Threshold-Free Cluster-Enhancement

Addressing the problem of threshold dependence in cluster inference

Steve Smith[1] & Tom Nichols[2,1]

[1]Oxford University FMRIB Centre, UK [2]GSK Clinical Imaging Centre, UK

Background

Cluster-level inference is generally found to be more sensitive than voxel-level inference, as it makes use of local spatial neighbourhood information to boost belief in extended areas of signal [Friston]. Standard cluster inference only makes use of the size of a cluster, whilst several alternatives to cluster size inference have been proposed which incorporate height information, including joint cluster-height inference [Poline,Hayasaka], and cluster-mass inference [Bullmore].

Existing cluster size methods, however, require an explicit choice of cluster-forming threshold. As users know well, minor adjustments in this threshold can result in dramatic changes in the resulting inferences, and there is no objective method by which a sensible threshold can be chosen.

In this work we propose a new approach to cluster inference which retains the power of the cluster size statistic while avoiding the specification of a cluster-forming threshold. The result is a continuous image where the intensity at a given voxel is a natural measure of local cluster support, or put another way, measures the signficance of all possible clusters that contain that voxel.

Method

The TFCE approach aims to enhance areas of signal that exhibit some spatial contiguity without relying on hard-threshold-based clustering.

For a particular cluster-forming threshold h, consider creating an image of cluster extent, $e(h)$. For a particular voxel, $e(h)$ is the extent of the cluster that contains that voxel (if h is greater than the voxels's intensity then $e(h)$ is zero). The crux of the method is how to combine the set of these extent images for all h's.

Most generally, we consider summing powers of the extents, weighted by a power of the threshold:

$$TFCE = \sum_h e(h)^E h^H. \quad (1)$$

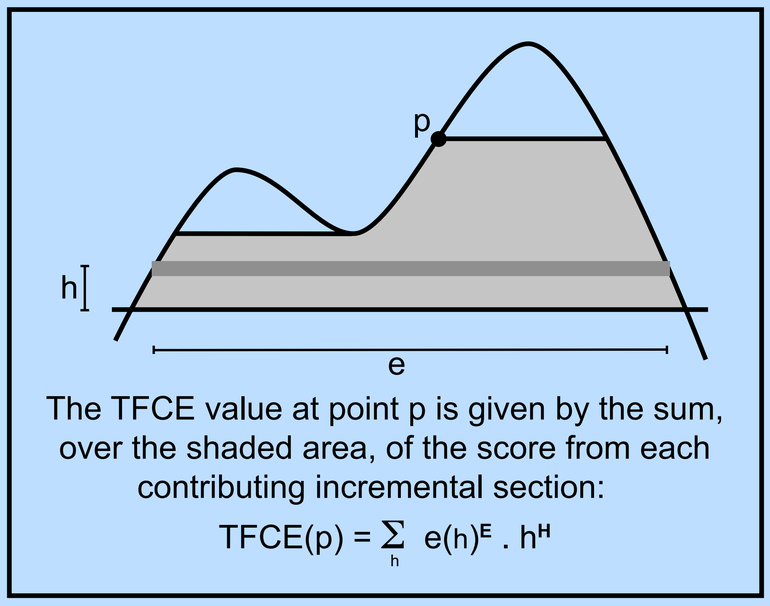
While this would seem to create two parameters (E & H) in the place of one (h), in fact there are principled choices for these parameters. For example, one approach is to choose E & H to equate significance over different h's, by transforming each $e(h)$ into a P-value, and then combining P-values over h's. For example, using Gaussian random field theory (RFT) for 3D images, one can show that, approximately,

$$P_{e(h)} \propto \exp(\ e(h)^{2/3}\ h^2\ ). \quad (2)$$

Combining all of the P-values with Fisher's P-value combining method, which sums the $-\log$ P-values, immediately results in the TFCE statistic, with $E = 2/3$ and $H = 2$.

The general algorithm is re-summarised for the 1D data case in the figure. Each voxel's TFCE score is given by the sum of the scores of all "supporting sections" underneath it; as the height h is incrementally raised from zero up to the height (signal intensity) of a given point p, the image is thresholded at h, and the single contiguous cluster containing p is used to define the score for that height h. This score is simply the height h (raised to some power H) multiplied by the cluster extent e (raised to some power E).

Note that this means that while p can get support from the lower parts of overlapping clusters (such as the left peak in the figure), only those parts of the overlapping clusters which lie below the point of inflection between the two clusters are allowed to contribute. This therefore achieves a balance between allowing overlapping clusters to contribute to each other's significance (desirable, particularly given that there is no unambiguous way of deciding at this stage whether they should be considered separately or jointly), while still giving separate scores for the distinct local maxima. The TFCE output image can easily be turned into true P-values (either uncorrected or fully corrected for multiple comparisons across space) via permutation testing.



The avoidance of hard cluster-forming thresholding means that TFCE should be more stable than traditional cluster-based thresholding, in terms of the potential for large changes in the output being caused by small changes in the input. The goal, however, was also to remove the dependence on the arbitrary choice of the cluster-forming threshold in the traditional approach, as there has never been a principled way of setting the cluster-forming threshold. Below we present evaluations on a wide range of settings for E and H, and it is primarily our empirical observations which have led us to select recommended values (as well as more qualitative considerations, and the theoretical ones presented above).

Note some conceptual similarities between such cluster "enhancement" and various previous work such as scale-space (Lindeberg, Worsley, Coulon) and wavelet pre/post-processing (Van De Ville, Fadili, Penny). In further work we will further investigate such similarities in theory and practice, as well as the interaction (e.g.) with spatial pre-smoothing.

Evaluations

We used ROC testing in two different ways to evaluate TFCE and other thresholding methods. First, ("FWE") we controlled the family-wise-error to set the FPR in the ROC testing, by evaluating in what fraction of pure noise images *any* false positives were seen for a given threshold; this is the most important measure as it corresponds to the way in which people generally threshold in practice (i.e., null-hypothesis testing, correcting for multiple comparisons over space, etc.). FWE ROC curves are quantified in terms of AUC (area under curve), integrated over the FPR range from 0:0.05. Second, ("simple") we set the FPR according to the proportion of false positive voxels found in non-signal data.

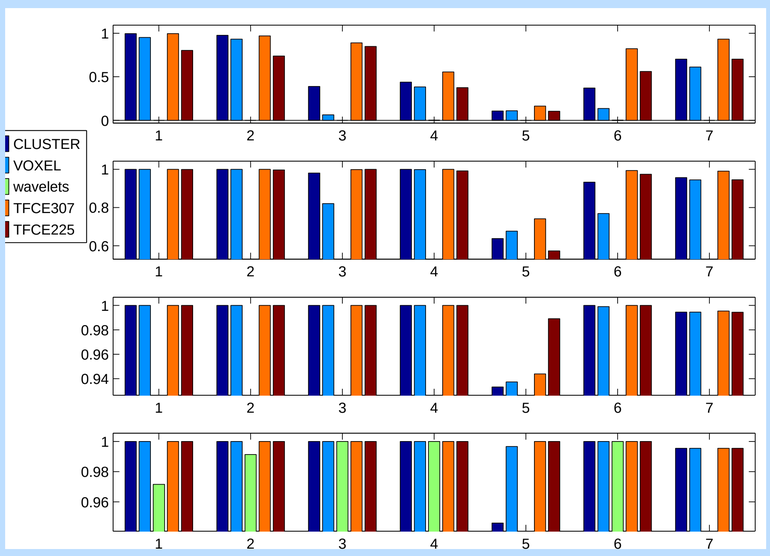
Slices through the 7 test signals are shown below (two nearly touching blobs, two distant blobs, several long thin bars, hollow cube with a dot in the centre, Penny's 3 blobs, Penny's cross [Flandin] and high-resolution real FMRI data). To the signal we added Gaussian white noise to give a range of peak SNR values: 1, 2, 5 and 10.

The first method that we tested was simple Gaussian smoothing followed by voxelwise thresholding ("VOXEL"). Gaussian kernels of FWHM of 0, 0.2, 1, 2, 4, 6 and 8 were applied. The second method that we tested was cluster-based thresholding ("CLUSTER"), comprising Gaussian-smoothing (same range as above), initial t-stat thresholding at a range of levels (0.5, 0.75, 1, 1.5, 2, 2.5, 3, 3.5, 4 and 5) and then using the cluster extent as the test statistic. The third method that we tested was TFCE. For completeness, we preceded the TFCE algorithm with the same range of data smoothing as described above, and for each smoothing extent, we varied E and H over the range 0.01, 0.1, 0.5, 1, 2 and 3.

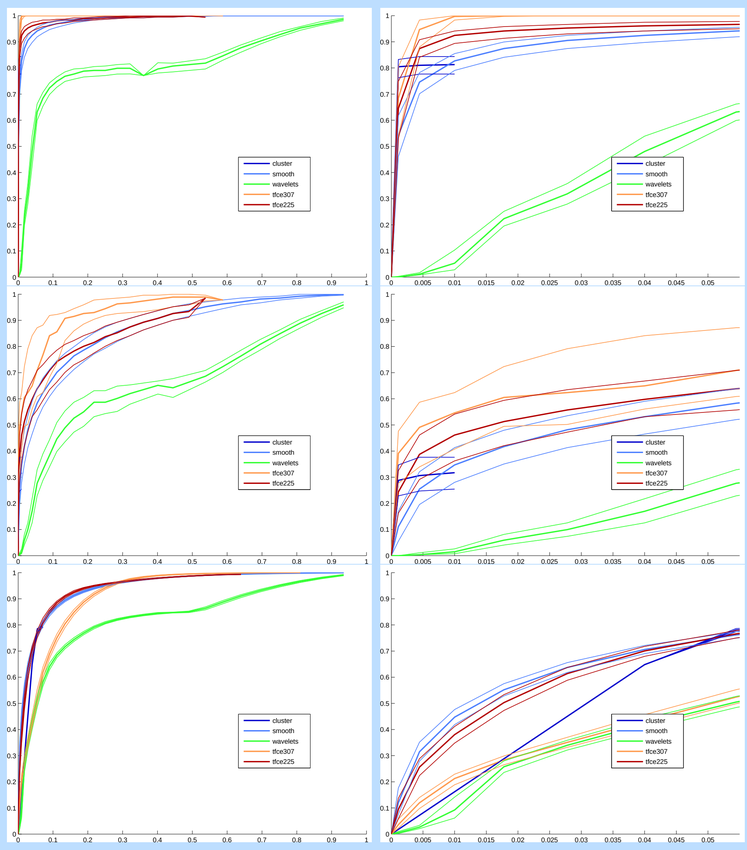
For the results shown below, we chose the settings for each method that gave the best results over all tests (all SNRs and all test images). For VOXEL, this was sigma=3mm, for CLUSTER, this was was sigma=1mm smoothing and cluster-forming threshold=2, for TFCE we selected two sets of parameters to present, "TFCE225"=presmoothing=1mm,E=0.5,H=1 and "TFCE307"=presmoothing=2mm,E=1,H=3.

The range of h thresholdings effectively used within TFCE captures signals of differing spatial scales, somewhat reminiscent of wavelet methods. Hence for our final method we tested wavelet enhancement/thresholding, as implemented in WSPM [Van De Ville] to obtain de-noised statistic images, which we then thresholded voxelwise to feed into the ROC evaluations. We used the default options with slight modification (3D, redundant 'multiple', 2-level 'iters' decomposition), as recommended in personal communication with the toolbox authors.

The figure below shows the results for these settings of the different algorithms. The plots show FWE-ROC normalised AUC values for the 7 test signal shapes (left-to-right) and 4 SNR levels (top-to-bottom). At the lowest SNR, TFCE with more smoothing is overall the best method. At the highest SNR, TFCE with less smoothing is the best.



The figure below shows "simple" ROC curves for (top-bottom): test signal 2, SNR 1; test signal 5, SNR 1; test signal 7, SNR 2. For each processing method, the mean and IQR (over 100 signal+noise images) ROC plots are shown. Right: expansion of the 0:0.05 FPR region of the same plots.



Acknowledgements

We are very grateful to the UK EPSRC for funding, to Matthew Webster for software coding, to Karla Miller for providing the high-resolution FMRI data and to Mark Woolrich and Adrian Groves for useful discussions.

References

Bullmore, IEEE-TMI 18:32-42 1999. Flandin, NeuroImage 34:1108-1125 2007. Friston, NeuroImage, 4:223-234 1996. Hayasaka, NeuroImage 23:54-63 2004. Poline, NeuroImage, 2:83-96 1997. Van De Ville, IEEE EMB Mag 25:65-78 2006.

[†]**Address for correspondence: Oxford Centre for Functional Magnetic Resonance Imaging of the Brain (FMRIB), Department of Clinical Neurology, University of Oxford, John Radcliffe Hospital, Headington, Oxford OX3 9DU, UK**