Parametric Cluster Size Tests: A Comparison between SPM Package and `fmri_stat` Package

Satoru Hayasaka

September 9, 2002

1 Introduction

The cluster size inference is a statistical test of the spatial extent of signals in brain image analyses. which is typically done in two steps. First, for a contrast of interest, the t-image is thresholded at a certain pre-defined threshold, and contiguous voxels above the threshold are grouped into clusters. Second, the spatial extent (i.e. volume) of each of these clusters is compared to the critical cluster size to test the null hypothesis.

There are different approaches to implement the cluster size inference, but perhaps the most widely known and used is the parametric cluster size test (Friston *et al.*, 1994; Worsley *et al.*, 1996) in SPM package. Despite its popularity, one down side of the SPM cluster size test is that it requires stationarity, or homogeneous spatial autocorrelation structure. Recently, this stationarity assumption has been overcame by Worsley (2002).

In this note, I will go over the two different approaches in the parametric cluster size inference, namely, SPM and `fmri_stat`, their differences and similarities. In particular, I will go over the cluster size measures, distribution of cluster sizes, and distribution of the number of clusters.

2 Elements of Cluster Size Tests

In both approaches, the basic machinery in the test is the same. First, cluster sizes are measured. Then the distribution of a single cluster size is calculated. Finally, the critical cluster size adjusted for the family-wise error rate is calculated based by exploiting the distribution of the number of clusters.

2.1 Cluster Size Measure

The first step in the cluster size test is to measure cluster sizes. Under the stationarity assumption, this can be done simply by measuring the volume of clusters. More formally,

$$S = \int_C 1 \tag{1}$$

where C is a cluster. The volume can be in the unit of voxels or resels, a roughness adjusted volume metric. In any case, under the stationarity assumption, the volume is always proportional to the number of voxels, because all the voxels are considered to have the same smoothness and volume.

Under the non-stationarity assumption, however, measuring cluster sizes involve a more complicated process. First, the local roughness needs to be estimated. Worsley (2002) suggests calculating the resel density at each voxel. Let us call this resel density **RPV** (Resel Per Voxel), which can be calculated at a voxel $\mathbf{t} \in \mathbb{R}^D$ by

$$\widehat{RPV}(\mathbf{t}) = (4 \ln 2)^{D/2} |\dot{\mathbf{u}}(\mathbf{t})' \dot{\mathbf{u}}(\mathbf{t})|^{1/2} \tag{2}$$

where $\mathbf{u}$ is the normalized residual image calculated from the residual image $\mathbf{r}$ by

$$\mathbf{u} = \mathbf{r}/(\mathbf{r}'\mathbf{r})^{1/2}$$

Typically the dimension D is 3. Once the RPV is calculated at each voxel, then the size of a cluster C is calculated simply by

$$\widehat{R} = \int_C \widehat{RPV} \tag{3}$$

One thing to note here is that the size of each voxel is different, since each voxel are considered to have different smoothness.

2.2 Cluster Size Distribution

2.2.1 SPM Approach

The SPM approach in calculating the cluster size distribution is based on the following identity. Let N be the number of resels above the threshold u, L be the number of clusters in the search volume, and S be the size of a cluster, then it is assumed that

$$\mathbf{E}[S] = \mathbf{E}[N]/\mathbf{E}[L] \tag{4}$$

(Friston *et al.*, 1994). Here, if the contrast image is assumed to be a t-field with ν degrees of freedom, then $\mathbf{E}[N]$ is

$$\mathbf{E}[N] = Q(1 - T_\nu(u)) \tag{5}$$

where Q is the resel count for the search volume and T_ν is the cumulative density function (cdf) of a t-random variable with ν degrees of freedom. $\mathbf{E}[L]$ is approximated by Worsley *et al.* (1996)

$$\mathbf{E}[L] \simeq \sum_{d=0}^{D} Q_d \rho_d(u) \tag{6}$$

where Q_d's are d-dimensional resel counts and ρ_d's are d-dimensional resel densities. Details regarding these quantities are explained elsewhere (Worsley *et al.*, 1996). Once $\mathbf{E}[N]$ and $\mathbf{E}[L]$ are obtained, then using the identity (4) and the fact that, for large u, $S^{2/D}$ is exponentially distributed (Nosko, 1969; Friston *et al.*, 1994), the distribution of S is approximated by

$$Pr(S > s) \simeq exp(-\psi s^{2/D}) \tag{7}$$

where ψ is

$$\psi = \left[\frac{\Gamma(\frac{D}{2}+1)\mathbf{E}[L]}{\mathbf{E}[N]} \right] \tag{8}$$

In real data, the lower dimension terms in (6) are often larger since edges are jagged, though ideally they should be smooth. Thus, for more robustness, in the SPM package, (8) is calculated using only the highest dimension term in $\mathbf{E}[L]$, by

$$\psi = \left[\frac{\Gamma(\frac{D}{2}+1)\ Q_D\ \rho_D(u)}{\mathbf{E}[N]} \right]$$

2.2.2 `fmri_stat` Approach

The distribution of a cluster size can be approximated without using the identity (4). In `fmri_stat`, the cluster size distribution in Cao (1999) is directly used, which is known to be

$$R \sim cB^{1/2} \left(\frac{U_0^D}{\prod_{a=1}^{D} U_a} \right)^{1/2} \tag{9}$$

where B is a Beta random variable with parameters $(1, (\nu - D)/2)$, U_0 is a χ^2 random variable with degrees of freedom ν, and U_a's ($a = 1, 2, \ldots, D$) are independent χ^2 random variables with degrees of freedom $\nu + 2 - a$. c is a constant chosen so that

$$\mathbf{E}[R] \; Q_D \; \rho_D(u) = Q_D \; (1 - T_\nu(u))$$

In other words, the resels above u (right hand side) matches the sum of expected cluster sizes (left hand side).

In real data, the true cluster size R (in resels) is never known since the true RPV is never known. It is thus estimated as in (3). However, like any other estimates, $\widehat{R}$ has some estimation variability. If this variability is assumed to be negligible, then the distribution of $\widehat{R}$ can be considered as the same as that of R in (9). Or, the estimation variability can be incorporated by

$$\widehat{R} \sim R \left(\frac{\prod_{b=1}^{D} V_b}{V_0^D} \right) \tag{10}$$

where V_0 is a χ^2 random variable with degrees of freedom ν, V_b's ($b = 1, 2, \ldots, D$) are independent χ^2 random variables with degrees of freedom $\nu - b$, and R is the true cluster size which is distributed as (9). In this case, the constant c is a value such that

$$\mathbf{E}[\widehat{R}] \; Q_D \; \rho_D(u) = Q_D \; (1 - T_\nu(u))$$

In `fmri_stat` package, the distribution function of $\widehat{R}$ is calculated by taking the logarithm, then by transforming into Fourier domain. Taking a log transforms a product of random variables into a sum or random variables, and a Fourier transformation changes a convolution of random variables into a sum of Fourier transforms.

2.3 Family-Wise Error Rate Correction

Even if the distribution of each cluster size is known, we still need to know the joint distribution of all the clusters in order to adjust the family-wise error rate at a desired level. In both methods, clusters are assumed to be independent, which is a reasonable assumption (Adler, 1980; Friston *et al.*, 1994). In that case, the number of clusters larger than s has a Poisson distribution with the mean $\mathbf{E}[L](1-F(s))$, where $F(\cdot)$ is the cdf of a cluster size, either from SPM or `fmri_stat`.

The critical cluster size adjusted for family-wise error can be found by simply finding the distribution of the largest cluster size. The rational for the family-wise error correction and its relationship to the largest cluster size can be found elsewhere. The critical cluster size at α-level family-wise error rate is

$$Pr(S_{max} > s) \simeq 1 - exp\{-\mathbf{E}[L](1-F(s))\} \tag{11}$$

3 Conclusion

In this note, I reviewed the two approaches for parametric cluster size inferences. Whether or not these methods perform well is determined by simulation studies which will follow.

References

Adler, R J. 1980. *The Geometry of Random Fields*. New York: Wiley.

Cao, J. 1999. The Size of the Connected Components of Excursion Sets of χ^2, t, and F Fields. *Advances in Applied Probability*, **31**, 579–595.

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.

Nosko, V P. 1969. Local structure of Gaussian random fields in the vicinity of high level shines. *Soviet Mathematics: Doklady*, **10**, 1481–1484.

Worsley, K J. 2002. Non-stationary FWHM and its effect on statistical inference of fMRI data. *NeuroImage Human Brain Mapping 2002 Meeting*, **16**, 779–780.

Worsley, K J, Marrett, S, Neelin, P, Vandal, A C, Friston, K J, & Evans, A C. 1996. A Unified Statistical Approach for Determining Significant Signals in Images of Cerebral Activation. *Human Brain Mapping*, **4**, 58–73.