cases, the critical region can be defined by a single p-value for both partial tests<sup>1</sup>. If the smallest p-value of the two partial tests falls below this p threshold, the null hypothesis is rejected. In Poline *et al.*'s approach, the joint distribution of the intensity and cluster size is approximated by RFT, and the p threshold is found by this approximated joint distribution. For our combined test with  $W_i^T$ , once the critical value of  $W_i^T$  is found, then a p-value which produces that critical value is sought. Such p-value becomes the critical p-value threshold for both  $P_i^t$  and  $P_i^s$ .

The cluster mass combining function  $W_i^M$  is defined as

$$W_i^M = \sum_{v \in C_i} (T(v) - u_c) \quad (4)$$

where  $C_i$  is a set of voxel indices for cluster  $i$ . In other words,  $W_i^M$  is the mass of cluster  $i$  above the cluster defining threshold  $u_c$ . It is calculated by summing voxel intensity above  $u_c$  for all the voxels in that cluster.

Based on the values of the above combining functions, another permutation test is performed. From the distributions of the largest peak intensity and cluster size, corrected p-values from partial tests are assessed at each cluster in each permutation. Then a combining function  $W_i$  is calculated at each cluster in each permutation. The largest value of  $W_i$  is recorded for each permutation. Finally, together from all the permutations, the null distribution of the largest  $W_i$  is obtained. Based on this distribution, corrected p-values are calculated, and the null hypothesis is rejected at clusters if their corrected p-values are smaller than the desired level of significance. More detailed outline of the combined test is found in 2.2.5.

---

<sup>1</sup>Alternatively, it is possible to use max function instead of min function used in (2). In such case, the critical region is also defined by a single p-value, but both partial tests' p-values have to be below this critical p-value in order to be rejected. Unlike  $W_i^T$ , this test is not consistent.

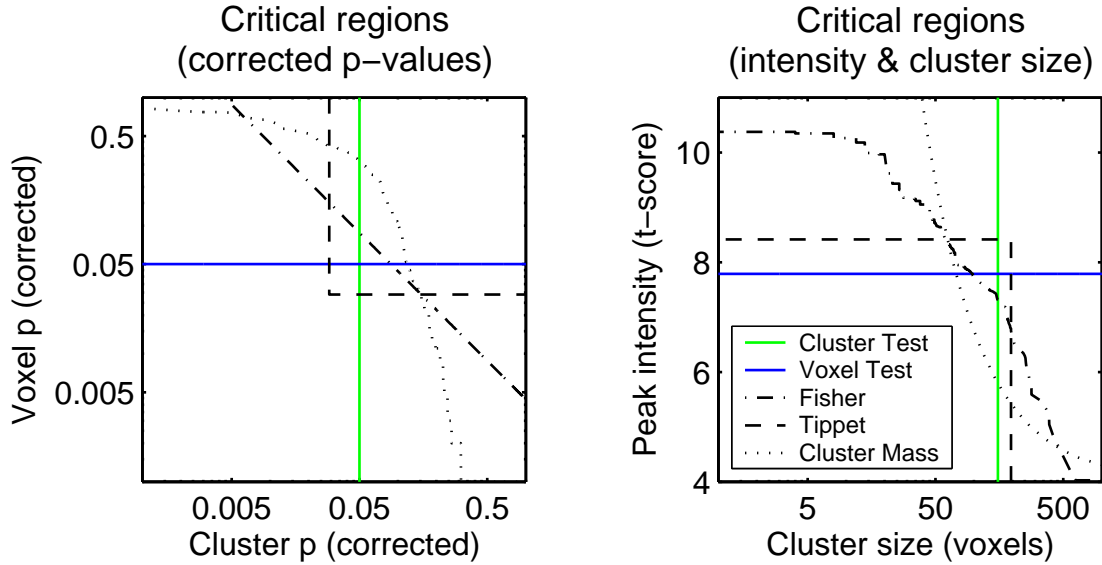


Figure 1: An example of critical regions for the three combining tests, as well as that of the voxel and cluster size tests, based on a multi-subject fMRI data on working memory (Marshuetz *et al.*, 2000). The left panel shows the critical regions in terms of partial p-values, whereas the right panel shows in terms of intensity and cluster size. For cluster mass, clusters are assumed to have a spherical shape with its intensity having a concave parabolic shape.

Each combining function is specialized for a certain type of signals. An example of critical regions for the above combining functions are shown in Figure 1, to demonstrate where the strengths of these combining functions lie. The critical regions for the three combining functions considered in this study represent three possible scenarios in combined inference. The Tippet combining function  $W_i^T$ , analogous to that of Poline *et al.*'s (1997) method, is sensitive to clusters when either peak intensity or cluster size is significant, but does not have extra sensitivity for clusters when both intensity and extent are marginally significant. On the other hand, the cluster mass function  $W_i^M$ , Bullmore *et al.*'s (1999) approach, has great sensitivity for a combination of marginally significant intensity and extent, but may not be able to detect signals when either intensity or extent is highly significant and the other is not. Therefore, strictly speaking, this test is not consistent (Pesarin, 2001), as mentioned in Introduction. In practice, however, the test could behave con-

sistently, meaning that the test is able to reject the null hypothesis even if only one of the partial tests produces an unusually small p-value. This is because all the clusters have mass even if their intensity or size is very small. For example, if the peak intensity of a cluster is arbitrarily large, say  $T(v) = 3000$ , then the cluster mass is large even if the cluster consists of a single voxel, resulting in rejection of the null hypothesis at that cluster. The Fisher combining function  $W_i^F$  is somewhere in between the Tippett and cluster mass combining functions. It can detect clusters if one of the partial tests is significant, and also has some extra sensitivity to marginally significant intensity and extent combination.

### 2.2.3 Calibration

In practice, an investigator might have a prior belief on the shape of signals; he or she might expect localized high intensity signals, or wide spread low intensity signals. When there is such a prior belief, it may be beneficial to calibrate the combining function to the signals of interest. For example, when localized signals are expected, the combining function should emphasize information from the voxel test, since the voxel test is more sensitive to localized signals. On the other hand, if signals are believed to be wide-spread, then the emphasis should be on the cluster size test in the combining function, since it is sensitive to spatially extended signals. The strength of such calibrated combining test is that, though it emphasizes either intensity or extent, it still utilizes information from the other.

Such calibration can be easily implemented by modifying the Tippett and Fisher combining functions since the contributions from the partial tests are standardized in terms of p-values between 0 and 1. Using a weight  $\theta \in (0, 1)$ , (2) and (3) can be modified as

$$W_i^T(\theta) = 1 - \min(2\theta \log P_i^t, 2(1 - \theta) \log P_i^s) \quad (5)$$

$$W_i^F(\theta) = -2(2\theta \log P_i^t + 2(1 - \theta) \log P_i^s). \quad (6)$$

In  $W_i^T(\theta)$  and  $W_i^F(\theta)$ ,  $\theta = 0.5$  corresponds to the unweighted statistics, as in (2) and (3). For  $\theta > 0.5$ , the  $P_i^t$  term dominates and a combining function becomes more sensitive to high intensity

peaks, whereas for  $\theta < 0.5$ , the  $P_i^s$  term dominates and the function becomes more sensitive to large clusters. For  $\theta = 0$ , the test becomes a cluster size test, and for  $\theta = 1$ , the test becomes a peak intensity test.

#### 2.2.4 Meta-Combining

Once p-values are calculated from the combining functions (2), (3), and (4), denoted by  $P_i^T$ ,  $P_i^F$ , and  $P_i^M$  respectively, then the meta-combining function is defined as

$$W_i^A = 1 - \min(\log P_i^T, \log P_i^F, \log P_i^M) \quad (7)$$

Based on the value of  $W_i^A$ , another permutation test is performed. In this meta-combining step, the largest value of  $W_i^A$  for each permutation is used as the test statistic, controlling the FWE rate. Note that (7) has the form of the Tippet combining function. It is also possible to use the Fisher combining function as a meta-combining function.

#### 2.2.5 Implementation

An actual combined test is carried out in four steps: partial tests, calculation of the combining function, estimation of combining function distribution, and calculation of p-values for the combined test. First, a permutation voxel intensity test and a permutation cluster size test are carried out, using the same permutations for both tests. During this permutation test, peak intensity and cluster size information from all the clusters in all the permutations are recorded. The same permutations are used in order to reduce computation time. At this step, corrected p-values for the peak intensity and cluster size,  $P_i^t$  and  $P_i^s$  respectively, are recorded for all the clusters in all the permutations. In the cluster mass combining function, cluster mass is calculated for all the clusters from all the permutations at this step. Once corrected partial p-values are obtained, then for each cluster in each permutation, the combining function  $W_i$  is calculated and recorded. After  $W_i$  is obtained from all the clusters, then for each permutation, the largest value of  $W_i$  is sought. Together from all the permutations, an empirical distribution of the largest  $W_i$  is obtained. Notice that in order to calculate the empirical distribution of  $W_i$ , the same permutations as the partial tests are

used so that no further permutations are necessary. Finally, based on this empirical distribution, FWE-corrected p-values can be found for the clusters from the original labeling.

More steps are necessary to perform a meta-combined test, which are very similar to that of a combined test. First, p-values from all three combined tests need to be calculated for all the clusters in all the permutations. This can be done by comparing each  $W_i^T$ ,  $W_i^F$ , or  $W_i^M$  to the empirical distribution of its largest values obtained above. By doing so, corrected p-values  $P_i^T$ ,  $P_i^F$ , and  $P_i^M$  are calculated. From these p-values, the meta-combining function  $W_i^A$  is calculated for all the clusters in all the permutations as in (7). The empirical distribution of the meta-combining test statistic  $W_i^A$  is obtained by recording the largest  $W_i^A$  from each permutation. Based on this empirical distribution of  $W_i^A$ , FWE-corrected p-values are found for all the clusters from the original labeling.

## 2.3 Simulation

To validate and to examine the performance of our combined and meta-combined tests, a simulation was carried out. For each realization in the simulation, a set of 15  $76 \times 76 \times 60$ -voxel Gaussian random noise images was generated by convolving a Gaussian white noise image with a 3D Gaussian kernel of FWHM (full-width at half-maximum) 4.5 voxels. The outer 16 voxels were truncated in order to avoid non-stationarity at the edge, resulting in the image size of  $48 \times 48 \times 32$  voxels. A sphere-shaped signal having a uniform intensity was then added to the center of the simulated noise images. Figure 2 shows the signals used in the simulation. To the generated images, partial permutation tests (peak intensity and cluster size), combined tests (Tippet, Fisher, and cluster mass), and the meta-combined test, all with 500 permutations, were then applied. The cluster defining threshold  $u_c$  was set to the uncorrected  $p = 0.01$  threshold of  $t_{14}$  ( $u_c = 2.6245$ ), and the significance level of the tests was set to 0.05. This was repeated for 2,000 realizations, and rejection rates were recorded for all the tests in all the settings. The Monte Carlo standard deviation of rejection rates was 0.0049.

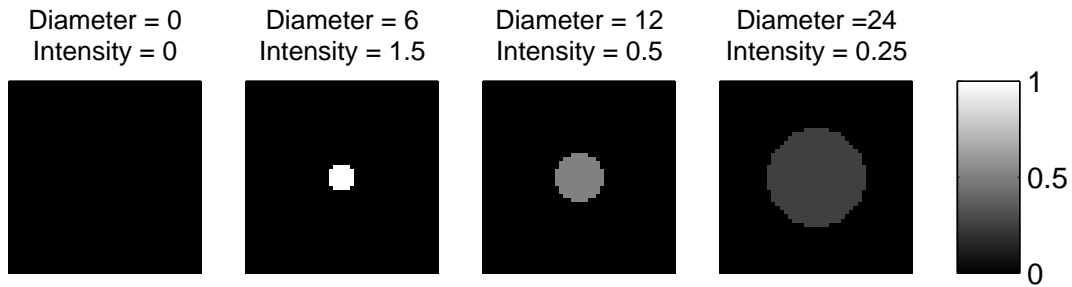


Figure 2: The diameter and the intensity of signals used in the simulation.

## 2.4 Applications

Combined voxel-cluster size tests using the above combining functions, as well as the meta-combining test were applied to two multi-subject fMRI data sets, the emotional response data and the working memory data. The performance of the combined tests is compared to that of the voxel test and the cluster size test. Furthermore, calibration of the Tippett and Fisher combining functions is examined for different values of calibration weight  $\theta$ .

### 2.4.1 Emotional Response Data

Taylor *et al.* (2003) studied emotional response among schizophrenia patients. The fMRI images were acquired from eight controls (Ct) and six schizophrenia patients (Pt) participated in this study, while they were presented with aversive (AV) and non-aversive (NA) IAPS (International Affective Picture System) images, and five blank (BL) gray images with a centered fixation cross. For each subject, a contrast image of AV-NA was calculated. Then the resulting contrast images of size 44,552 voxels in a  $53 \times 63 \times 45$  space with  $3 \times 3 \times 3$ mm voxels were compared in a two-sample  $t$ -test (Ct-Pt) to assess the difference in emotional responses between controls and schizophrenics. The permutation test with 1,000 permutations was employed in the analysis, and for each permutation, the statistic image was thresholded at  $u_c = 3.0$  to define clusters.

### 2.4.2 Working Memory Data

This data set is from Marshuetz *et al.* (2000) used as an example in Nichols & Holmes (2002). Order effects in working memory were examined using fMRI.

Each of 12 subjects participated in eight fMRI acquisitions. Images were acquired from 12 subjects under three different conditions presented as blocks in one of two randomized orders. For each of three conditions used, 528 images were acquired at  $TR = 2s$ . Two of the three conditions, item recognition and control were considered in this example. For the item recognition condition, each subject was shown five letters, then a probe letter after a two-second interval. The subject was asked to respond “yes” or “no” if the probe letter was among the five letters presented. For the control condition, the subject was shown five X’s, then two seconds later either “y” or “n”. They were asked to respond “yes” for “y” and “no” for “n”.

The data set was analyzed using a random effect model (Holmes & Friston, 1999). For each subject, a contrast image of difference between item recognition and control was calculated. The resulting contrast images of size 122,659 voxels in a  $79 \times 95 \times 68$  space with  $2 \times 2 \times 2$ mm voxels were analyzed in a one-sample  $t$ -test to assess effects associated with the item recognition task among the subjects. The permutation test with 1,000 permutations was used, and for each permutation, the statistic image was thresholded at the 0.001 critical value of a  $t_{11}$  random variable ( $u_c = 4.02$ ) to define clusters. Furthermore, critical regions for the weighted combining functions  $W_i^T(\theta)$  and  $W_i^F(\theta)$  were examined at  $\theta = 0.4, 0.5, \text{ and } 0.6$ .

### 2.4.3 Computing Environment

For both analyses, a DELL PC with dual 2.4GHz Xeon processors and 2GB of RAM with MATLAB 6.5 (MathWorks Inc., Natick, MA) running on a Linux platform was used. The calculation time was 14 minutes for the emotional response data and 21 minutes for the working memory data, with one of the two processors.

### 3 Results

#### 3.1 Simulation

The rejection rates of the partial, combined, and meta-combined tests from the simulation are shown in Table 1. For the high intensity signal (diameter/intensity = 6/1.5), the cluster mass combining function  $W_i^M$  is found to be more powerful than the voxel test. Furthermore, because of  $W_i^M$ , the meta-combined test  $W_i^A$  is found to be as powerful as  $W_i^M$ . For signals with large extent (12/0.5 and 24/0.25), none of the combined tests is quite as powerful as the cluster size test, but all combined tests are much more powerful than the voxel test. In any settings, the meta-combined test is just slightly less powerful than the best combined test. Thus, when the form of signals is unknown, the meta-combined test is an ideal choice.

Tests	Signal (diameter/intensity)			
	(0/0)	(6/1.5)	(12/0.5)	(24/0.25)
Partial				
Cluster	0.039	0.069	0.543	0.585
Voxel	0.048	0.881	0.148	0.117
Combined				
Tippet $W_i^T$	0.036	0.800	0.478	0.502
Fisher $W_i^F$	0.045	0.825	0.479	0.477
Mass $W_i^M$	0.040	0.992	0.495	0.489
Meta-combined				
$W_i^A$	0.041	0.986	0.480	0.486

Table 1: Results from the simulation. Rejection rates of the partial, combined, and meta-combined tests for different signals.

#### 3.2 Emotional Response Data

The results from the emotional response data analysis are shown in Figure 3 and Table 2. The corresponding empirical null distributions of the test statistics are shown in Figure 4. One significant cluster is found by  $W_i^F$  and  $W_i^M$  in the medial pre-frontal cortex (MPFC), the area associated with processing of emotions (Phan *et al.*, 2002). This indicates that the controls have greater magnitude BOLD response while viewing aversive images, than the schizophrenics. The combined tests with

$W_i^F$  and  $W_i^M$  are able to detect this cluster, while both partial tests are only marginally significant for this cluster (peak  $p=0.060$  and cluster size  $p=0.050$ ). On the other hand, the combined test with  $W_i^T$  produces a larger p-value than each of the partial tests. This may be because the Tippett combining function corrects for multiple testing for peak intensity and cluster size. The meta-combining function  $W_i^A$  produces a p-value slightly larger than that of  $W_i^F$  and  $W_i^M$ , but is still able to detect this cluster.

### 3.3 Working Memory Data

The results from the working memory data analysis are shown in Figure 5 and Table 3. The corresponding empirical null distributions of the test statistics are shown in Figure 6. All the combining functions are able to detect five clusters, which are very similar to the results in Nichols & Holmes (2002) on the same data set. Activations are found in the bilateral posterior parietal (clusters 2 and 4), left thalamus (cluster 1), and anterior cingulate (cluster 3) regions, which are typical of working memory studies (Marshuetz *et al.*, 2000), as well as in the left pre-motor region (cluster 5).

Though clusters 4 and 5 are significant by their size but not by peak intensity, all the combined tests are able to detect these clusters. For such clusters where only one of the partial tests is significant,  $W_i^T$  produces the smallest p-values of the three, as expected by the strength of the Tippett combining function. However, compared to p-values from individual partial tests, p-values from  $W_i^T$  can be larger than the minimum p-value of the two partial tests, as seen in clusters 2 and 3 as well as the emotional response results above. The meta-combining function  $W_i^A$  is able to cover this weakness very well. Though it produces a slightly large p-value than  $W_i^T$  in cluster 5,  $W_i^A$  produces p-values as comparably small as  $W_i^F$  and  $W_i^M$  in clusters 2 and 3.

Figure 7 shows critical regions for the weighted combining functions  $W_i^T(\theta)$  and  $W_i^F(\theta)$  with various weights ( $\theta = 0.45, 0.5, \text{ and } 0.55$ ), along with the corresponding critical regions of the partial tests. For  $\theta = 0.45$ , critical regions from both combining functions include more areas from the cluster size test's critical region than that of the voxel test. On the other hand, for  $\theta = 0.55$ , the

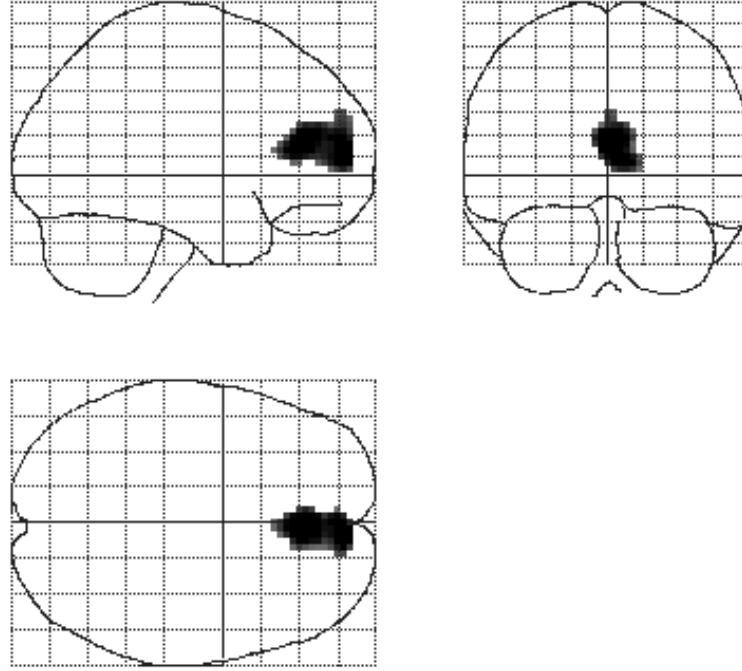


Figure 3: Results from the emotional response data analysis. A cluster was found in the medial pre-frontal cortex MPFC by the combined tests with  $W_i^F$  and  $W_i^M$

Cluster $i$	Size (voxels)	p-values						t-score	Location ( $x, y, z$ mm)
		Cluster	$W_i^T$	$W_i^F$	$W_i^M$	$W_i^A$	Peak		
1	342	0.050	0.082	0.036	0.028	0.039	0.060	7.11	(3, 36, 15)

Table 2: P-values from the various tests in the emotional response data analysis

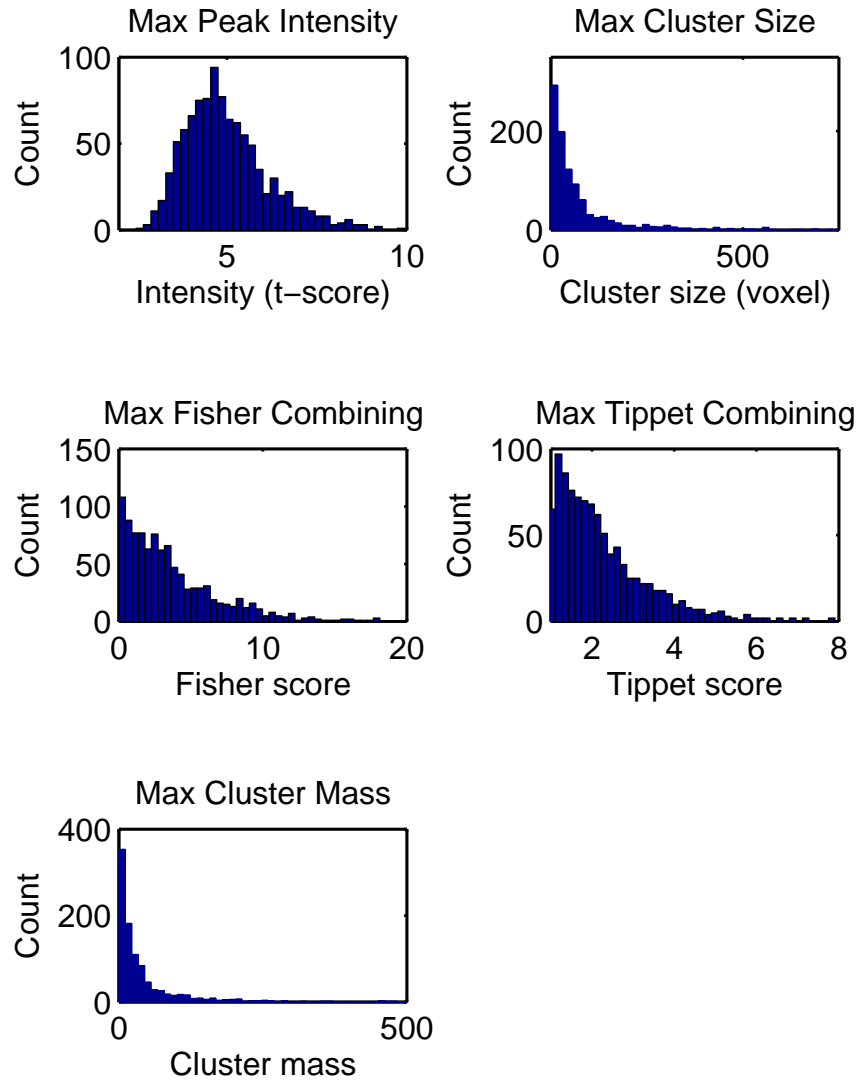


Figure 4: The empirical null distribution of the test statistics from the emotional response data analysis.

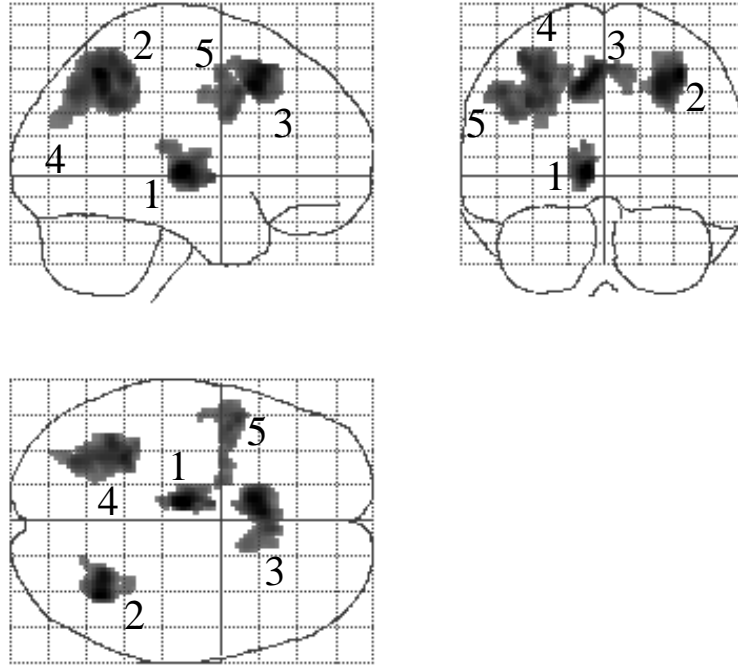


Figure 5: Results from the working memory data analysis. Activations are found in the bilateral posterior parietal (2,4), left thalamus (1), and anterior cingulate (3) regions, which are typical of working memory studies (Marshuetz *et al.*, 2000), as well as in the left pre-motor region (5).

Cluster $i$	Size (voxels)	p-values						t-score	Location ( $x, y, z$ mm)
		Cluster	$W_i^T$	$W_i^F$	$W_i^M$	$W_i^A$	Peak		
1	345	0.010	0.001	0.001	0.003	0.001	0.001	13.15	(-8, -18, 2)
2	529	0.005	0.009	0.002	0.001	0.001	0.007	10.19	(36, -58, 48)
3	520	0.005	0.009	0.002	0.002	0.003	0.012	9.37	(-10, 16, 44)
4	1138	0.001	0.001	0.004	0.001	0.001	0.083	7.36	(-30, -46, 48)
5	436	0.006	0.011	0.021	0.012	0.016	0.208	6.31	(-48, 8, 40)

Table 3: P-values from the various tests in the working memory data analysis

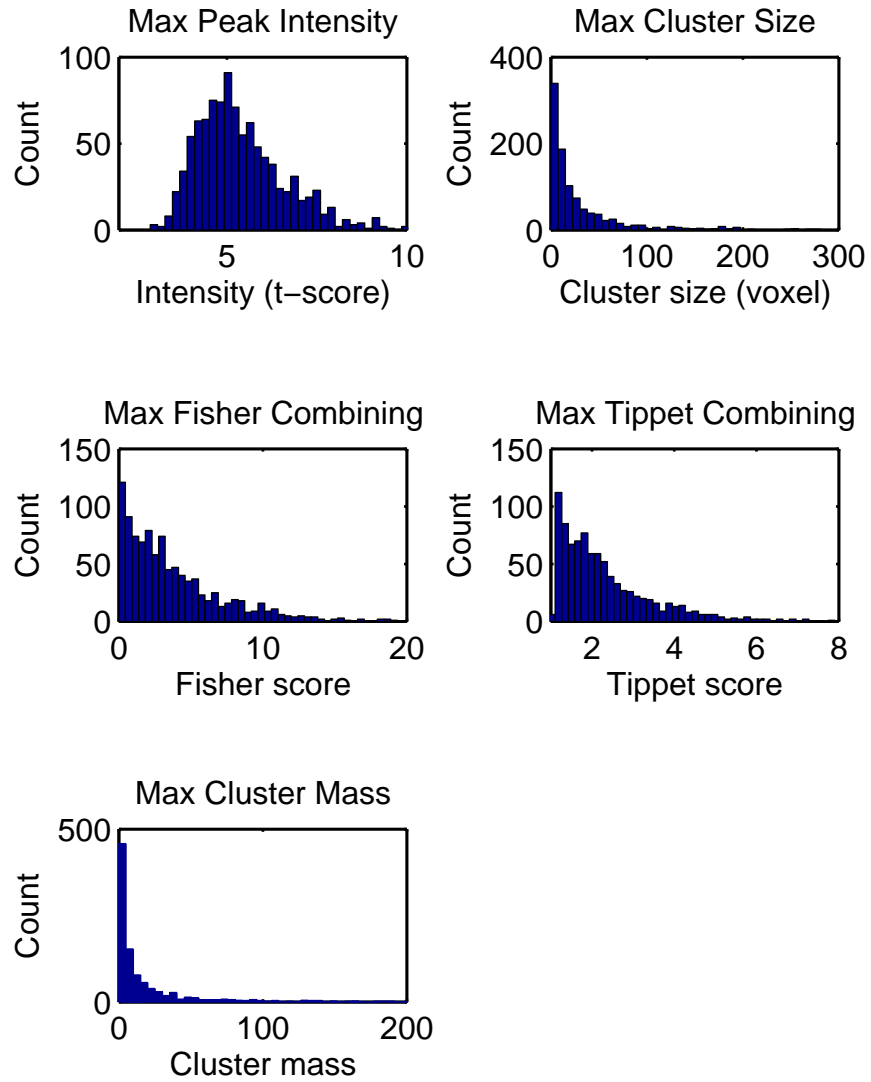


Figure 6: The empirical null distribution of the test statistics from the working memory data analysis.

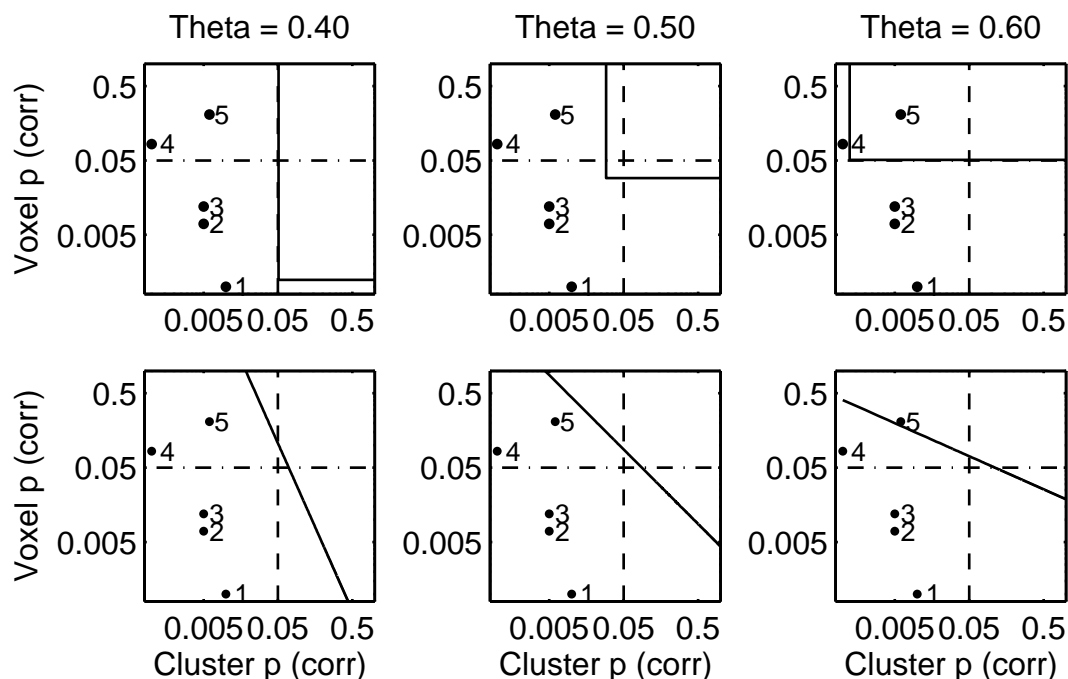


Figure 7: Critical regions at 0.05 significance level for the tests with the weighted Tippett (top, solid lines) and weighted Fisher (bottom, solid lines) combining functions with  $\theta = 0.4, 0.5,$  and  $0.6$  (from left to right) from the working memory data example. Also noted in the plots are the critical regions of the voxel intensity test (dash-dot lines) and the cluster size test (dashed lines). Clusters are indicated according to their partial test p-values.

combining functions' critical regions include more areas from the voxel test's critical region. This indicates that both weighted combining functions are able to calibrate the critical regions with the value of  $\theta$ . A value of  $\theta < 0.5$  emphasizes p-values from the cluster size test, making a combined test more sensitive to spatially extended signals. Conversely, if  $\theta > 0.5$ , the test becomes more sensitive to high intensity signals. For example, though clusters 5 is in the critical regions of both combined tests at  $\theta = 0.45$  and  $0.5$  due to its size, it is not in the critical region at  $\theta = 0.55$  because its peak intensity p-value is not small enough.

## 4 Discussion

We have developed combined voxel-cluster size tests using three combining functions. Our simulation demonstrated that our combined tests and meta-combined test perform well for any types of signals examined. In particular, the meta-combined test is consistently powerful, thus ideal for

situations where the shape of expected signals is unknown. From the results of the multi-subject fMRI analyses, the Tippett combining function  $W_i^T$  is found to be less sensitive than the partial tests. A p-value from this test is usually greater than the smaller one of the two partial tests. However, this test is sensitive to clusters whose only one of the partial tests is significant (localized high intensity signals or spatially extended low intensity signals). The other combining functions, Fisher  $W_i^F$  and cluster mass  $W_i^M$ , can be more significant than individual partial tests when both partial tests are significant. Even when both partial tests are marginally significant, these tests can detect clusters, because their critical regions cover some outside portion of the partial tests' critical regions (see Figure 1). Between  $W_i^F$  and  $W_i^M$ ,  $W_i^F$  should be still sensitive to clusters of which only one of the partial test is highly significant while the other is non-significant.

The meta-combining function  $W_i^A$  is able to combine the strength of the above combining functions into one. There are infinitely many combining functions for peak intensity and cluster size information, but our meta-combining function can cover three typical scenarios represented in the three combining functions. The Tippett test is sensitive to a rejection in either one of the partial tests. The cluster mass test is sensitive to moderately high intensity peaks and large clusters occurring simultaneously. The Fisher test is a compromise of the two. Hence by combining these three, the meta-combined test is sensitive to signals significant in one of the partial tests, AND signals marginally significant in both partial tests.

When used with the permutation framework, our combining function strategy provides an easy way to implement voxel-cluster combined inference. One of the strengths of combined approach is its ability to make an inference “without regard to the underlying dependence relations” (Pesarin (2001), p134). Calculating a combining function from partial p-values at each cluster implicitly incorporates the dependence structure among these p-values.

The main reason for our use of corrected p-values from the partial tests, rather than uncorrected p-values, is to reduce computational burden. It is easy to find the uncorrected cluster size distribution using permutations since at most a few hundreds of clusters could occur for each permutation, but finding uncorrected voxel intensity distribution requires a large memory to record all the voxel

intensities above the cluster defining threshold  $u_c$  for all the permutations.

In some cases, uncorrected p-values from the partial tests are more desirable than corrected p-values. When uncorrected p-values from the partial tests are used, then the distribution of the Fisher combining function  $W_i^F$  can be approximated by a  $\chi^2$  random variable with  $df = 4$ , and this test can be implemented parametrically. In such case, uncorrected cluster size p-values can be found from the permutation test and uncorrected voxel p-values can be found based on  $t$  distribution with appropriate  $df$ . Since the cluster size distribution based on RFT may be biased (Hayasaka & Nichols, 2003), it is preferable to use of permutations to find an empirical cluster size distribution, which is known to be almost exact (Holmes *et al.*, 1996). For the intensity distribution, it is reasonable to assume the distribution of each voxel intensity of  $T$  as a  $t$  random variable due to the central limit theorem.

In any case, when p-values are to be used in the combining functions, partial test p-values have to be both corrected or both uncorrected. Otherwise, if p-values from one test were corrected and p-values from the other test were uncorrected, then the uncorrected p-value could dominate in the combining function, since they are not corrected for multiple comparisons and could be considerably smaller than the corrected counterpart.

We suspect that there is very little effect on the sensitivity and the specificity of the test with our use of partial p-values, compared to using the actual peak intensity and cluster size information directly in a combined function. With a sufficient number of permutations, there is an almost one-to-one correspondence between the peak intensity (or cluster size) and its p-values based on permutations. Thus our use of marginal p-values simply maps the peak intensity (or the cluster size) to an interval  $(0, 1)$ .

Our attempt to calibrate the combined test becomes possible with use of weight  $\theta$  and permutation framework. In the working memory data example, we are able to tune down low intensity clusters (see Figure 7). Except the weight, since the contributions from both partial tests in the weighted combining functions are the same. Hence, in theory, these tests also should be able to filter out high intensity localized signals with small  $\theta < 0.5$ . Furthermore, it is possible to modify

the cluster mass combining function  $W_i^M$  so that it can also be calibrated using a weight  $\theta$ . Details are found in Appendix A.

The optimal value of  $\theta$  could depend on various factors, such as image smoothness, signal intensity and widths. Since traditional random field theory only models the noise and not the form of the signal, we have left noise & signal modeling and estimation of  $\theta$  for future work.

If there is a prior belief about the data, then  $\theta$  can be adjusted accordingly *a priori*. For example, if strong localized activations are expected, then  $\theta$  can be slightly augmented from 0.5. One could analyze a dataset multiple times with different  $\theta$ 's, but this adds an additional dimension to the search volume (i.e., spatial dimensions  $x, y, z$  and the parameter  $\theta$ ). Such addition of an extra dimension to the search space could lead to reduced sensitivity, as seen in scale space search (Worsley *et al.*, 1996). In any case, one should not choose value of  $\theta$  based on p-values of partial tests *a posteriori* after examining partial p-values of each cluster.

In summary, we developed & implimented combined voxel-cluster size tests using combining functions and permutation framework. We extended this combined inference into meta-combining, by combining strengths of different combining functions. We also developed weighted combining functions which adjust the combined test according to signals of interest. As seen in our simulation and data analyses, these methods provide a way of detecting a wider variety of signals than existing intensity-based or extent-based inference methods.

## 5 Acknowledgments

This Human Brain Project / Neuroinformatics research is funded by the National Institute of Mental Health, the National Institute on Aging, and the National Institute of Biomedical Imaging and Bioengineering. The authors would like to express their appreciation to Dr. Stephan Taylor and Dr. Chrsity Marshuetz for providing us the data sets used in this paper.

## References

- Adler, R J. 1980. *The Geometry of Random Fields*. New York: Wiley.
- Bullmore, E T, Suckling, J, Overmeyer, S, Rabe-Hesketh, S, Taylor, E, & Brammer, M J. 1999. Global, Voxel, and Cluster Tests, by Theory and Permutation, for a Difference between Two Groups of Structural MR Images of the Brain. *IEEE Transactions on Medical Imaging*, **18**, 32–42.
- Friston, K J, Worsley, K J, Frackowiak, R S J, Mazziotta, J C, & Evans, A C. 1994. Assessing the Significance of Focal Activations Using Their Spatial Extent. *Human Brain Mapping*, **1**, 210–220.
- Friston, K J, Holmes, A, Poline, J-B, Price, C J, & Frith, C D. 1996. Detecting Activations in PET and fMRI: Levels of Inference and Power. *NeuroImage*, **4**, 223–235.
- Hayasaka, S, & Nichols, T E. 2003. Validating Cluster Size Inference: Random Field and Permutation Methods. *NeuroImage*, **20**, 2343–2356.
- Holmes, A P, & Friston, K J. 1999. Generalizability, random effects, and population inference. *Proceedings of the 4th International Conference on Functional Mapping of the Human Brain, June 7-12, 1998, Montréal, Canada*. *NeuroImage*, **7**, S754.
- Holmes, A P, Blair, R C, Watson, J D G, & Ford, I. 1996. Nonparametric Analysis of Statistic Images from Functional Mapping Experiments. *Journal of Cerebral Blood Flow and Metabolism*, **16**(1), 7–22.
- Lazar, N A, Luna, B, Sweeney, J A, & Eddy, W F. 2002. Combining Brains: A Survey of Methods for Statistical Pooling of Information. *NeuroImage*, **16**, 538–550.
- Marshuetz, C, Smith, E E, Jonides, J, DeGutis, J, & Chenevert, T L. 2000. Order information in working memory: fMRI evidence for parietal and prefrontal mechanism. *Journal of Cognitive Neuroscience*, **12**(S2), 130–144.

- Mudholkar, G S, & George, E O. 1979. The logit method for combining probabilities. *Pages 345–366 of: Rustagi, J (ed), Symposium on Optimizing Methods in Statistics.* New York: Academic Press.
- Nichols, T. E., & Hayasaka, S. 2003. Controlling the Familywise Error Rate in Functional Neuroimaging: A Comparative Review. *Statistical Methods in Medical Research*, **12**(5), 419–446.
- Nichols, T E, & Holmes, A P. 2002. Nonparametric Permutation Tests for Functional Neuroimaging: A Primer with Examples. *Human Brain Mapping*, **15**, 1–25.
- Pesarin, F. 2001. *Multivariate Permutation Tests.* New York: Wiley.
- Petersson, K M, Nichols, T E, Poline, J-B, & Holmes, A P. 1999. Statistical Limitations in Functional Neuroimaging II. Signal Detection and Statistical Inference. *Philosophical Transaction of the Royal Society of London. Series B*, **354**, 1261–1281.
- Phan, K L, Wager, T, Taylor, S F, & Liberzon, I. 2002. Functional Neuroanatomy of Emotion: A Meta-Analysis of Emotion Activation Studies in PET and fMRI. *NeuroImage*, **16**, 331–348.
- Poline, J-B, Worsley, K J, Evans, A C, & Friston, K J. 1997. Combining Spatial Extent and Peak Intensity to Test for Activations in Functional Imaging. *NeuroImage*, **5**, 83–96.
- Stouffer, S A, Suchman, E A, DeVinney, L C, Star, S A, & Williams, R M. 1949. *The American Soldier: Vol. I. Adjustment During Army Life.* Princeton, NJ: Princeton University Press.
- Taylor, S F, Welsh, R C, Phan, K L, & Liberzon, I. 2003. Medial prefrontal cortex dysfunction in schizophrenia. *Page 146 (57) of: Annual Meeting of the American College of Neuropsychopharmacology Abstracts. San Juan, Puerto Rico, December 2003.*
- Worsley, K J, Evans, A C, Marrett, S, & Neelin, P. 1992. Three-Dimensional Statistical Analysis for CBF Activation Studies in Human Brain. *Journal of Cerebral Blood Flow and Metabolism*, **12**, 900–918.

Worsley, K J, Marrett, S, Neelin, P, & Evans, A C. 1996. Searching Scale Space for Activation in PET Images. *Human Brain Mapping*, **4**, 74–90.

# Appendices

## A Weighted Cluster Mass

For Cluster  $i$  consisting of set of voxels  $C_i$ , the cluster mass  $W_i^M$  is defined as

$$W_i^M = \sum_{v \in C_i} (T(v) - u_c) \quad (8)$$

Since  $T(v) > u_c$  for any  $v$  in Cluster  $i$ , (8) can be rewritten as

$$W_i^M = \sum_{v \in C_i} |T(v) - u_c| \quad (9)$$

which is a sum of  $L^1$  distance between  $u_c$  and  $T(v)$  at each voxel. From (9), let us consider squared mass using  $L^2$  distance, such that

$$W_i^{M'} = \sum_{v \in C_i} |T(v) - u_c|^2$$

Since the distance between  $u_c$  and  $T(v)$  is squared, this squared mass favors peaked clusters. Also from (9), let us consider square-root mass using  $L^{\frac{1}{2}}$  distance

$$W_i^{M''} = \sum_{v \in C_i} |T(v) - u_c|^{\frac{1}{2}}$$

which favors a large cluster size. Thus, if the cluster mass is calculated with  $L^m$  distance,  $m \in (0, \infty)$ , the sensitivity of the cluster mass test can be adjusted according to signals of interest by

$$W_i^M(m) = \sum_{v \in C_i} |T(v) - u_c|^m$$

If peaked clusters are sought, then  $m \gg 1$  should be used, and for clusters with large extent,  $m \ll 1$  should be used. To be consistent with the two other weighted combining functions  $W_i^T(\theta)$  and  $W_i^F(\theta)$ , let  $m = \theta/(1 - \theta)$  with  $\theta \in (0, 1)$ , so that the weighted cluster mass function can be written as

$$W_i^M(\theta) = \sum_{v \in C_i} |T(v) - u_c|^{\frac{\theta}{1-\theta}}$$

For  $\theta = 0.5$ ,  $W_i^M(\theta)$  becomes the cluster mass combining function. As  $\theta \rightarrow 0$ , the cluster area dominates and the test becomes a cluster size test. On the other hand, as  $\theta \rightarrow 1$ , the largest voxel intensity in the cluster dominates and the test becomes a peak intensity test.

Interestingly cluster mass  $W_i^M$  is related to the Fisher combining function  $W_i^F$ . Let  $\bar{T}_i$  be the mean of  $T(v)$  in cluster  $i$ . Then (8) can be rewritten as

$$W_i^M = (\bar{T}_i - u_c)S_i \quad (10)$$

where  $S_i$  is the size of cluster  $i$ . Then after taking the logarithm, (10) becomes

$$\log(W_i^M) = \log(\bar{T}_i - u_c) + \log(S_i) \quad (11)$$

This is similar to the Fisher combining function

$$W_i^F = -2(\log P_i^t + \log P_i^s)$$

since there is a close relationship between  $(\bar{T}_i - u_c)$  and  $P_i^t$ , and also between  $S_i$  and  $P_i^s$ , though this is not a one-to-one transformation. Notice, however, while both partial p-values are in a range of  $(0, 1)$ ,  $(\bar{T}_i - u_c)$  and  $S_i$  are in ranges of  $(0, \infty)$  and  $\{1, 2, 3, \dots, \infty\}$ , respectively. Since the magnitude of  $S_i$  is considerably larger than that of  $(\bar{T}_i - u_c)$  in practice, the contribution from the cluster size is likely to dominate in (11).

## B Alternative Functional Forms for Partial P-values

Instead of  $\log P_i$ , different functional forms of partial p-values can be used in calculation of the Tippett and Fisher combining functions. One possibility is to use  $\Phi(\cdot)$ , the cumulative distribution function of a standard normal random variable. The function  $\Phi^{-1}(1 - P_i)$  can be used to transform a p-value into a number in  $(-\infty, \infty)$  (Stouffer *et al.* (1949) cited in Lazar *et al.* (2002)). Another possibility is to use the log odds  $-\log\left(\frac{P_i}{1-P_i}\right)$ , which transforms a partial p-value into a number in  $(-\infty, \infty)$  (Mudholkar & George (1979) cited in Lazar *et al.* (2002)).