A Cluster Overlap Measure for Comparison of Activations in fMRI Studies

Guillermo A. Cecchi, Rahul Garg, and A. Ravishankar Rao

IBM T. J. Watson Research Center, NY, USA

Abstract. Most fMRI studies use voxel-wise statistics to carry out intra-subject as well as inter-subject analysis. We show that statistics derived from voxel-wise comparisons are likely to be noisy and error prone, especially for inter-subject comparisons. In this paper we propose a novel metric called *weighted cluster coverage* to compare two activation maps. This metric is based on the intersection of spatially contiguous clusters of activations. It is found to be more robust than voxel-wise comparisons and could potentially lead to more statistical power in fMRI-based group studies.

1 Introduction

Theoretical considerations as well as ample experimental evidence suggest that brain function is, to some extent, supported by the correlated activity of *groups* of neurons. These groups, in turn, tend to be spatially contiguous and consistent with the spatial continuity of anatomical patterns of connectivity and neuronal identity. However, the field of fMRI image analysis is dominated by techniques that use a voxel-wise linear model, the General Linear Model (GLM) approach. As evidence, Grinband *et al.* [1] identified 170 papers published in leading journals just in the first six months of 2007 that used this approach.

An open problem in the fMRI analysis literature is how to appropriately conduct inter-subject studies and summarize their results. The typical approach is to apply the GLM method, which results in candidate regions responsive to the particular experimental stimulus or protocol. This is followed by voxel-wise comparison amongst different subjects. A major drawback of this approach is that voxel-based methods do not give robust inter-subject results. By robustness we mean the sensitivity of the results with respect to small changes in the positions of the voxels or their contents. Several reasons contribute to the lack of robustness across subjects, and even across sessions for the same subjects: misalignment, movement, field distortions and noise, and morphological differences in individual brains. Usually, these problems are addressed by the application of significant spatial smoothing, which effectively leads to a loss of resolution.

A number of recent publications have proposed different algorithms to identify functional clusters. Thirion *et el.* [2,3] have explored the issue of comparing inter-subject fMRI activation maps. They developed a *functional* parcellation technique based on spectral clustering, to group similar regions across multiple

subjects. Most other methods are heavily based on a priori knowledge of the anatomy and connectivity, in particular sulcal identification [4–8]. This peak of interest highlights the need for a consistent measure to compare cluster-based maps.

At the same time, recently developed techniques for functional network analysis have shown promise. They are based on extracting graph structures from spatio-temporal datasets [9], and then computing relevant statistical properties of these graphs. For instance, hub density maps can be computed, which capture the degree of linkage of voxel nodes in the graph. This measure of activity tends to be more highly localized, or compact, than that of the more diffuse activity captured via GLM maps.

The above observations lead naturally to the concept of *resolution*, to describe how well localized the measured brain activity is. We consider a higher resolution technique to be one that potentially generates a larger number of spatially compact candidate regions. Hence there is a need for an improved method to perform group studies and compare maps across subjects. With this aim in mind, we have created the following set of desired capabilities that we seek, and which would be present in an ideal method.

1. Robustness with respect to misregistration. An ideal method is one whose results do not degrade with small spatial shifts in the data, especially when they arise from differences in the brain anatomies of subjects.
2. Robustness with respect to different runs for the same subject. This issue addresses temporal shifts, including temporal partitioning of the data sets, i.e. the analysis of different time segments should yield consistent results if the experimental paradigm is time-invariant.
3. Robustness across multiple scales of resolution. An ideal method would be able to perform comparisons between both high and low resolution data.
4. Generic applicability. An ideal method would be applicable across multiple maps, including maps obtained via GLM or other methods.

The main contribution of this paper is to propose a method called the *weighted cluster coverage* method, which is able to satisfy the four requirements that we seek in an ideal method, as identified above. The cluster coverage method overcomes many of the limitations of applying simple voxel-based image difference or correlation metrics that are typically used. The method is described in detail in Section 2.3.

2 Comparison of statistical maps

In this section we present three metrics, namely *voxel correlations*, *weighted set overlap* and *weighted cluster coverage* to compare two fMRI maps. The voxel correlation and set overlap metrics arise naturally under the assumptions made in typical voxel-wise group analysis methods. The weighted cluster coverage metric designed by us, uses spatial contiguity information to overcome the limitations of voxel-wise methods.

2.1 Voxel correlations

In a typical fMRI-based study, a set of maps is computed for each subject using the GLM analysis. Each map in the set represents "brain activations" in a specific experimental condition, suitably transformed into a normalized statistic such as "Z", "t" or an "F" score. These maps are then aligned to a standardized brain atlas (such as the MNI or Talairch atlases). A voxel-wise comparison of the aligned maps of each subject is carried out using various statistical techniques (e.g., ANOVA) to arrive at the final conclusions of the study [10].

An implicit assumption in such an approach is that value represented by each voxel in the aligned map remains the same (modulo the noise) in a given experimental condition for all subjects in a given group. The voxel-correlation metric is designed to evaluate this assumption.

Given a map μ of N voxels, define $Top(\mu, p)$ as the set of Np voxels with the highest values in the map μ. Here, p denotes a percentile. The *voxel correlation* between two maps μ_1 and μ_2 at a percentile p is defined as the Pearson correlation coefficient between μ_1 and μ_2, restricted to the set $S = Top(\mu_1, p) \cap Top(\mu_2, p)$. Intuitively, voxel correlation represents the similarity between the voxel values of the active voxels of the two maps. Note that if $\mu_1 = \mu_2$, then the voxel correlation is 1; if $\mu_1 = -\mu_2$ then voxel correlation is -1 and if μ_1 and μ_2 are independent then the voxel correlation is zero for all values of p[1].

2.2 Weighted set overlap

The machine-learning based fMRI data analysis techniques (such as MVPA [11] or [12]) rely upon extracting a set of meaningful "features" that are used to build models of the data. The model and the corresponding features are used to interpret the results of the experiment and derive scientific conclusions.

An implicit assumption in all such models is that each voxel represents the same physical (or physiological) process in different experimental conditions and subjects. The *weighted set overlap* metric is designed to evaluate this assumption.

Define the weight of a map μ with respect to the set S as the sum of the map values of voxels in the set S. Formally $W(\mu, S) = \sum_{i \in S} \mu_i$. The weighted set overlap of two maps μ_1 and μ_2 at percentile p is given by

$$\frac{W(\mu_1 + \mu_2, S_1 \cap S_2)}{W(\mu_1, S_1) + W(\mu_2, S_2)}$$

where $S_1 = Top(\mu_1, p)$ and $S_2 = Top(\mu_2, p)$. Intuitively, this measure represents the degree of overlap of voxels found active in the two maps. Note that if $\mu_1 = \mu_2$ then the weighted set overlap is 1 for all the values of p. It is 0, if the top p fraction of the voxels of μ_1 and μ_2 do not have any common voxel.

[1] The choice $S = Top(\mu_1, p) \cup Top(\mu_2, p)$ which gives more voxels, was not used since it introduces an artificial negative bias in the correlations.

2.3 The cluster coverage metric

The cluster coverage metric is designed to overcome the limitations of voxel-wise methods by taking spatial contiguity of voxels into account.

Consider two sets of voxels in space, S_1 and S_2. These could arise from two fMRI trials on two subjects, or from two trials on the same subject, for instance. The voxels in the two sets are first clustered into groups of spatially connected 3-d components (two voxels are said to be connected if they share a common face, i.e. we use 6-connectedness in three dimensions). Let these components be denoted by the sets $\{r_{1j}\}$ and $\{r_{2j}\}$. The *weighted cluster coverage* of S_1 by S_2 is defined as the ratio of the weight of clusters of S_1 that intersect with clusters of S_2, to the total weight of clusters of S_1. Formally,

$$C_w(S_1, \mu_1, S_2, \mu_2) = \frac{\sum_{j:r_{1j} \cap S_2 \neq \phi} W(\mu_1, r_{1j})}{\sum_j W(\mu_1, r_{1j})} \tag{1}$$

The weighted cluster coverage $C_w(S_1, \mu_1, S_2, \mu_2)$ is an asymmetric measure, which ranges from 0 to 1. If $C_w(S_1, \mu_1, S_2, \mu_2)$ is zero, this implies that S_1 and S_2 have no voxels in common. As $C_w(S_1, \mu_1, S_2, \mu_2) \rightarrow 1$, this indicates that many connected components of S_1 intersect with connected components in S_2. A cluster coverage of one implies that every spatially contiguous component in S_1 intersects with a component in S_2.

The mean weighted cluster coverage of two maps μ_1 and μ_2 at a percentile p is defined as $(C_w(S_1, \mu_1, S_2, \mu_2) + C_w(S_2, \mu_2, S_1, \mu_1))/2$ where $S_1 = Top(\mu_1, p)$ and $S_2 = Top(\mu_2, p)$. This is a symmetric variant of the cluster coverage metric.

3 Evaluation and Results

3.1 Evaluation methodology

We used fMRI data from a simple finger tapping experiment. The data consists of fMRI scan of three sessions of six healthy subjects. The sessions, which lasted for 800 seconds, consisted of blocks of a self-paced finger-tapping task. Each session was split into two sub-sessions corresponding to the first and the second half of the session. For each of the sub-sessions, GLM analysis [13] was carried out to find the areas in the brain active during the finger-tapping task. The resulting maps of Z-statistics were registered to the MNI atlas, and then used for comparisons.

Three types of comparisons were carried out. The *intra-session* comparisons refer to the comparisons between the first half and the second half of the sessions (leading to 6×3 comparisons). The *inter-session* comparisons refer to comparisons between two different sessions of the same subject (leading to $6 \times {}^3C_2 \times 2$ comparisons. Here ${}^nC_r$ refers to the number of ways in which r objects can be chosen from n distinct objects). The *inter-subject* comparisons refer to the comparisons between the same sub-sessions of different subjects (leading to ${}^6C_2 \times 3 \times 2$ comparisons). For each of the comparison, seven different values of

p ranging from 0.2 to 0.002 were considered. All the three comparison metrics (discussed earlier) were computed for all the p-values considered. For each value of p, the mean and standard deviation of the similarity metrics were computed.

3.2 Results

Figure 1(A) shows the voxel correlations of the Z-maps of the finger tapping experiment at different values of p. The mean value of intra-session, inter-session and inter-subject overlap (using the voxel correlation metric) are plotted as a function of p. The error bars represent the standard deviation of the metric.

The mean voxel correlations for inter and intra-session comparison remain in the range 0.87 to 0.70 and 0.85 to 0.55 respectively. The intra-session overlap is slightly higher and within 27 percent of the inter-session overlap for all values of p. In contrast, the mean inter-subject overlap starts at 0.36 and becomes negative for values of p less than or equal to 0.02. This shows that the actual values of the Z-statistic, although fairly consistent within a subject, are highly inconsistent across multiple subjects.

Figure 1(B) shows the weighted set overlap of the Z-maps for different values of p. The results are similar to those of the voxel-correlation metric. For the intra and inter-session comparisons the mean value of weighted set overlap is in the range 0.82 to 0.76 and 0.78 to 0.68 respectively, with inter-session overlap slightly higher than the intra-session overlap (but within 12% of each other). However, the weighted set overlap in the inter-subject comparisons starts at 0.55 for $p = 0.2$ and becomes as small as 0.067 for $p = 0.002$, indicating its inconsistency across the subjects.

The mean weighted coverage measure performs much better as shown in Figure 1(C). For inter- and intra-session comparisons, the mean weighted coverage is almost the same (within 5%) with a low standard deviation. For inter-subject comparison, it is in the range 0.97 to 0.71. This indicates that the weighted cluster coverage measure, which takes into account the spatial contiguity of active voxels, is much more robust for intra- and inter-subject comparisons as compared to the voxel-wise methods discussed above.

Overlap with misregistration: In order to evaluate the impact of misregistration, the second sub-session of each session was spatially shifted by different amounts. For each shift, all the three overlap measures were computed for $p = 0.05$. Figure 1(D) shows the overlap as a function of the shift amount. The voxel correlations and weighted set overlap are very sensitive to misregistration. A shift of less than 5mm (which is not uncommon across subjects) reduces the voxel correlation from 0.63 to 0.23 and weighted set cover from 0.70 to 0.40. The weighted cluster coverage metric is more tolerant to shifts up to 25mm (primarily because the areas functionally active in the finger tapping task have a similar spatial extent). If the amount of shift is more than 25mm, then all the metrics fall to zero as there is not overlap in the functionally active areas of the first sub-session with the misregistered second sub-session.

Random Maps: The weighted cluster coverage metric has a positive bias since it computes the overlap of spatially contiguous clusters. If a large fraction of voxels are chosen, many clusters of the one map may intersect with the second map by chance, leading to an unreasonably high weighted cluster coverage value. To evaluate the extent of this bias, we compared randomly generated maps using the three metrics.

To make a random map, first each voxel was assigned a normally distributed random value. To make the map similar to fMRI activation maps, a spatially smoothing 2.5mm Gaussian filter was applied to the randomly generated voxel values. The resulting map was masked using the brain mask of one of the subject in the study. Now pairs of random maps were compared with each other.

Figure 1(E) shows the overlap of random maps. The voxel correlations remain close to zero for all the values of p, as expected. The bias in weighted set overlap follows the p-value as expected. The weighted cluster coverage shows biases of $0.95, 0.73, 0.39$ for values of p in $0.2, 0.1, 0.05$ respectively. However, for $p = 0.02, 0.01, 0.005, 0.002$, this bias rapidly diminishes to $0.1, 0.05, 0.02, 0.01$ respectively. These results are not unexpected since the spatially contiguous clusters become large when large fraction of voxels is chosen. This increases the chance of intersection of clusters, making the weighted cluster coverage metric more liberal at large values of p. However, when the fraction of voxels chosen is small (say ≤ 0.02), this bias quickly reduces to acceptably small values.

In general, random maps induce a probability distribution on the value of the weighted cover metric. This distribution may be used to convert the cover metric to a p-value representing the chance likelihood of the event.

Maps generated using different techniques: Brain maps can be generated using several other techniques in addition to GLM. For example the network analysis technique [9] generates *hub maps* representing the density of functional network connections. The functionally active areas are also found to have high connectivity and therefore high link-density in the hub maps. In addition, several other areas of high link-density which are not found active in the GLM analysis are also present in the hub maps. Figure 1(G) shows one such slice of a subject in the finger tapping experiment, where the threshold was held at the same percentile for both maps. The Supplementary Motor Area, which is found active by GLM analysis (colored blue), also has high link-density (red). However, one can see other areas that have high link-density (red) but not found active by the GLM analysis; in particular, there is a prominent cluster centered in the Posterior Cingulate that is not locked to the experimental protocol.

Figure 1(F) shows the overlap of GLM maps and the hub maps using different metrics. The overlap using voxel correlations and the weight set cover is symmetric and in the range 0.5 to 0.3. However, the cluster coverage metric shows much larger overlap starting from close to 1 for high values of p to close to 0.6 for low values of p. The asymmetric nature of the cluster coverage measure also uncovers the property of these maps discussed above, also visible in Figure 1(G). This confirms that most of the clusters found active by GLM analysis also contain voxels with high connectivity, and that the converse is not true.

4 Conclusions and Future Work

We have shown that the cluster coverage metric gives better consistency, lower standard deviations and robustness with respect to activation thresholds than metrics that do not exploit the a priori information about the spatial contiguity of brain function. The metric clearly reduces intra-subject variability, but its effect is more dramatic for inter-subject comparisons, making it particularly promising for large group studies. The effectiveness of the metric is also reflected in the slow degradation upon misregistration, when compared to non-cluster metrics, a feature of particular importance for both inter- and intra-subjects studies. Our metric meets the four requirements for an ideal comparison method, as described in the introduction.

This metric could be extended to the surface-based methods [7] by computing spatially contiguous clusters (and their intersections) on the manifold defined by the cortical surface.

References

1. J. Grinband et al. Detection of time-varying signals in event-related fMRI designs. *Neuroimage*, 43(3):509–520, 2008.
2. B. Thirion et al. Dealing with the shortcomings of spatial normalization: Multi-subject parcellation of fMRI datasets. *Human Brain Mapping*, 27(8):678–693, 2006.
3. T. Vincenf et al. Sensitivity analysis of parcellation in the joint detection-estimation of brain activity in fMRI. In *ISBI*, pages 568–571, 2008.
4. J.W. Meyer et al. MRI-based topographic parcellation of human brain cerebral white matter: 1. Technical foundations. *Neuroimage*, 9:1–17, 1999.
5. J.-F. Mangin et al. From 3D magnetic resonance images to structural representations of the cortex topography using topology preserving deformations. *Journal of Mathematical Imaging and Vision*, 5(4):297–318, 1995.
6. P. Thompson et al. High-resolution random mesh algorithms for creating a probabilistic 3D surface atlas of the human brain. *Neuroimage*, 3(1):19–34, 1996.
7. G. Lohmann et al. Automatic labelling of the human cortical surface using sulcal basins. *Medical Image Analysis*, 4(3):179–188, 2000.
8. A. Cachia et al. A generic framework for parcellation of the cortical surface into gyri using geodesic voronoi diagrams. *Medical Image Analysis*, 7(4):403–416, 2003.
9. V.M. Eguiluz et al. Scale-free functional brain networks. *Physical Review Letters*, 94(018102), 2005.
10. K. Friston et al. *Statistical Parametric Mapping, The Analysis of Functional Brain Images*. Academic Press, 2007.
11. K. A. Norman et al. Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, 10(9):424–430, September 2006.
12. T. M. Mitchell et al. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195, May 2008.
13. *FSL Release 3.3*, http://www.fmrib.ox.ac.uk/fsl (2006).

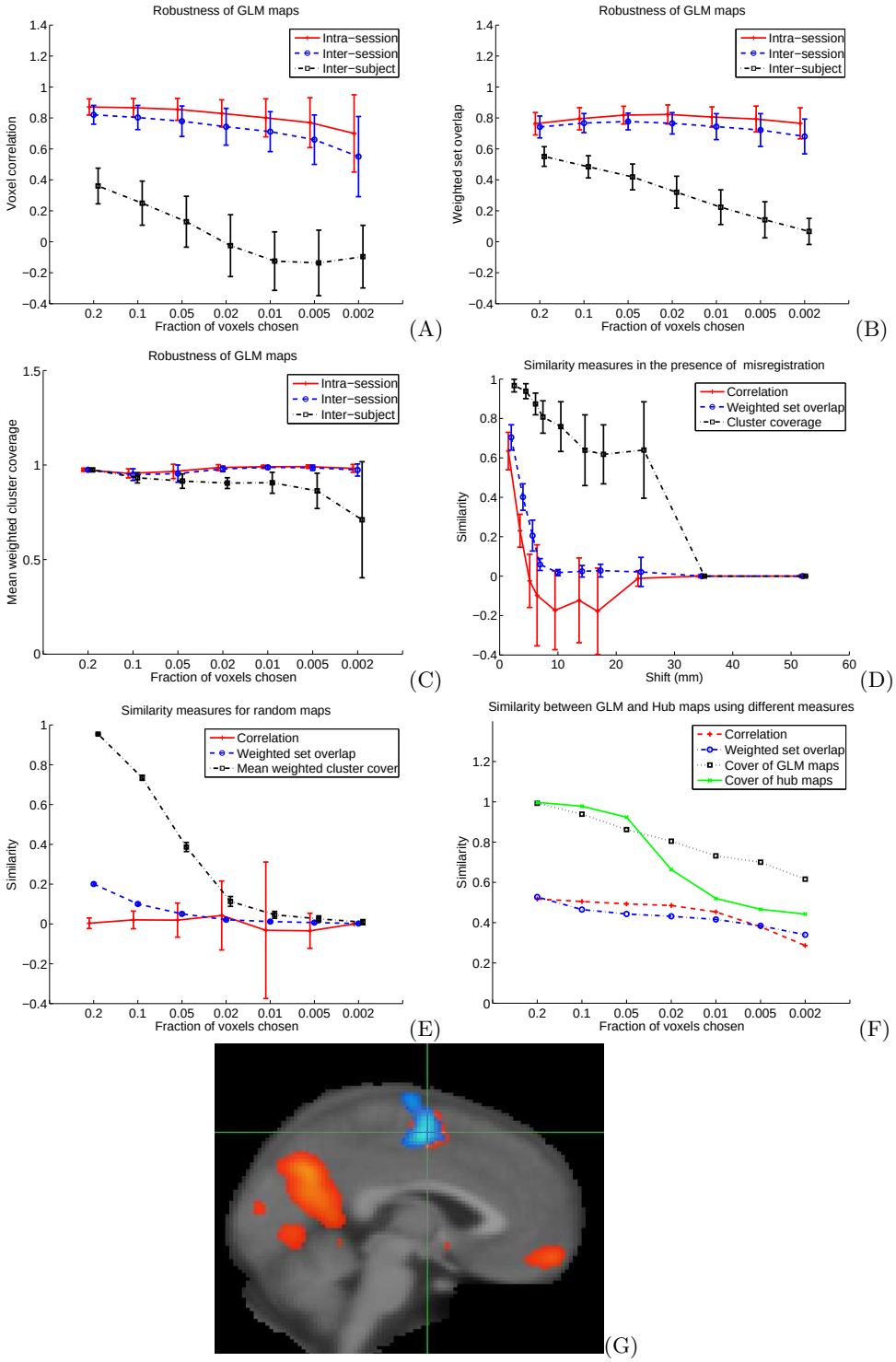
Robustness of GLM maps
Intra−session
Inter−session
Inter−subject
Voxel correlation
Fraction of voxels chosen
(A)
Robustness of GLM maps
Intra−session
Inter−session
Inter−subject
Weighted set overlap
Fraction of voxels chosen
(B)
Robustness of GLM maps
Intra−session
Inter−session
Inter−subject
Mean weighted cluster coverage
Fraction of voxels chosen
(C)
Similarity measures in the presence of misregistration
Correlation
Weighted set overlap
Cluster coverage
Similarity
Shift (mm)
(D)
Similarity measures for random maps
Correlation
Weighted set overlap
Mean weighted cluster cover
Similarity
Fraction of voxels chosen
(E)
Similarity between GLM and Hub maps using different measures
Correlation
Weighted set overlap
Cover of GLM maps
Cover of hub maps
Similarity
Fraction of voxels chosen
(F)
(G)

Fig. 1.