Cluster based Analysis of FMRI Data

Ruth Heller
Department of Statistics and Operations Research
Tel Aviv University, Tel Aviv 69978, Israel
rheller@post.tau.ac.il
Damian Stanley
Center for Neural Science
New York University, 4 Washington Place, New York, New York 10003
das@cns.nyu.edu
Daniel Yekutieli
Department of Statistics and Operations Research
Tel Aviv University, Tel Aviv 69978, Israel
yekutiel@post.tau.ac.il
Nava Rubin
Center for Neural Science
New York University, 4 Washington Place, New York, New York 10003
nava.rubin@nyu.edu
Yoav Benjamini*
Department of Statistics and Operations Research
Tel Aviv University, Tel Aviv 69978, Israel
ybenja@post.tau.ac.il

December 15, 2005

Abstract

We propose a method for the statistical analysis of fMRI data that tests cluster units rather than voxel units for activation. The advantages of this analysis over previous ones are both conceptual and statistical. Recognizing that the fundamental units of interest are the spatially contiguous clusters of voxels that are activated together, we set out to approximate these cluster units from the data by a clustering algorithm especially tailored for fMRI data. Testing the cluster units has a two-fold statistical advantage over testing each voxel separately: the signal to noise ratio within the unit tested is higher, and the number of hypotheses tests compared is smaller. We suggest controlling FDR on clusters, i.e. the proportion of clusters rejected erroneously out of all clusters rejected, and explain the meaning of controlling this error rate. We introduce the powerful adaptive procedure to control the FDR on clusters. We apply our cluster based analysis (CBA) to both an event-related and a block design fMRI vision experiment, and demonstrate its increased power over voxel-by-voxel analysis in these examples as well as in simulations.

Key words : fMRI, brain imaging, spatial clustering, FDR, adaptive FDR, multiple testing, power, inference on clusters.

1 Introduction

The typical analysis of fMRI data uses one or both of two main analysis approaches. The single-voxel approach creates activation maps by testing each voxel separately (possibly after spatial pre-processing, e.g. smoothing) for correlation with the experimental paradigm (predictor) and declaring a voxel active if the p-value is less than some threshold. (The threshold value may be pre-decided, or it may be adjusted adaptively by the data, e.g. using FDR.) The second common approach is to pre-define a region of interest (ROI), based on either anatomical or functional data (by an already-established paradigm known to activate that region), and then to perform statistical analysis on the ROI time-course obtained from the new experiment.

Both of these analysis approaches have produced a wealth of important findings. Nevertheless, they have several limitations. Activation maps obtained by single-voxel analysis are inherently limited by the SNR of individual voxel data, which is typically low. Furthermore, the very large number of statistical tests (a typical acquisition involves tens of thousands of voxels) requires adjusting the p-values for multiple comparisons, imposing high statistical thresholds that may reveal only the voxels with the very highest SNR but mask others that do have real effects. To avoid this loss in sensitivity, activation maps are often presented with 'raw' p-values, i.e. without adjusting for multiple comparisons, choosing the threshold on a case-by-case basis; but this hampers the replicability of the results because it makes it hard to compare results from different experiments and/or observers. The ROI approach overcomes the low SNR inherent in single-voxel data, but introduces other serious shortcomings. The most obvious problem is that it thwarts researchers' ability to discover effects of the experimental manipulations in brain regions other than those already hypothesized and pre-defined. In addition, the chosen ROI itself may be comprised of sub-regions that behave differently, but current ROI analysis methods do not allow researchers to discover such microstructure. Finally, there are methodological problems: the quality of the ROI data depends heavily on how reliably the region(s) could be defined prior to the critical experiment. The pre-defined ROI is likely to contain a mixture of voxels that do co-vary with the experimental manipulation with voxels that do not, and the latter add noise without adding any signal.

In this paper we present a novel fMRI analysis method, a 'cluster-based analysis' (CBA) method. The approach can be thought of as a 'hybrid' between the single-voxel and the ROI analyses, combining some of the advantages of each while avoiding many of their pitfalls. Like the single-voxel approach, CBA creates complete activation maps: every voxel in the acquisition volume has an a priori chance of being 'discovered'. The important difference from the single-voxel approach is that the units of analysis are now contiguous clusters of voxels, taking advantage of the increased SNR of multi-voxel data, as in the ROI approach.

Our approach is based on a central tenet articulated by Penny and Friston (2003): "the fundamental quantities of interest to the neuroimager are the location, shape, and temporal signature of clusters of voxels showing task-related activity." The clusters may be large, containing many voxels, or they may be small such that they are comprised of only a few voxels (e.g., the V1 "blind spot"). In both cases, the unit of a 'voxel' is arbitrarily determined by the measurement technique and does not represent a primary neural entity. Although this is implicit in the way results are reported in most studies, there is lack of adequate analysis methods to deal with functionally-significant clusters. The correlation among neighboring voxels is well recognized, and is commonly incorporated into the analysis by applying a spatial filter prior to the tests for significance. The spatial resolution of the resulting statistical maps depend in such cases on the spatial filters used, losing the opportunity to capture a more refined microstructure of correlations that may exist in the data.

The approach we propose here makes use of the correlation between neighboring voxels while retaining the spatial resolution of the data. We first identify clusters based on correlation between voxel time series during a preparatory scan (e.g., a functional ROI localizer). We then perform on each of the clusters a test for significant activation during the target experiment. The null hypothesis we test is that all voxels within the cluster are non-active. We define a cluster as "active" if it contains at least one voxel that is active, and as "non-active" if it contains no active voxels. We will use the terms "detected cluster" or "a cluster that is declared active" to refer to a cluster whose null hypothesis is rejected. When testing whether to declare a cluster as active, we use the time-course signal constructed from the average of the constituent voxels' time-courses. Other than that, testing proceeds as in voxel-by-voxel analysis (e.g. a generalized linear model (GLM) analysis of the correlation between the signal and the experimental paradigm). This approach guarantees that each p-value is uniformly distributed under the null hypothesis, thus validating our testing procedure.

Our approach has several advantages: (1) averaging data from multiple voxels increases the SNR of each statistical comparison; (2) since the statistical testing is now performed on clusters, the total number of tests is reduced; (3) controlling the proportion of erroneously-detected clusters is more relevant than merely the proportion of erroneously-detected voxels. Indeed, a common practice is to eliminate from the activation maps isolated voxels (mini-clusters) even if they passed the threshold. Similarly, smoothing the activation map introduces signal into non-active voxels but creates no new regions (i.e. aggregates of contiguous voxels). Both widely used procedures reveal the preference of investigators for inference on regions rather than on individual voxels.

Note that the above procedure is based on two experimental stages, so the clusters are defined on a different data set than the one used to test for activation under the paradigm of interest. Furthermore, since the first experimental stage is essentially the same as that used in the traditional ROI approach, ROIs can still be pre-defined and, as we shall see later, used in conjunction with the clusters in adaptive, more powerful statistical testing. When a localizer experiment is not available or not possible to conduct, it is still possible to use the CBA approach by performing the experiment twice, using the data from the first experimental run to generate clusters and the results of the second experiment are to test the clusters for paradigm-related activation.

In section 2.1 we describe the first stage of the analysis, how to define the units of analysis using a clustering method. In section 2.2 we describe the second stage of the analysis, how to discover which clusters are active. We may seek clusters of activity in the entire brain, or within a predefined ROI. If the search is constraint to a relatively small ROI, we suggest a further improvement in the statistical analysis in section 2.3, that describes a more powerful method for controlling for multiple testing. This method may be successfully applied also to test voxels of activity rather than clusters of activity within the ROI. Next, we apply our analysis to both real and simulated fMRI data. The results are detailed in section 4 and our conclusions in section 5.

2 The CBA Algorithm

2.1 Clustering Method

The CBA approach is based on using data driven clusters as the units of analysis. The first step is therefore to form these clusters, based on fMRI data other than the data used to test the experimental paradigm of interest. We constrain the clusters to be contiguous regions in the brain. This is in contrast to many clustering methods in fMRI that ignore the contiguity constraint when grouping together similar voxels (e.g. Goutte et al. (1999), Windischberger et al. (2003)).

The gain when using the CBA approach in the ability to detect a larger proportion of truely active brain areas (the gain in power) will be larger as the degree of homogeneity of the clusters (the proportion of the voxels in the cluster that are truly active) is larger. The scientific importance of homogenous similarly activated clusters was discussed in the introduction, here we will focus on its importance in terms of the ability to discover activated clusters and the interpretability of cluster FDR. The level of noise in the cluster average time course is, by definition, smaller than the noise in the a voxel-by-voxel time course. At the same time, if the cluster contains only a few activated voxels then the average time course SNR can still be smaller than the SNR in the activated voxels in the cluster, making it harder to detect the cluster activation than it is to detect individual voxel activation. Weighing these opposing factors, we favored a clustering algorithm that will produce small clusters, even if this means that the activated clusters will not correspond to whole activated modules but only to subsets of modules.

The spatial structure of the data is taken into account by allowing only neighboring voxels to belong to the same clusters. The neighborhood is taken per volume, ie where every voxel has 26 neighbors. Since the neighborood extends across slices, the slices need to be corrected for different acquisition times prior to clustering. This correction is especially important if the slice acquisition order is interleaved.

The clustering algorithm is as follows:

1. For each voxel, the correlation with each of its neighbors is computed.

2. For every voxel, the neighbor with the highest correlation is found (after adjusting correlation values for distance on acquisition grid, see below). Note that this is not a symmetric property: given a voxel i that is maximally correlated with neighbor j, voxel j may be maximally correlated with another of its neighbors, k.

3. Each voxel and its maximally-correlated neighbor define an initial region, and if the same voxel is in two or more regions these regions are joined together, iteratively until the process terminates in non-overlapping clusters.

Figure 1 is a schematic display of step 3 in the algorithm. (For visualization reasons it is shown on a slice rather than volume.) Note that the resulting clusters' sizes and shapes are data driven, unlike most smoothing methods where neighboring voxels are joined into neighborhoods of fixed sizes and shapes over which the signal is averaged.

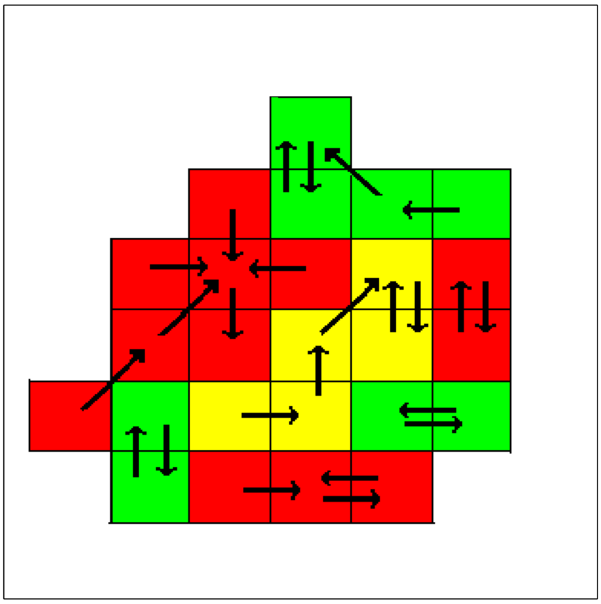

Figure 1: A graphical display of step 3 of the clustering algorithm. Each arrow starts from a different voxel and points to its maximally-correlated neighbor. The different colors denote the clusters obtained for this set of voxels.

To counteract biases in the comparison of correlations among nearest neighbors that have unequal distances on the acquisition grid, the raw correlation values were adjusted as follows. Let $\rho(d)$ be the correlation at distance d and let $\rho(0) \equiv 1$. Our goal is to keep the ratio of attenuations of the correlations between $\rho(1)$ and $\rho(d)$ constant relative to the attenuations between $\rho(0)$ to $\rho(1)$ for any $(d > 1)$. We estimate this constant robustly using the median correlations at each distance. For example, for $d = \sqrt{2}$ (eg, diagonal neighbors on the same slice), the constant is $c_{\sqrt{2}} = \frac{1/m_1}{m_1/m_{\sqrt{2}}}$, where m_1 and $m_{\sqrt{2}}$ are the median correlation values of horizontal/vertical neighbors and $\sqrt{2}$ distance diagonal neighbors respectively. Our adjusted correlation, $\hat{\rho(1)}$, should therefore satisfy $c_{\sqrt{2}} = \frac{1/\hat{\rho(1)}}{\hat{\rho(1)}/\rho(\sqrt{2})}$, i.e. $\hat{\rho(1)} = \sqrt{\rho(\sqrt{2})m_1}\sqrt{\frac{m_1}{m_2}}$. Similarly, for between slice neighbors that are a distance $d = \sqrt{3}$ apart,

$\hat{\rho(1)} = \sqrt{\rho(\sqrt{3})m_1}\sqrt{\frac{m_1}{m_3}}$ where m_3 is the median correlation of the between slice diagonal neighbors.

Figure 2 shows the correlation adjustment process for correlations between a representative voxel and its neighbors. The voxel is paired with the neighbor with whom the adjusted correlation is maximal.

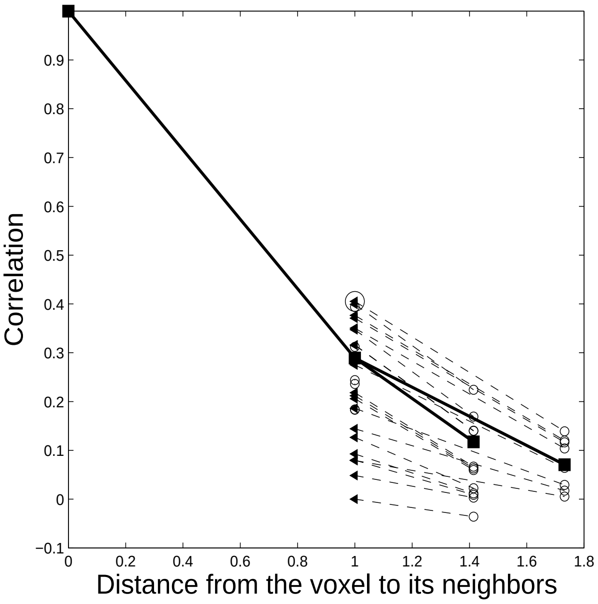

Figure 2: A graphical display of the procedure used to adjust the nearest-neighbor correlation values for unequal distances on the acquisition grid. The raw correlations are noted by a circle and displayed as a function of the distance from the voxel at study. The adjusted correlations are noted by a black triangle at distance 1, connected with a dashed line to the corresponding raw value. The maximal correlation voxel is highlighted by an enclosing large circle (in this example, it belonged to a $d = \sqrt{3}$ nearest-neighbor). The median correlations are noted by black squares.

2.2 FDR on Clusters

The resulting clusters from the preparatory scan serve as our units of analysis in subsequent analysis of the experimental data. For each cluster we calculate the average time-course of its constituent voxels, and use it as the cluster's signal. Then, p-values for the clusters are calculated. Each p-value is uniformly distributed under the null hypothesis that none of the voxels in the cluster are active.

Even though using clusters rather than voxels reduces the extent of the problem of multiple hypotheses testing , this analysis still involves testing thousands of clusters (eg, if we test at the 0.05 level of significance, then even if all 1000 brain clusters are (in truth) non-active, we will declare, on average, 50 clusters as active). One way to tackle this increased probability of making false discoveries is to control the False Discovery Rate (FDR). The FDR is the expected proportion of false discoveries among the discoveries. Setting our threshold level at 0.05, this means that we expect that no more than 5% of the discoveries to be false discoveries on the average. In our case a discovery may be a detected voxel in single-voxel analysis or a detected cluster in CBA. Note that for a rightfully-detected cluster, one can only conclude that it contains at least one active voxel, not that all voxels are truely active.

The BH procedure (Benjamini and Hochberg, 1995) has been adopted in the fMRI community for controlling the FDR at any desired level q while testing voxels (see Genovese et al. (2002), Stanley and Rubin (2003)), with implementations in software packages such as SPM and Brain Voyager. The results of Storey (2003) provide a Bayesian interpretation to the FDR. For fMRI this implies that the posterior probability that the cluster is not active given that it was detected is less than q.

The BH procedure makes use of the m p-values, calculated one for each voxel for testing its activation. Sorting these p-values we get $P_{(1)} \leq \ldots \leq P_{(j)} \leq \ldots P_{(m)}$. Then find the largest p-value among all those satisfying $P_{(j)} \leq q\frac{j}{m}$, call it $P_{(k)}$, and declare the k voxels whose p-value is less or equal to $P_{(k)}$ as active.

The procedure can be equivalently presented, and motivated, by describing it as an "adjustment" made to the raw p-values. If we choose $P_{(j)}$ as the threshold to separate activated voxels from not activated ones, j voxels will be chosen as active. Denoting by m_0 the number of voxels that are not (truly) activae (out of the entire sample of m voxels), on average $P_{(j)}m_0$ non activated voxels will be (falsely) declared as active. Thus a crude estimate of the proportion of false discoveries is $\frac{P_{(j)}m_0}{j}$. There may be a larger p-value for which the crude estimate may be even smaller. Hence, taking a greedy approach we may replace the estimate by $min\{\frac{P_{(i)}m_0}{i}|i \geq j\}$. Since we expect the number of truly active voxels to be a small fraction of m (as typical in brain imaging experiments), we may (pessimistically) approximate m_0 by its upper bound m without great loss in sensitivity. Thus, we get the "BH FDR-adjusted p-values" as follow:

$$P_{(j)}^{BH} = min\{\frac{P_{(i)}m}{i}|i > j\}$$

The BH adjusted p-values can now be compared with the desired level of FDR, say $q = .05$, and all those voxels for which $P_{(j)}^{BH} \leq q$ be declared as active.

The BH procedure, in either form, controls the expected proportion of falsely discovered voxels among all voxels discovered at the desired level. Note that while the above provides an intuitive argument that the procedure controls the FDR, the actual proof is much more involved (see Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001)). In Benjamini and Yekutieli (2001) it is proved to hold when the P-values at the different voxels are independent and under a technical condition, called positive regression dependence, that holds when the noise in the data is Gaussian with nonnegative correlation across voxels and the tested hypotheses are one-sided. According to Genovese et al. (2002) this is a reasonable assumption in fMRI. They argue that while strict independence is hard to verify and will often fail with neuroimaging data, it is often approximately true in the sense that the correlations are local and tend to be positive. Nichols T. (personal communication) has verified the assumption of fMRI data. This assumption is widely accepted in fMRI literature. Moving to clusters, the resulting test statistics also satisfy this technical condition, since if voxels are positively correlated so are cluster averages:

$$cov(\sum_{i=1}^{n} a_i X_i, \sum_{i=1}^{m} b_i Y_i) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j cov(X_i, Y_j) \geq 0 \text{ if}$$
$$cov(X_i, Y_j) \geq 0 \text{ and } a_i > 0, i = 1, \ldots, n, b_j > 0, j = 1, \ldots, m$$

where $a_i = 1/n, b_j = 1/m$ for the averages of clusters of size n and m respectively and the covariance between every pair of voxels $cov(X_i, Y_j)$ is non-negative $i = 1, \ldots, n, j = 1, \ldots, m$. We checked this assumption for our analysis reported in section 4.1 and found that the average correlation between the clusters was 0.57 and 0.25 in the block design and event related experiments respectively, with only one statistically nonsignificant negative correlation, at -0.02 ($p - value > 0.4$).

In CBA we use the same procedure on the p-values obtained for the clusters, replacing the total number of clusters m_c for the total number of voxels used above. Thus, the procedure controls the expected proportion of falsely discovered *clusters* among all clusters declared active. Note that a falsely discovered cluster is a cluster that contains no active voxels, and correspondingly a truely discovered cluster is a cluster that contains at least one active voxel. The point of view taken here is similar to the one taken in (Pacifico et al., 2004) in the sense that a cluster is considered a discovery rather than an individual voxel, although the methods proposed there and here are very different in principle as well as in detail.

Of course one should bear in mind that with CBA we give up the control of FDR on voxels. Thus, the FDR on voxels may be in certain situations higher than the FDR on clusters, especially if there are many non-homogenous clusters that contain both activated and non-activated voxels. We believe that researchers are interested in these flexible units of analysis for which conclusions are taken, rather than in the artificially generated voxel units. Thus we emphasize the control of FDR of clusters rather than the FDR of voxels.

The FDR methodology is geared to handle a predetermined family of hypotheses of fixed size. Here the number of hypotheses and their identity may vary from one realization of the preparatory scan to the other. But since the clustering step is performed on the preparatory scan it is independent of the analysis step, so using a conditioning argument the FDR is still controlled.

2.3 Adaptive FDR

The testing of clusters rather than voxels reduces the extent of the multiple hypotheses testing problem as the number of clusters tested m_c is smaller than the number of voxels tested m. In fact the reduction when using the clustering method in Section 2.1 is to at least $\frac{m}{2}$. The number of tests conducted can be further reduced by restricting the analysis to clusters within regions of interest (ROI) rather than searching over the entire brain for activity. Such ROI can either be predefined (eg anatomically), or extracted from the experiment that is already being used to define clusters. The restriction to an ROI is the same as in the common ROI analysis approach, but while the common approach tests for activation of the entire ROI as a single unit, our testing units remain clusters, those ones which are within the ROI. Thus, CBA in combination with ROI analysis can be viewed as helping us search for activation within subregions of the ROI.

In a successful choice of ROI, the potential proportion of activated clusters out of all tested clusters within the ROI is much larger than when analyzing the entire brain. This offers an opportunity to use an adaptive method that estimates this proportion, and use it instead of the more conservative value m_c (for a similar observation see Genovese et al. (2002)). This will increase the proportion of the clusters detected as active out of the active ones (the power). Recall that when we motivated the BH procedure in Section 2.2 we bounded the number of non-active units by the total number of units tested. In the adaptive procedure, we try to estimate the number of non-active units and plug in the estimate.

In particular, we make use of the adaptive two stage procedure introduced by Benjamini, Krieger and Yekutieli (Technical Report RP-SOR-01-03, URL http://www.math.tau.ac.il/st/) on cluster units. The ordered cluster p-values are $P_{(1)} \leq \ldots \leq P_{(m_c)}$. First, we run the BH procedure as before, and get $k_1 = \max\{i : P_{(i)} \leq q\frac{i}{m_c}\}$ clusters declared as active. Next, we estimate the number of null clusters, m_{0c}, by $\hat{m_{0c}} = (1+q) * (m_c - k_1)$. Finally, we use the BH procedure with $q^* = q\frac{m_c}{\hat{m}_{0c}}$, i.e. $k_2 = \max\{i : P_{(i)} \leq q\frac{i}{\hat{m}_{0c}}\}$. As the proportion $\frac{m_{0c}}{m_c}$ is smaller, the gain in power in using the adaptive procedure rather than the BH procedure is expected to be larger.

Benjamini, Krieger and Yekutieli prove that it controls the FDR under independence of the test statistics, and argue that this is also the case under the PRDS assumption (see Section 2.2). In section 3.4 we show that the adaptive FDR procedure preserves an FDR level of 0.05 for simulated signals that take into account the fMRI dependency structure. Other adaptive procedures exist in the literature (see Benjamini et al. (2005) for a review). The only other method with proven FDR control under independence of the test statistics, making use of a different estimator of m_0, is in Storey et al. (2004).

3 Methods

3.1 fMRI data acquisition

Scanning was performed on a 3 Tesla head-only Siemens Allegra MRI Machine. A head coil was used for structural scans (transmit/receive; Nova, MA). Functional data were acquired with a flexible four element array of surface coils (receive only; Nova, MA) fit into the head coil (transmit); the array elements were placed over the occipital lobe and temporal lobes to maximize signal from these regions. A set of 16 high-resolution slices oriented parallel to the lateral fissure were acquired using a T1-weighted spin echo sequence (TR=600 ms, TE=9.1 ms, flip angle=90?). Interslice distance was 4 mm (no gap, interleaved acquisition); resolution was 128 x 128, FOV

192 mm, resulting in 4 x 1.5 x 1.5 mm voxels. Functional (T2*-weighted) EPI images (TR=2s in the localizer and block-design experimental runs, TR=1s in event-related experimental runs; TE=30ms; flip angle=90?) were acquired using the same slice prescription as the T1-weighted spin echo images, except that the in-plane resolution was 64 x 64, resulting in 4 x 3 x 3 mm voxels. The slices completely covered the ventral occipital and temporal lobes. Functional data were superimposed on the T1-weighted spin echo images so that regions of activation could be anatomically localized. The number of whole volume acquisitions varied between experiments (see below).

3.2 Experimental Design and Visual Stimuli

Lateral Occipital Complex (LOC) localizer. Observers viewed grayscale images of objects and phase-scrambled controls (maintaining the amplitude spectrum of each images Fourier components but randomizing their phase rendered the objects unrecognizable). The intact and phase-scrambled images were presented in a pseudorandomized order in an event-related design for X sec followed by X sec blank each. There were 32 exemplars from each category, repeated X times. In addition, a third trial type consisting of an X second blank interval was intermixed X times, providing temporal jitter to increase the efficacy of the design. Order of the three trial types was counterbalanced and optimized using m-sequences (Buracas and Boynton, 2002). The stimulus presentation was preceded by X seconds and followed by X seconds of fixation. A fixation point was present on the screen at all times and the observer was asked to maintain fixation for the duration of the experiment. Each image was 11.25 x 11.25 degrees of visual angle and successive images were jittered +/- 0.6 degrees. Observers performed a 1-back task (X probes per run). Two runs were performed during the scanning session.

Illusory contour (IC) and Salient Region (SR) stimuli. The IC stimulus was a Kanizsa square: four pacman inducers arranged so that they create the impression of a large central square in front of four circular disks (Kanizsa, 1979). The corresponding control, no-IC stimulus, consisted of the same inducers flipped outwards so that the illusory square disappeared. The SR stimulus consisted of inducers resembling those of the IC except that their corners were rounded and they were misaligned so that crisp bounding ICs were no longer perceived, although the impression of an enclosed region remained. The corresponding control, no-SR stimulus, consisted of the same inducers flipped outwards. (For more details and figures of the stimuli see

Stanley and Rubin (2003).) In the block-design experimental runs, observers viewed alternating 16 sec blocks of experimental and control conditions (separate runs for ICs and SRs). Within each block, the image reversed contrast every 1 sec. Eight blocks of experimental and control stimuli were presented in a single run. In the event-related design observers viewed all four stimulus types in a pseudorandom order. Each trial consisted of the presentation of a stimulus for 1 second, followed by a two second blank interval. In addition to one trial type for each condition, a fifth trial type consisted of a 3 second blank interval, providing temporal jitter to increase the efficacy of the design. There were 25 trials from each of the 4 experimental conditions and 24 blank trials, preceded by 10 seconds and followed by 6 seconds of fixation. Trial order was counterbalanced and optimized using m-sequences (Buracas and Boynton, 2002). A fixation point was present on the screen at all times and the observer was asked to maintain fixation for the duration of the experiment. On

Stimulus presentation. Visual Stimuli were generated using Matlab and Psychtoolbox (Brainard, 1997), (Pelli, 1997) and fed into an Eiki LC-XG100/4267 LCD projector (1024 x 768 pixels, 60 Hz) with an extra focusing lens installed. The projected image appeared on a plastic rear-projection screen, and observers viewed it in a mirror mounted on the head coil.

3.3 Data analysis

Functional data were corrected for head motion using a customized MCFLIRT (Jenkinson et al., 2002) script. Each scan was then corrected for differences in slice acquisition time using the FSL function slicetimer. Finally, time course data were preprocessed to remove linear trends using the robust loess method (Cleveland and Devlin, 1988) from each voxel independently.

Data from the localizer runs were processed using the clustering algorithm in 2.1, producing the clusters to be used in statistical testing on the experimental runs. In addition, we defined as our LOC ROI all clusters discovered by the BH procedure on clusters (section 2.2), at level 0.001. (Note that since the clustering and the testing were performed on the same part of the experiment, the expected FDR on clusters may be greater than 0.001, but this should be of no concern at this stage.) Using the clusters and ROI from the localizer runs, data from the experimental runs were analyzed as follows. On the entire brain, we performed both CBA and voxel-by-voxel analysis using the BH procedure. On the ROI, we performed in addition the two analyses using the adaptive procedure.

The procedure for calculating the p-values basically follows Worsley et al.

(2002). The calculation is based on the average cluster time series in CBA, and on the voxel time series for voxel-by-voxel analysis.

A general linear model with AR(1) errors was used (see e.g. Worsley et al. (2002)). The hemodynamic response was modelled as a difference of two gamma functions

$$h(t) = (\frac{t}{5.4})^6 e^{-\frac{(t-5.4)}{0.9}} - 0.35(\frac{t}{10.8})^{12} e^{-\frac{(t-10.8)}{0.9}}$$

and convolved with the external stimulation which was modelled as $\sum_{i=1}^{k} X_{ti}\beta_i$ with $i = 1, \ldots, k$ different stimuli. We followed Worsley et al. (2002) for estimation of the coefficients β_i and the calculation of the p-value without spatially smoothing the AR(1) parameter.

Matlab code which implements the CBA algorithm and procedures described above, as well as the data for the fMRI example presented in this paper, is available in http://www.math.tau.ac.il/∼ybenja/CBAforFMRIstat.

3.4 Validation using Simulations

The simulations we performed had two purposes. First, we wanted to validate that the adaptive FDR procedure does not exceed the predefined FDR rate under typical fMRI dependency. Second, we wanted to compare the performance of cluster based and voxel based analysis on data where ground truth was known.

Setting A 64×64 slice was chosen for the comparison. The slice contained several hundred clusters, with an average size of 16 voxels per cluster. We designated n clusters containing overall m voxels to have activations in the first part of the experiment, and approximately half these clusters were designated to have activations in the second part of the experiment. The values of (n, m) examined were $(2, 1), (5, 3), (10, 5)$, and $(20, 10)$.

The measured signal (i.e. signal+noise) of voxel v within cluster c at time t was

$$y_{cvt} = \mu_{ct} + a_{ct} + \epsilon_{cvt}.$$

The signal μ_{ct}, in an active cluster c at time t was set to $\mu = 3$ in the first part of the experiment. In the second part, we examined $\mu = 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75$. The signal level in null clusters was set to zero.

The components of signal variation between clusters a_{ct} were drawn independently from a normal distribution with mean zero and standard deviation 3.

The spatially correlated noise components ϵ_{cvt} were simulated (independently for every time point t) by convolving white Gaussian noise with a spatial Gaussian kernel (FWHM $15mm$, with a convention that a voxel is $1 \times 1 \times 1mm^3$) using 2 dimensional (64×64 pixels) processes. In our simulations ϵ_{cit} has mean zero and standard deviation 6.

The time series for each part of the experiment was of length 100. Under each configuration of signal and noise, 100 simulation repetitions were performed.

The simulations were performed in Matlab (version 6.5).

Simulation Data Analysis On the first part of each simulated experiment, we clustered the data using the clustering algorithm in 2.1. We also defined as our ROI all clusters discovered by the BH procedure on clusters at level 0.05. Note that since the clustering and the testing were performed on the same part of the experiment, the expected FDR on clusters may be greater than 0.05.

Using the clusters and ROI from the first part of the simulated experiment, we analyzed the simulated data in the second part as follows. On the entire slice, we performed both CBA and voxel-by-voxel analysis using the BH procedure. On the ROI, we performed in addition CBA and voxel-by-voxel analysis using the adaptive procedure. The FDR level was estimated by averaging over the simulations the proportion of false discoveries among the discoveries at the appropriate units: voxels or clusters (recall that a discovery of a cluster is false if it contains no active voxels). We compared the performance of the analysis methods in terms of power. Power can be defined in many ways. We measured power as the proportion of discoveries out of all potential true discoveries. In fMRI terms, this translates to the proportion of detected voxels that are truly active out of all truly active voxels. For CBA, power was also measured by the proportion of detected clusters out of all active clusters. (Recall that the FDR is taken to be the same for the CBA and the voxel-by-voxel analysis.) In general, other possible measures of power include: the probability of making at least one true discovery; the probability of finding all potential true discoveries; or finally as one minus the expected proportion of missed discoveries out of all nondiscoveries. While we believe the measure we chose is the most adequate for fMRI, the advantage of CBA over voxel-by-voxel analysis is likely to present itself also with the other measures.

4 Results

4.1 CBA Results on an fMRI Vision Experiment

To evaluate the advantage that CBA may offer over voxel-by-voxel analysis, as well as the advantage of the adaptive FDR procedure, we tested it on data acquired in an experiment that we knew to have yielded a relatively small difference between experimental and control conditions. In a previous publication (Stanley and Rubin, 2003) we analyzed the responses to illusory contour (IC) and salient region (SR) stimuli compared with their corresponding control stimuli presented in a block design. An ROI analysis performed on the average time courses of the observers functionally defined lateral occipital complex (LOC) showed significant effects for both types of stimuli, but a voxel-by-voxel analysis yielded activation in few individual voxels, and none for some subjects. We therefore tested CBA on this paradigm. In addition, we reran the experiment on one of the observers and added an event-related run which interleaved all four stimulus types (ICs, SRs and their corresponding controls; see also Methods). The results presented below are a representative sample.

The LOC ROI computed with CBA on data from the event-related localizer runs consisted of 207 voxels, grouped into 20 clusters. The clusters ranged in size from 2 to 47, with a median of 9 voxels per cluster and a standard deviation of 9.3. Figures 3 and 4 show the discoveries in the experiments of interest within the ROI from two representative slices, 10 and 11. Every cluster within the ROI is indicated with a different color; note that clusters extend across slices, which is why noncontiguous voxels within a slice can belong to the same cluster. Figure 3 shows the results obtained for the event related experiment using CBA with a BH procedure at the 5% FDR level. Activated clusters within the ROI are indicated by white outlines. Overall, 44 voxels were found to be active, grouped into 4 clusters of sizes 12, 12, 9 and 11. In contrast, voxel-by-voxel analysis performed on the same data with a BH procedure at the 5% FDR level failed to detect any activation.

Figure 4 shows the activated clusters in slices 10-11 in the block design IC vs. Control experiment computed with three different procedures. Again, activated clusters within the ROI are indicated by white outlines. The top panel shows the result of CBA with an adaptive FDR procedure at the 5% level. Overall, 131 voxels were found to be active, grouped into 9 clusters (sizes 47, 13, 9, 11, 8, 10, 12, 12 and 9). For comparison, the middle panel shows the activated clusters when the BH procedure was used instead of the adaptive procedure. This procedure discovered considerably less activation. For example, note that the pink and green ROI clusters in slices 10-11 are marked as discoveries only using the adaptive procedure (outlined in top panel but not middle panel). Overall, only 56 voxels were found, grouped into 5 clusters (a subset of the clusters using the adaptive FDR; sizes 13, 9, 10, 12 and 12). Finally, the bottom panel shows the result of a voxel-by-voxel analysis with an adaptive FDR procedure at the 5% level. Overall, 41 voxels were found to be active, coming from 14 different clusters. When the BH procedure was used instead of the adaptive procedure, 38 voxels, coming from 13 different clusters, were found to be active.

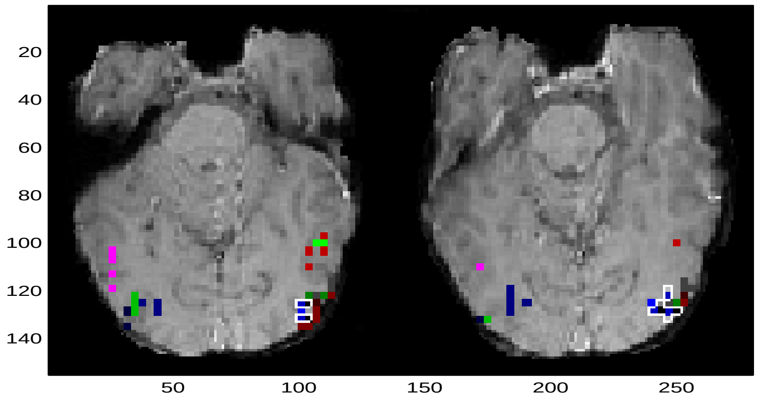

Figure 3: Activated clusters computed using CBA in two representative slices, 10 and 11. Colored voxels belong to clusters that comprise the LOC ROI, with different colors for different clusters (noncontiguous voxels in a slice can belong to the same cluster since those extend across slices.) White outlines indicate activated clusters within the ROI in the event-related experiment, obtained using CBA with the BH procedure at the 5% FDR level. Voxel-by-voxel analysis on the same data, with BH procedure at the 5% FDR level, yielded no activated voxels.

To further evaluate the gain in CBA over voxel-by-voxel analysis, we reanalyzed the data from the localizer runs as follows: a voxel-by-voxel FDR analysis at the 5% level of the two localizer runs detected 239 voxels; repeating the analysis on just one of the localizer runs detected only 15 voxels; a CBA at the 5% level on the same run detected 168 voxels grouped into 13 clusters (the event related IC experimental run was used to create the clusters). Treating the discoveries from the voxel-by-voxel analysis based on the two localizer runs as the ground truth, the power of CBA was 0.29 compared with only 0.06 for voxel-by-voxel analysis. Figure 5 shows the resulting activated regions in slice 09. The top panel shows the "ground truth" (dark blue). The middle panel shows the very few activations in voxel-by-voxel analysis on one localizer run only (dark blue). In contrast, the bottom panel shows the relatively large number of activations in CBA for the same run (each cluster is indicated with a different color).

4.2 Validation using Simulations

We present the analysis results for the following representative signal configuration: in the simulated localizer experiment, 5 clusters were active (82 voxels); in the simulated main experiment, 3 out of the 5 clusters were active (44 voxels). The results were similar for the other configurations.

Figure 6 shows that the FDR is below 0.05 for all analysis methods. Moreover, the right graph in figure 6 shows that the FDR of CBA using the adaptive procedure is higher than that of CBA using the BH procedure, and similarly the FDR of voxel-by-voxel analysis using the adaptive procedure is higher than that of voxel-by-voxel analysis using the BH procedure.

Figure 7 show the power improvement of all analysis methods over the voxel-by-voxel analysis using the BH procedure, as a function of signal size μ. When μ is extremely low, then both CBA and voxel-by-voxel analysis were barely able to detect activations. However, as μ increased, the CBA analysis detected more activations than the voxel-by-voxel analysis. When μ was very large, both methods of analysis performed equally well. Clearly, the advantage of CBA over voxel-by-voxel analysis is largest when μ is not too low or too high - the zone of interest in practice.

5 Discussion

We presented an algorithm to calculate activation maps based on analysis units which are independently defined clusters of voxels. The clusters are defined as contiguous volumes of voxels which were correlated with each other more than with their other (contiguous) neighbors in an independent run (eg, a localizer run). the method is based on the proposition that the units of testing for activation in the brain should be larger than a voxel but smaller than an entire region of interest. We argued that fMRI analysis is likely to be more meaningful at the cluster level than at the voxel level. The regions constructed by the clustering method are more likely to be related to functional modules in the brain, leading to increased SNR per unit tested. This improves our ability to detect activations, and may enable a fruitful search for the interactions between brain regions.

We showed that CBA discovered larger and more contiguous activation areas than voxel-by-voxel analysis. On the other hand we do not argue that our clustering algorithm is optimal according to some well defined criterion. How to define the relevant criterion and develop the optimal segmentation is a question for further research. For example, incorporating prior knowledge from the anatomical image into the clustering algorithm may result in a much more powerful procedure (e.g. by applying a grey matter extraction procedure based on the anatomical image prior to CBA).

The approach we currently take defines the units of testing conservatively, so the units are fairly small and are more likely subunits of the true activated modules than unions of such. This reduces the possibility of introducing more noise than signal into the clusters. Also, the small size of clusters will keep low the number of voxels that are not truly active in discovered regions (aggregates of clusters). Moreover, although the number of clusters (and their identity) may vary from one realization of the experiments to the other, the discovered regions created from these small "building blocks" can still be quite similar. At the same time, bear in mind that the inclusion of a few voxels that were not truly active in a detected cluster may be a small price to pay for the gain CBA offers in terms of increased discoveries. Note that commonly used pre-processing steps such as smoothing can also introduce into discovered regions voxels that are not truly active. Furthermore, when the statistical testing is performed on individual voxels, in such cases (after smoothing) the expected proportion of erroneously rejected voxels can be higher than the nominal FDR level. In contrast, in CBA, although discovered clusters may contain voxels that are not truly active, one is still able to control the FDR at the cluster level.

We explained the meaning of FDR on clusters, and introduced the adaptive FDR on clusters. In assessing the FDR we give the same weight to every cluster. Benjamini and Hochberg (1997) also introduced the weighted FDR, which may be especially appropriate here. For example, we may want to control a weighted FDR with weights proportional to the size of the clusters, which means on the one hand that it is important to reject a large cluster since it considerably increases the weight of the total discoveries, but on the other hand it also increases the weight of the errors if in fact it is an error. We are currently exploring size weighted FDR procedures and their suitability for fMRI data.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B Met.*, 57 (1):289–300.

Benjamini, Y. and Hochberg, Y. (1997). Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics*, 24:407–418.

Benjamini, Y., Krieger, A. M., and Yekutieli, D. (2005). Adaptive linear step-up false discovery rate controlling procedures. *Technical Report RP-SOR-01-03, URL http://www.math.tau.ac.il/st/.*

Benjamini, Y. and Yekutieli, Y. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29 (4):1165–1188.

Brainard, D. (1997). The psychophysics toolbox. *Spatial Vision*, 10:433–436.

Buracas, G. and Boynton, G. (2002). Efficient design of event-related fmri experiments using m-sequences. *NeuroImage*, 16:801–813.

Cleveland, W. and Devlin, S. (1988). Locally-weighted regression: An approach to regression analysis by local fitting. *J. Am. Statist. Assoc.*, 83:596–610.

Genovese, C., Lazar, N., and Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage*, 15:870–878.

Goutte, C., Toft, P., Rostrup, E., Nielsen, F., and Hansen, L. (1999). On clustering fmri time series. *NeuroImage*, 9 (3):298–310.

Jenkinson, M., Bannister, P., Brady, J., and Smith, S. (2002). Improved optimisation for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17(2):825–841.

Kanizsa, G. (1979). *Organization in Vision.* Praeger, New York.

Pacifico, M., Genovese, C., Verdinelli, I., and Wasserman, L. (2004). False discovery control for random fields. *Journal of the American Statistical Association*, 99 (468):1002–1014.

Pelli, D. (1997). The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, 10:437–442.

Penny, W. and Friston, K. (2003). Mixtures of general linear models for functional neuroimaging. *IEEE Transactions on Medical Imaging*, 22:504–514.

Stanley, D. and Rubin, N. (2003). fmri activation in response to illusory contours and salient regions in the human lateral occipital complex. *Neuron*, 37:323–331.

Storey, J. (2003). The positive false discovery rate: a bayesian interpretation and the q-value. *Annals of Statistics*, 31:2013–2035.

Storey, J., Taylor, J., and Siegmund, D. (2004). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society, Series B*, 66:187–205.

Windischberger, C., Barth, M., Lamm, C., Schroeder, L., Bauer, H., Gur, R., and Moser, E. (2003). Fuzzy cluster analysis of high-field functional mri data. *Artificial Intelligence in Medicine*, 29 (3):203–223.

Worsley, K., Liao, C., Aston, J., Petre, V., Duncan, G., Morales, F., and Evans, A. (2002). A general statistical analysis for fmri data. *NeuroImage*, 15:1–15.

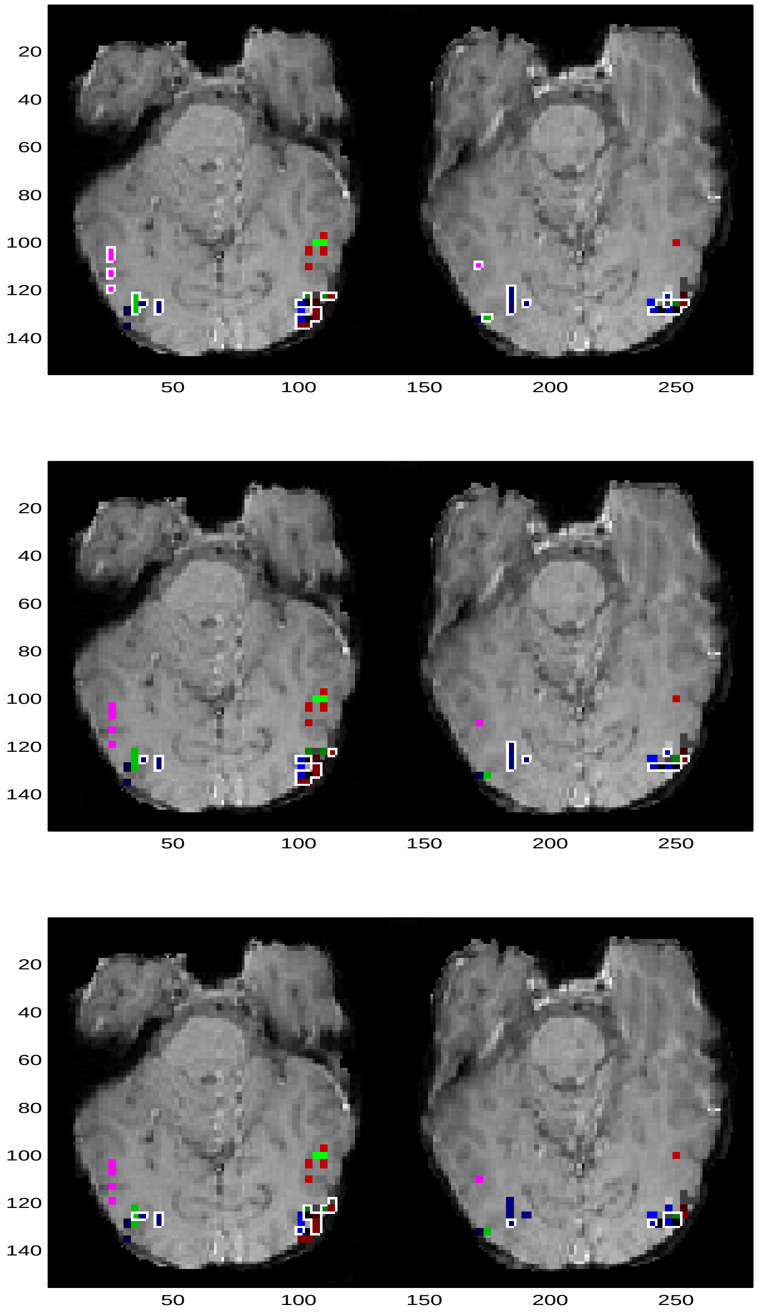

Figure 4: Activated clusters with the LOC ROI in the block-design IC vs. Control experiment computed with three different procedures (clusters indicated in white outlines; sample slices 10-11). Top panel, CBA with an adaptive FDR procedure at the 5% level. Middle panel, CBA with BH procedure at the 5% FDR level. Bottom panel, voxel-by-voxel analysis with adaptive FDR at the 5% level.

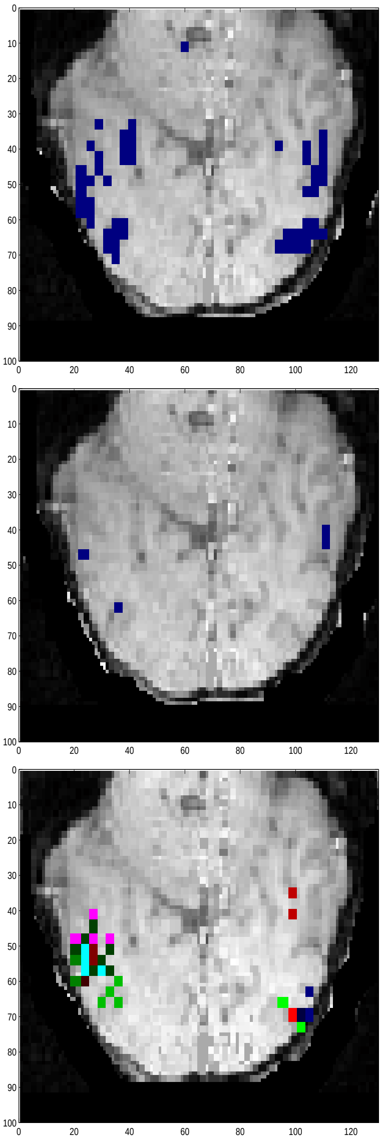

Figure 5: Illustration that the enhanced discoveries made by CBA when limited data are available correspond well with the activation discovered with voxel-by-voxel analysis when more data are available. Top panel, results of voxel-by-voxel analysis on data from both localizer runs (here and below, FDR at 5%). Middle panel, the same analysis on data from only one of those runs yields few discoveries. Bottom panel, results of CBA on the single-run data (clusters derived from the event related IC experimental run).

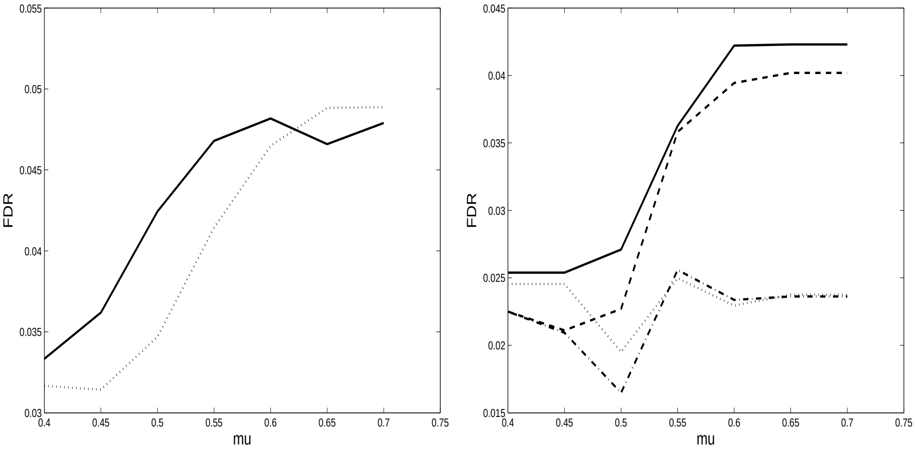

Figure 6: FDR as a function of μ when the analysis is done on the entire slice (Left) for (1)cluster based analysis (solid line) and (2) voxel-by-voxel analysis (dotted line) and on the ROI (Right) for (1) CBA using the adaptive procedure (solid line); (2) CBA using the BH procedure (dotted line); (3)voxel-by-voxel analysis using the adaptive procedure (dashed line); (4) single voxel analysis using the BH procedure (dot and dashed line). Note that the FDR is always below 0.05, and that for both CBA and single voxel analysis the FDR using the adaptive procedure is closer to the desired 0.05 than when using the BH procedure, making the adaptive procedure more powerful.

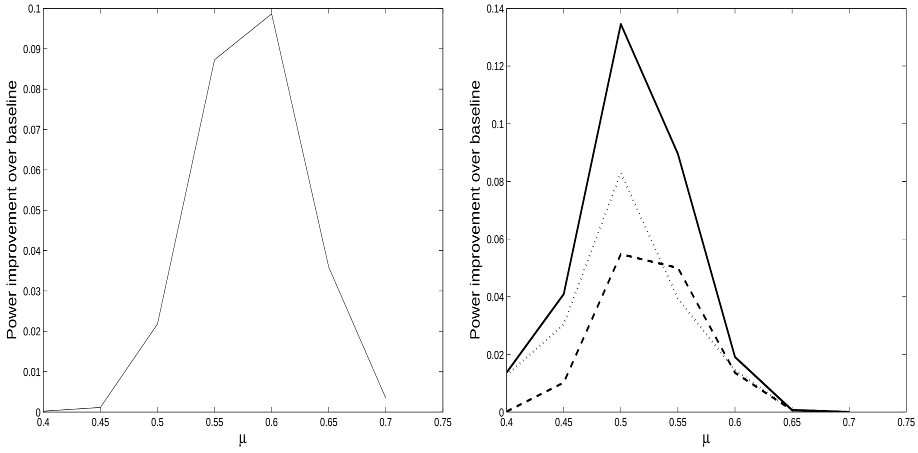

Figure 7: Power improvement over voxel-by-voxel analysis using the BH procedure as a function of signal size μ when the analysis is done on the entire slice (Left) of CBA and on the ROI (Right) of (1) CBA using the adaptive procedure(solid line) (2) CBA using the BH procedure (dotted line) and (3) voxel-by-voxel analysis using the adaptive procedure (dashed line). Note that the power advantage is largest when μ is not too small and not too large. The most powerful analysis method is clearly (1).