Threshold-Free Cluster Enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference

FMRIB Technical Report TR08SS1

(A related paper has been accepted for publication in NeuroImage)

Stephen M. Smith[1] **Thomas E. Nichols**[2,1]

[1]**FMRIB (Oxford University Centre for Functional MRI of the Brain), Dept. Clinical Neurology, University of Oxford**

[2]**GlaxoSmithKline Clinical Imaging Centre, UK**

Abstract

Many image enhancement and thresholding techniques make use of spatial neighbourhood information to boost belief in extended areas of signal. The most common such approach in neuroimaging is cluster-based thresholding, which is often more sensitive to finding true signal than voxelwise thresholding. However, a limitation is the need to define the initial cluster-forming threshold. This threshold is arbitrary, and yet its exact choice can have a large impact on the results, particularly at the lower (e.g., t,z$<$4) cluster-forming thresholds frequently used. Furthermore, the amount of spatial pre-smoothing is also arbitrary (given that the expected signal extent is very rarely known in advance of the analysis). In the light of such problems, we propose a new method which attempts to keep the sensitivity benefits of cluster-based thresholding (and indeed the general concept of "clusters" of signal), while avoiding (or at least minimising) these problems. The method takes a raw statistic image and produces an output image in which the voxelwise values represent the amount of cluster-like local spatial support. The method is thus referred to as "threshold-free cluster enhancement" (TFCE). We present the TFCE approach and discuss in detail ROC-based optimisation and comparisons with cluster-based and voxel-based thresholding; we find that TFCE gives generally better sensitivity than other methods over a wide range of test signal shapes and SNR values. We also show an example on a real imaging dataset, suggesting that TFCE does indeed provide not just improved sensitivity, but richer and more interpretable output than cluster-based thresholding.

Keywords

Clustering, thresholding, inference, Gaussian field theory

1 Introduction

Many image enhancement and thresholding techniques make use of spatial neighbourhood information to boost belief in extended areas of signal. The motivation for considering neighbouring voxels is to increase sensitivity to regions of signal that are more spatially extended than the noise coherence. The most common such approach in neuroimaging is cluster-based thresholding, which is generally implemented as 2 stages, generally assuming that some spatial smoothing has previously been applied: first, threshold the raw statistic image (or parametric map) and identify resulting clusters of contiguous supra-threshold voxels, then calculate a p-value for each cluster on the basis of its size/mass (e.g., using Gaussian field theory or permutation testing).

Cluster-based thresholding is popular as it is often perceived to be more sensitive to finding true signal than voxelwise thresholding; for example, cluster-based inference is more powerful when the spatial correlation length of signal exceeds that of noise and vice-versa for inference on the height of maxima [Friston et al., 1996]. However, a limitation is the need to define the initial cluster-forming threshold (e.g., threshold the raw t-statistic image at $t>2.5$). This threshold is arbitrary, and yet its exact choice can have a large impact on the results, particularly at the lower (e.g., t,z$<$4) cluster-forming thresholds frequently used. It has not been possible to give more objective advice than "broader signals are best detected by low thresholds and sharp focal signals are best detected by high thresholds" [Friston et al., 1994]. A second problem is that the initial *hard* thresholding introduces instability in the overall processing chain; small variations in the data around the threshold level can have a large effect on the final output.

A third problem, common also to simple voxel-based thresholding, is that the amount of spatial smoothing is itself arbitrary, given that the expected signal extent is very rarely known in advance of the analysis. Furthermore, one may well want to optimise sensitivity to different shapes and sizes of signal within one dataset simultaneously. An early attempt to address these issues was work on scale-space [Worsley et al., 1996a], but such approaches have not become widely-used, possibly because of the resulting increase in potential over-fitting of noise (and the associated increase in the number of multiple comparisons). Finally, a fourth problem is that it can be hard to directly interpret the meaning of (what may ideally be) separable sub-clusters or local maxima within very extended clusters (see the example shown in Section 5, though see also the discussion regarding multi-level inference in [Friston et al., 1996]).

In this paper we suggest a new method which attempts to keep the sensitivity benefits of cluster-based thresholding (and indeed the general concept of "clusters" of signal), while avoiding (or at least minimising) the problems listed above. The method takes a raw statistic image and produces an output image in which the voxelwise values represent the amount of cluster-like local spatial support. The method is thus referred to as "threshold-free cluster-enhancement" (TFCE). It is simple (Figure 1): each voxel's new value is given by the sum of the "scores" of all "supporting sections" underneath it; each section's score is simply its height h (raised to some power H) multiplied by its extent e (raised to some power E). The output value is therefore a weighted sum of all of the local clustered signal, without the need for a hard cluster-forming thresholding. For inference, the TFCE image can easily be turned into voxelwise p-values (either uncorrected, or corrected for multiple comparisons across space) via permutation testing.

In summary, to optimise the detection of both diffuse, low-amplitude signals and sharp, focal signals, we propose a simple but generic form of non-linear image processing that boosts the height of spatially distributed signals without changing the location of their maxima. This enables us to apply standard permutation testing to the height of the maxima of the resulting statistic image, while maintaining strong control over family-wise error. Critically, this avoids specifying a threshold on clusters, while sensitising the inference to a wide range of signal shapes; see Figure 1 (right). Although the form of this threshold-free non-linear enhancement may seem ad hoc, we provide a series of heuristics in the appendices, which motivate its form and the algorithm's parameters.

This paper first presents the TFCE approach in detail, and provides some illustration of its characteristics. Detailed optimisation and validation is then presented, using a range of simulated signal types and noise levels, and careful ROC testing to compare the power of TFCE, cluster thresholding, voxel thresholding and a spatial wavelet approach. In the ROC testing we investigate both the control of family-wise error (comparing sensitivity when correcting for multiple comparisons), and also voxelwise accuracy. TFCE appears to give sensible results, with generally better sensitivity and stability than the other methods. Appendicies are included which further justify the TFCE method and specific parameter choices.

2 Description of the TFCE approach

2.1 TFCE definition

The TFCE approach aims to enhance areas of signal that exhibit some spatial contiguity without relying on hard-threshold-based clustering. The image is passed through an algorithm which enhances the intensity within cluster-like regions more than background (noise) regions. The output image is therefore not *intrinsically* clustered/thresholded, but the hope is that after TFCE enhancement,

thresholding will better discriminate between noise and spatially-extended signal.

The TFCE algorithm is illustrated in Figure 1 (left). The solid curve shows a 1D profile through an example statistic image (e.g., an unthresholded t- or z-statistic image resulting from a neuroimaging data analysis). 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) h_p 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). Precisely, the TFCE output at voxel p is

$$TFCE(p) = \int_{h=h_0}^{h_p} e(h)^E h^H dh, \tag{1}$$

where h_0 will typically be zero, and E and H, 0.5 and 2 respectively (these choices are discussed in detail below). In practice this integral is estimated as a sum, using finite dh (for example, dh=0.1 if the input is a raw t or Z image). While this may seem like an arbitrary transformation, it can be seen as a generalization of the cluster mass statistic [Bullmore et al., 1999] (E=1, H=0), a type of weighted norm of cluster size (Appendix B) and a multi-threshold meta-analysis of Random Field Theory cluster P-values (Appendix C).

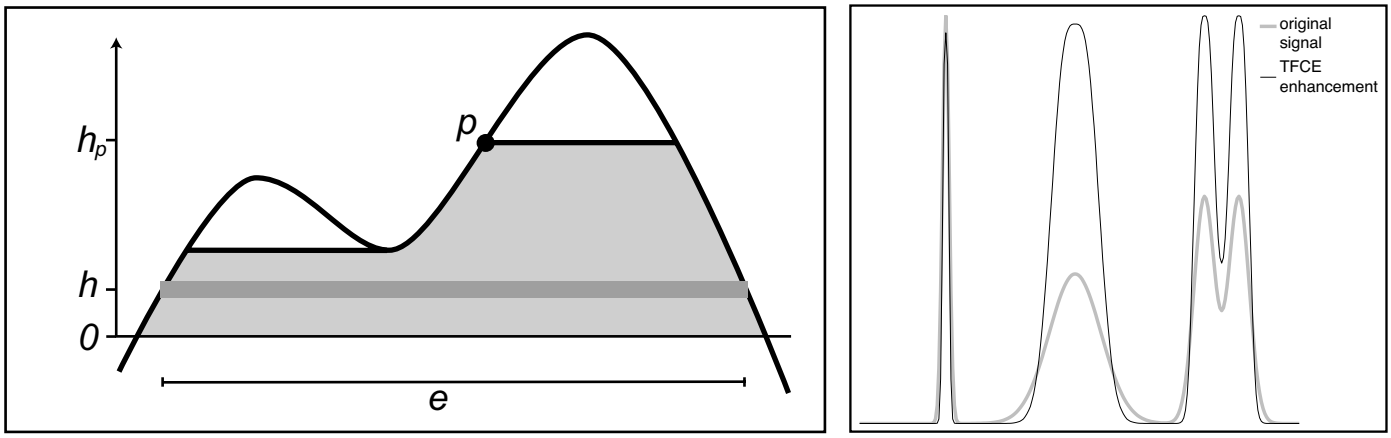

Figure 1: Illustration of the TFCE approach. Left: The TFCE score at voxel p is given by the sum of the scores of all incremental supporting sections (one such is shown as the dark grey band) within the area of "support" of p (light grey). The score for each section is a simple function of its height h and extent e. Right: Example input image and TFCE-enhanced output. The input contains a focal, high signal, a much more spatially extended, lower, signal and a pair of overlapping signals of intermediate extent and height. The TFCE output has the same maximal values for all three cases, and preserves the distinct local maxima in the third case.

Note that this means that although 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 local minimum between the two clusters are allowed to contribute. This therefore achieves a balance between allowing overlapping clusters to contribute to each other's belief (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.

By default, voxels that were originally negative will be given zero TFCE score. However, if it is desired to apply the enhancement to both negative and positive values, the original image can be simply negated, passed through the enhancement, re-negated and combined (via addition) with the positive enhancement.

Although TFCE creates a potentially distinct output value for every voxel, once could choose to only be interested in local maxima in the output, and consider that their position and value are descriptive of their whole local "cluster". One appealing property of the approach is that every local maximum in the input image is a local maximum in the TFCE outout, and vice-versa. This desirable property is unusual in signal transformations - for example, linear smoothing typically changes local maxima.

Another related property is monotonicity about local maxima; related to this, an iso-contour in the original image corresponds to an iso-contour in the TFCE image. This property guarantees that for any cluster in the input image (for any threshold h), there exists a threshold on the TFCE image which will produce the same cluster. Hence, while the TFCE transformation is nonlinear, it retains basic and important features of the data.

The cluster-enhanced output image can easily be turned into p-values (either uncorrected or fully corrected for multiple comparisons across space) via permutation testing. For example, to correct for multiple comparisons, one simply has to build up the null distri-

bution (across permutations of the input data) of the maximum (across voxels) TFCE score, and then test the actual TFCE image against that. Once the 95th percentile in the null distribution is found then the TFCE image is simply thresholded at this level to give inference at the $p<0.05$ (corrected) level.

2.2 Parameter selection

The avoidance of hard-thresholding-based cluster formation 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.

However, in TFCE we still have to choose parameters E and H. Our hope is that we can choose values for these parameters which give good results over a wide range of signal and noise characteristics, and the values can then be pre-fixed in a way that has never been possible with cluster-based thresholding. We show later in the evaluations section that this is indeed possible.

The obvious effect of increasing the height parameter H is to give more weight to higher clusters. If $H > 1$ then the TFCE score scales more than linearly with increasing statistic image intensity, which we consider likely to be desirable. Likewise, increasing E gives more weight to spatially larger clusters. If $E < 1$ then the TFCE score scales less than linearly with cluster size. When considered in the context of the different "supporting sections" evaluated within the TFCE summation, this is probably desirable; at the lowest values of h the sections can become very large, but these large low areas of support are not providing very useful spatial specificity, so one would not want their effect to scale with size in a strong way.

In a later section 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 such as those presented above). However, we note that these are in fact closely in line with the values suggested by an approximate analytical approach deriving from concepts in Gaussian field theory (see Appendix C).

2.3 Role of data smoothing

Both voxel-based and cluster-based thresholding rely, for maximum sensitivity, on smoothing of the data. The extent of the smoothing should ideally be matched to the signal extent, though of course this is rarley known in advance of the data analysis and so is impossible to optimise (though see more on this in Section 6). In contradistinction, the TFCE algorithm explicitly does not require spatial smoothing in order for a point in space to gain support from its neighbours that are part of the same "cluster"; the region of support can be identified without any recourse to regularisation through smoothing (see again Figure 1 for simple illustration of this). Furthermore, because TFCE avoids any hard thresholding steps, sensitivity to noise (and hence the need for spatial smoothing) should be reduced compared with hard-cluster-based thresholding.

It cannot be ignored, however, that the signals anticpated in neuroimaging are band-limited. Specifically, the signals lack high-frequency features which would suffer from smoothing. Hence, a small amount of smoothing may reduce noise without substantially degrading the signal. We show below both illustrations of the effect of a small amount of smoothing and full evaluations of this effect over a wide range of smoothing extents.

2.4 Illustrative signal+noise examples

We now show simple examples of applying smoothing-only, TFCE-only and smoothing-followed-by-TFCE to signal+noise data. The two examples are chosen to be towards the two extremes of "typical" SNR regimes; Figure 2 shows high-noise simulated data (SNR approximately 1) and Figure 3 shows fairly low-noise simulated data (SNR approximately 5). Note that if one expects to apply TFCE to a t-statistic image, these SNR values correspond to the peak t values, not the raw data SNR. The simulated data is 3D; each 2D image slice shown is taken from the centre of the relevant 3D image, while each 1D profile is taken from the line down the centre of that central 2D slice.

1000 random noise images were added to the original signal, so that for each processing method the mean output, and the variability about that, can be shown (the first column in the figures). The variability is represented by plotting the upper and lower bounds of the IQR (interquartile range). The row of "raw signal+noise" results shows clearly that the mean of 1000 signal+noise images gives a clean, unbiased estimate of the true signal; however, individual signal+noise images vary greatly about this mean (for these illustrations the data was not restricted to being positive). The corresponding example single image, and the 1D profile at the top

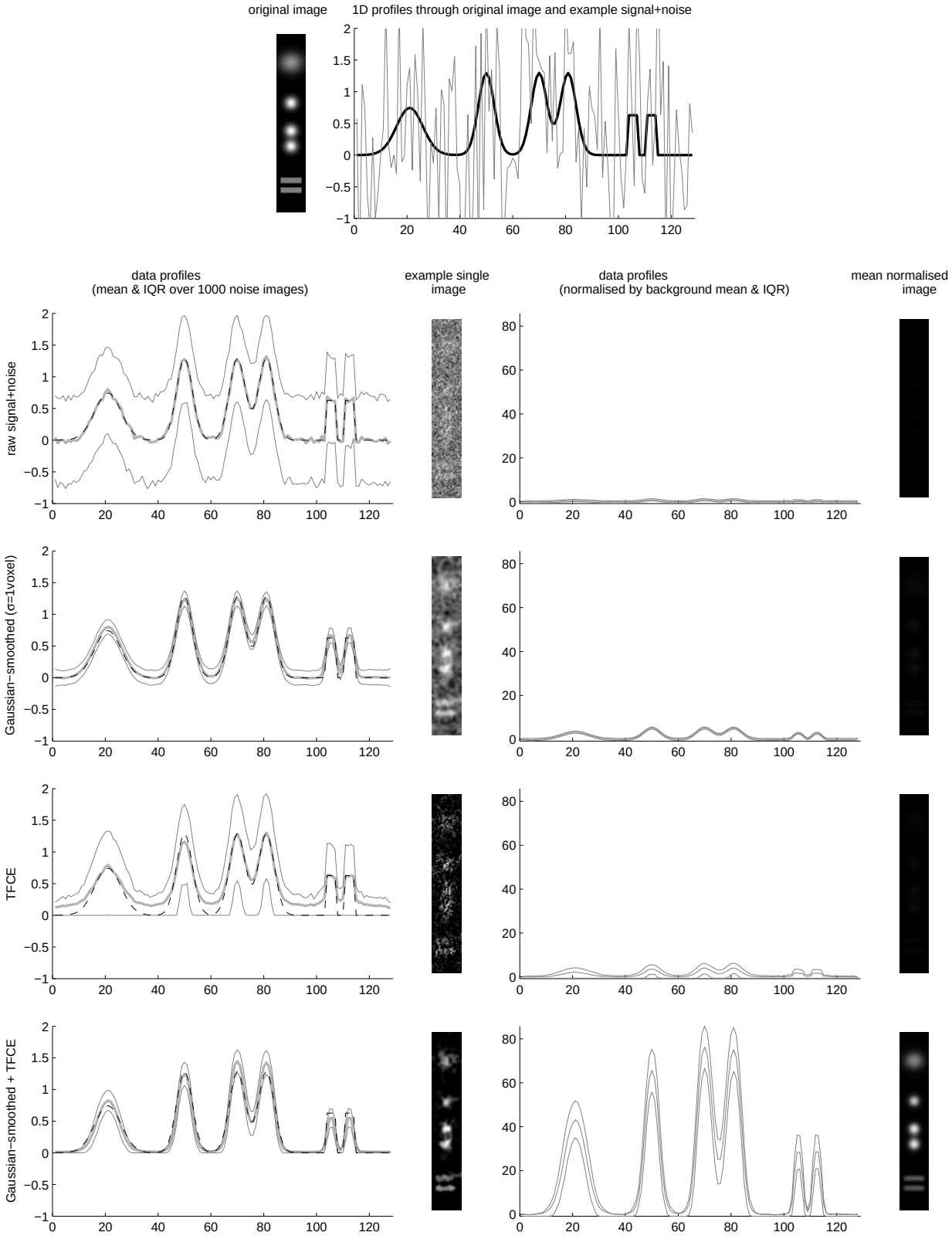

Figure 2: The effects of different processing options on low SNR data. The top row shows a slice from the 3D pure signal image, a 1D profile through this, and an example profile after adding noise. Row 2 shows the unprocessed signal+noise data, row 3 shows the effect of a small amount of smoothing, row 4 shows the output from TFCE, and row 5 shows the output from smoothing followed by TFCE. The first column shows the mean and IQR profiles over the 1000 different noise images. The original signal is shown as a dashed line. Although the variability in the signal after smoothing is comparable with smoothing+TFCE, the suppression of the background noise is much better with smoothing+TFCE. Example images are shown in column 2. Column 3 shows the mean±std profiles after normalising the data on the basis of the background mean and IQR. This shows how well separated the signal is from the noise after the different processing options. Smoothing+TFCE is considerably better than the other options at boosting signal while suppressing noise. Column 4 shows mean normalised images (all using the same intensity display range).

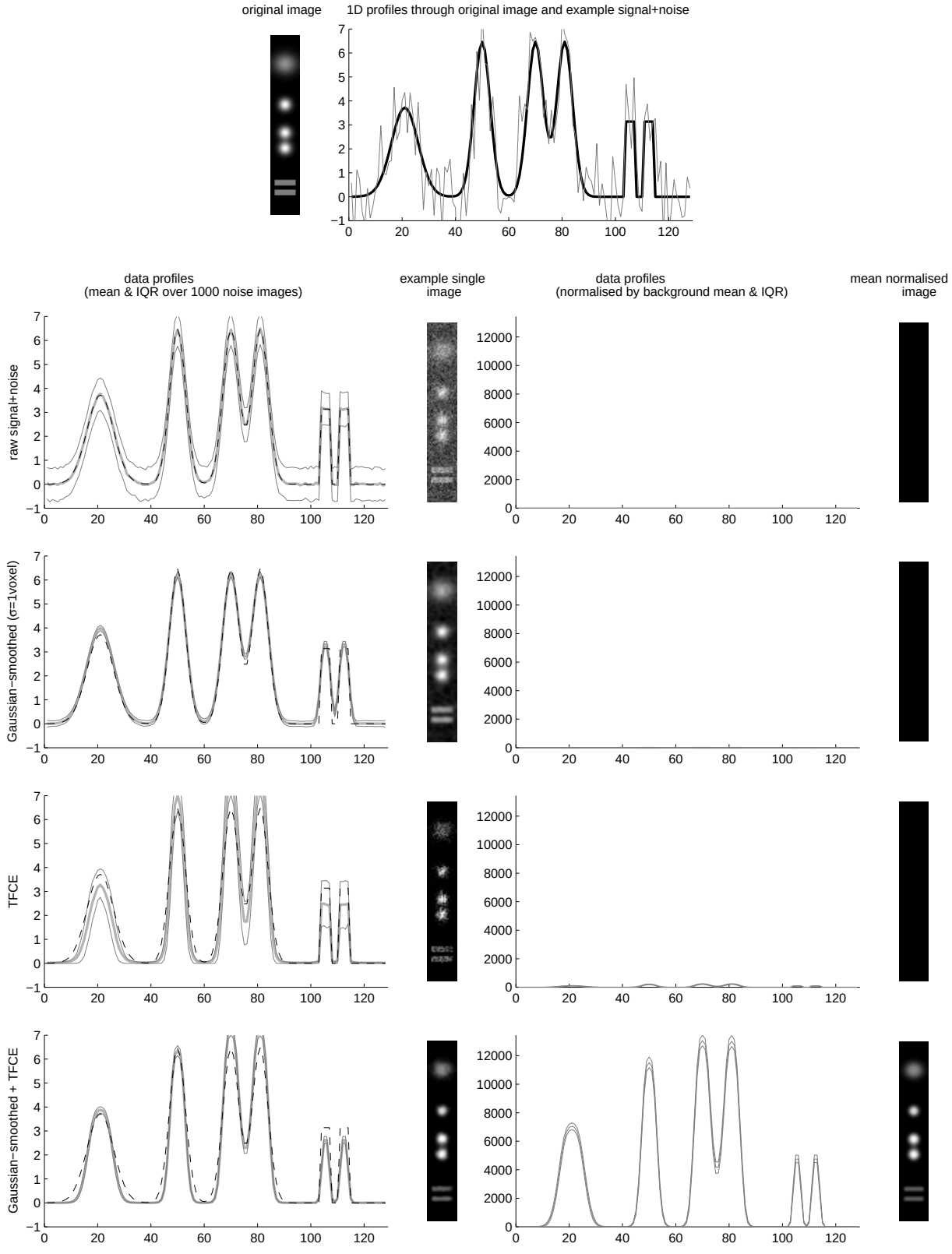

Figure 3: The effects of different processing options on high SNR data. See Figure 2 caption for details.

show just how difficult this input data is to recover the signal from, in the high-noise case.

Column 3 of the figures shows 1D plots of the mean and IQR for each processing option after normalisation using the background mean and IQR of the method. Here we took some of the zero-signal data points and estimated the average mean and IQR in the presence of no signal; the original mean and IQR plots were then normalised by this background mean and variation, to give $normdata = (data - mean_0)/IQR_0$. This normalisation therefore makes the different processing options' plots directly comparable, in terms of showing how well separated the signal is from the noise (this is an empirical, non-parametric analog to forming a Z-statistic).

The figures show that TFCE does a good job of shifting the signal away from the noise, in the case of the low SNR data, further aided by pre-smoothing.

3 ROC-based evaluations using simulated datasets

In this section, simulated data comprising several test image shapes are used to compare various enhancement/thresholding methods against each other, with ROC evaluations giving objective combined measures of specificity and sensitivity.

3.1 ROC methodology

An ROC (receiver-operator characteristic) curve, given a signal+noise image and the known ground truth, plots true positive rate (TPR) against false positive rate (FPR), as one varies a threshold applied to binarise the image. An ideal algorithm gives perfect true positive rate at zero false positive rate, i.e., the perfect ROC curve jumps immediately up to TPR=1 (y axis) for FPR=0 (x axis) and stays at 1 for all values of FPR. Hence a commonly-used single summary measure of the whole ROC curve is the AUC (area under curve); the higher the AUC, the better.

3.1.1 ROC taking into account correction for multiple comparisons

Standard ROC methodology was designed for a single inferential decision (e.g., tumour present somewhere in the image, "yes" or "no"). When multiple tests are considered, various TPR and FPR measures can be defined and plotted. For example, the free-response receiver operating characteristic (FROC) [Bunch et al., 1978] plots the proportion of true positive tests (among all possible positive tests) on the y axis versus the expected number of false positives per image on the x axis. Another approach, the Alternative FROC (AFROC) plot [Chakraborty and Winter, 1990] uses a different value for the x axis, the probability of any false positive detections anywhere in the image. As neuroimaging analyses typically seek to control the familywise error rate (the chance of one or more false positives anywhere, under the null), we use the AFROC plot to characterise our method. We limit our AUC calculation to FPR values between 0 and 0.05, since performance at FWE FPR in excess of 0.05 is not of interest.

Hence we take the following approach, using the AFROC definition for ROC testing, as this is the most appropriate way to objectively compare different spatial enhancement/thresholding methods:

1. Generate the raw ground truth 3D image ("signal").
2. Generate 1000 random Gaussian noise images (mean zero, unit variance) that, when added to the signal, give a specified SNR in the resulting 1000 signal+noise images.
3. Pass all noise-only and signal+noise images through the algorithm being tested.
4. Threshold the processed noise-only and signal+noise images at the full range of possible threshold values, computing the FPR and TPR at each:
    - **FPR**: For each threshold level, count the number of processed noise-only images which contain *any* supra-threshold voxels. This count (divided by 1000) gives the family-wise FPR for this threshold level (i.e., achieves full correction for multiple comparisons across space).
    - **TPR**: For each threshold level, use each of the 1000 processed signal+noise images (along with the original ground truth signal) to obtain an estimate of the TPR. We use the raw voxelwise TPR (fraction of non-background signal voxels

correctly reported), averaged over the 1000 signal+noise images (we also record the IQR of the TPR across the 1000 images, as a measure of the stability of the various algorithms being tested).[1]

5. Take the resulting ROC curve, and, using only the FPR range of 0 to 0.05, calculate the AUC.

The above approach, by estimating FPR from processed pure-noise data, avoids the need to determine what is "real" background in the signal+noise data after passing through a given algorithm, which can be problematic if the algorithm has a spatial aspect (e.g., smoothing followed by clustering, or TFCE). It is exactly what we want in the standard scenario of null-hypothesis testing which aims to explicitly control the FPR *in the presence of no true signal*; it tests sensitivity when the specificity is being controlled in the way that we generally require in practice.

This method of calculating TPR ignores the FP voxels in the signal+noise images that are spatially close to the true signal (as distinct from "real" FP voxels in the noise-only data), and in doing so does not weight, for example, against the smearing of estimated signal into background if smoothing is being applied. We partly make this choice because we are following as closely as possible a sensible definition of FWE, partly because we are primarily interested in signal detectability (as opposed to recovery/spatial-accuracy, though see below), and partly because this avoids problematic non-mononotonic ROC behaviour.

Note however, that because we are also interested in detecting as many of the signal voxels as possible (partly because we are interested in the spatial information in the thresholded images, and, even more importantly, because we are interested in detecting as many distinct "clusters" of signal as are truly present), we do not wish to use a family-wise TPR measure (alternatively known as a set-level approach [Friston et al., 1996]). Such a measure would only tell us whether there is any significant signal present *somewhere in the image* (or, put another way, whether the complete image, taken as a whole, can be considered significant), which is not of primary interest to us here, as we are still interested in keeping information about signal localisation; we are considering spatial statistical methods and wish to maximise sensitivity to localised information while making optimal use of any spatial extent present in the signal. We did in fact also carry out ROC testing using a family-wise estimation of true positives, but (in addition to the interpretational limitations described above) found the results quite unuseful, as such an approach is too sensitive to signal to usefully discriminate between methods tested.

We used this approach to compare various algorithms using a range of settings of algorithm-controlling parameters (such as initial threshold level in the case of cluster thresholding). We then selected the optimal overall algorithm parameter(s) for each method, and present plots comparing ROC AUC results across the different methods.

3.1.2 More "traditional" ROC testing

We also carried out ROC testing using simpler measures of FPR; these are not as meaningful in our context, for the reasons described above (primarily, we believe that one should care about true FWE when calculating FPR as this matches common neuroimaging practice), but do have the advantage of allowing more complete quantitation of signal recovery (*accuracy* as opposed to detectability), as well as matching the more "traditional" ROC approach. Here the voxelwise FPR (fraction of background voxels incorrectly reported as signal) is calculated in the presence of signal, and any bleeding of the signal into the background degrades the ROC results. For each algorithm we generated a separate ROC curve for each of the 1000 signal+noise test datasets, so that we could plot both the mean and IQR ROC curves, hence showing not just the mean ROC characteristics, but also the variability about this.

3.2 Simulated data

We generated 7 3D test signal shapes for feeding into the ROC analyses. These are shown in Figure 4; they cover a wide range of signal types, including small blobs, touching blobs and extended areas of activation.

The test signals do not contain signal strength very close to zero. For example, in the case of the real FMRI data, we thresholded a Z-statistic image at $Z>4$, deleted clusters with less than 50 contiguous voxels, truncated the Z values at 10, and divided by 10. The resulting non-background voxels therefore range from 0.4–1. The reason for not allowing the underlying signal to fall all the way to zero is that ROC testing fundamentally assumes that the "ground truth" is a predominantly binary concept; a given voxel

[1] To be specific: Let J be the number of noise realisations (in our case 1000), indexed by $j \in 1:J$. Let u be a given threshold, to be varied over the full range of statistic image intensity values. For TPR, using the signal+noise data, we compute $\text{TPR}_j = \sum_i [H_i \hat{H}_{ij}] / \sum_i H_i$, the average voxel-level true positive rate for realization j, where H_i is 1 if voxel i has a non-zero ("true") ground-truth signal, 0 otherwise, and $\hat{H}_{ij}$ is 1 if voxel i exceeds u in realization j of the processed signal+noise data. Then the ordinate value plotted is $\text{TPR} = \sum_j \text{TPR}_j / J$, the true positive rate averaged over signal voxels, and we also record the first and third quartiles. For FPR, using the noise-only data, we compute $\text{FPR}_j = \max(1, \sum_i \hat{H}_{ij})$, an indicator of any false positives for realization j. Then the abscissa value plotted is $\text{FPR} = \sum_j \text{FPR}_j / J$, the chance of any false positives searching over space.

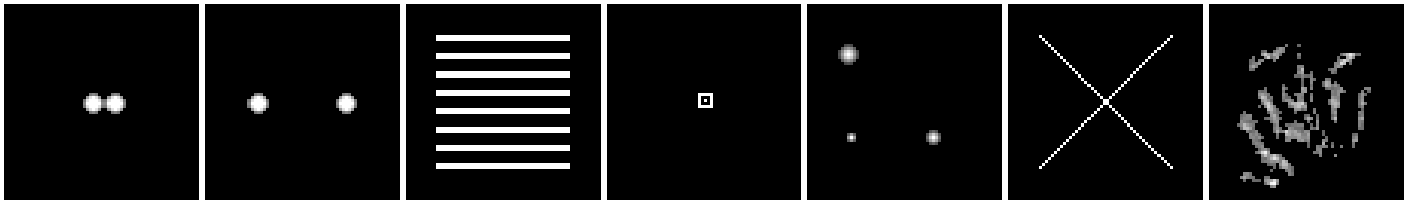

Figure 4: 7 3D test signals used in the ROC testing. Left to right: Two spherical blobs with slightly blurred edges, just touching; identical blobs, more considerably separated; several long thin rectangles; small hollow cube with a central dot; 3 Gaussian blobs (3D version of that used in [Flandin and Penny, 2007]); long thin cross (also matching that used in [Flandin and Penny, 2007]); real activation derived from high-resolution FMRI data [Miller et al., 2006].

either contains signal or it does not. Our initial analyses created test signals that were allowed to fall all the way to zero, rather than truncating them at some intermediate strength. However, it quickly became clear that this approach generates a highly undesirable test dataset; the presence of many low-strength voxels, which are hard to find in the presence of even low amounts of noise, adds significant variability to the entire ROC procedure, while not adding meaningful information about the success (relative or absolute) of the methods being tested. By allowing reasonable variation of signal strength within the test signals, and yet not allowing it to fall too close to zero, we believe that we have achieved a good compromise between meaningful, inpterpretable and precise ROC testing on the one hand, and realistic data simulation on the other.

Each test signal has a background value of 0 and a peak height of 1. We then scaled the signal by a factor or 1, 2 or 5 and added unsmoothed Gaussian white noise of standard deviation 1, to give a range of peak SNR values: 1, 2 and 5.

3.3 Methods tested

Voxel: The first "thresholding" method that we tested was simple Gaussian smoothing followed by voxelwise thresholding. Gaussian smoothing kernels of FWHM of 0, 1, 2, 3, 4 and 6 voxels were applied. After smoothing, the data was scaled so as to keep the noise standard deviation equal to 1, so that the images were still analogous to T/Z images. (Note that we would in practice advocate smoothing the original data, not the statistic image. In our simulations, the smoothing of the simulated "statistic image" followed by variance renormalisation is equivalent to smoothing original data.)

Cluster: The second method that we tested was standard cluster-based thresholding, comprising three stages. First, the data is Gaussian-smoothed (using the same same range as listed above). Second, the smoothed data is thresholded at a given level; we tested a range of 8 different cluster-forming thresholds: 0.75, 1, 1.5, 2, 2.5, 3, 3.5 and 4. Finally, contiguous clusters of supra-threshold voxels are formed (using 26-neighbour connectivity) and each cluster's test statistic is given by its extent (number of voxels in the cluster). This approach is a very commonly-used method of thresholding neuroimaging data, with cluster extent normally turned into cluster p-value using Gaussian field theory [Friston et al., 1994]. (In our ROC testing, where we know the true distinction between signal and noise, we have *exact* control of the FPR, whereas Gaussian field theory estimates p-values only approximately). The combination of 6 smoothing levels and 8 cluster-forming thresholds means that 48 different cluster "algorithms" are tested.

TFCE: 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. For each smoothing extent, we varied the E and H parameters over the range 0.1, 0.5, 1, 2 and 3, and also tested both 6-connectivity and 26-connectivity.

Wavelets: Finally, we considered a wavelet de-noising / signal-enhancing method [Van De Ville et al., 2007], as implemented in the WSPM toolbox. The fundamental goal in such approaches is very similar to that of TFCE, namely enhancing spatially-extended areas of signal in a flexible/adaptive way. In one analysis we used the default toolbox settings; in a separate analysis, we used custom parameter settings as advised by the toolbox authors (3D transform instead of 2D+Z, 2-level decomposition, and a redundant transform to increase shift invariance).

4 Simulation results

4.0.1 ROC taking into account correction for multiple comparisons

Each of the methods was tested using a range of parameter settings, as described above. By careful investigation of the complete set of results for every method (including looking across all possible settings, as well as looking at summary scores, taking the arithmetic and geometric means of AUC values across all 3D test signals and all SNR values), we selected the "optimal" parameter values for each method. For "voxel", the optimal smoothing was σ=3 voxels. For "cluster", we selected two settings, the first ("cluster-A": σ=1.5, cluster-forming-threshold=2.5) giving the best results overall, and the second ("cluster-B": σ=3, cluster-forming-threshold=1) chosen to illustrate interesting behaviour in some of the results. For TFCE, the optimal parameters were σ=1.5, H=2, E=0.5. With the "wavelet" testing, we were unable to achieve comparable ROC performance; AUC values were so far below all other methods tested that we have omitted these results in the figures. Results are shown in figure 5, where the y-axis is normalised AUC (i.e., 20$\times$AUC, as the area under the ROC curve was only estimated for FPR=0:0.05).

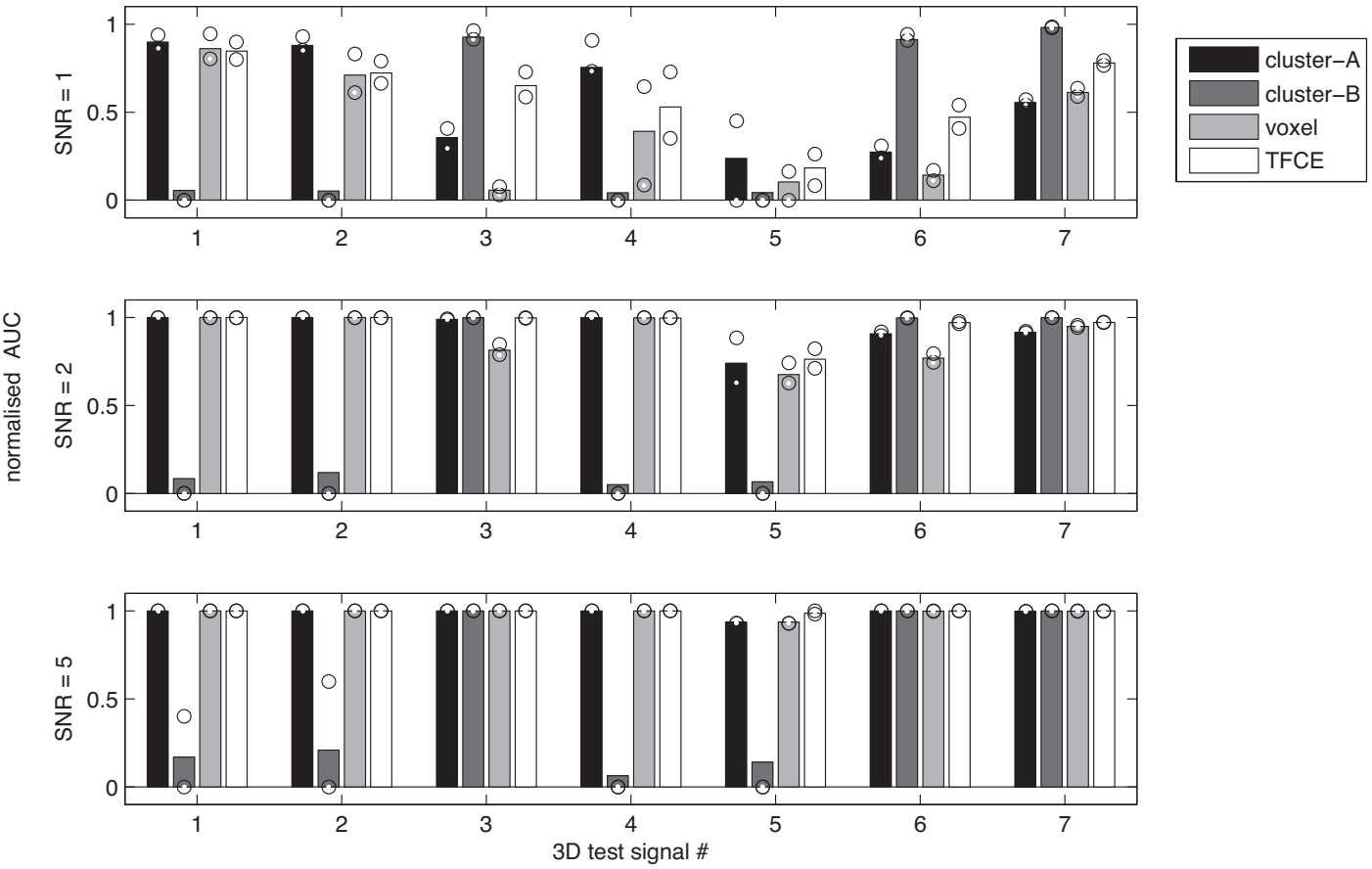

Figure 5: FWE-ROC area-under-curve values for 7 test signal shapes (left-to-right, in the same order as in Figure 4) and 3 SNR levels (top-to-bottom: 1, 2 & 5), derived from 1000 separate noise images add to the signals. The FPR is determined from the FWE, i.e. is the fraction of the 1000 noise images where *any* false positives are found in the noise-only data. The TPR is determined by the fraction of correctly labelled signal voxels, averaged over all 1000 signal+noise images. The circles mark the IQR of the AUC over the 1000 images. The AUC values are the area under the ROC curve for the FPR range 0:0.05, normalised by 1/0.05 so that the ideal value is 1.

Note that test signals 3, 6 and 7 are reasonably spatially extended, while signals 1, 2, 4 and 5 are more tightly focal. There are several points of interest in these results:

1. All methods work well at high SNR, except for the "cluster-B" (cluster-based thresholding using large spatial smoothing and a low cluster-forming threshold), when applied to more focal signal.

2. At medium SNR, again "cluster-B" performs very badly on more focal signal. Of the others, voxel-based thresholding performs less well on the more extended signals; TFCE performs best overall, with "cluster-A" nearly as good.

3. At low SNR, "cluster-B" performs well on the cases it is specifically tuned for (high noise, spatially-extended signal), and in these 3 cases (3, 6, 7) performs better than the other methods. However, it performs very badly on more focal signal.

4. At low SNR, voxel-based thresholding performs badly on extended signal.

5. At low SNR, TFCE performs slightly better on average than "cluster-A". Of interest are the results for test signal 5 (Penny's blobs), where TFCE and "cluster-A" have similar mean results, but cluster-based thresholding is significantly less stable, with the 25th percentile lying at zero.

4.0.2 More "traditional" ROC testing

Figure 6 shows more traditional (if less quantitatively useful, for reasons discussed above) ROC curves for a few interesting cases. In this testing, the FPR was calculated from the signal+noise data, and for each threshold level, a separate FPR and TPR was calculated for each of the 1000 signal+noise images, to allow the variability of the ROC curve to be visualised, as well as its median.

The cluster results do not extend all the way across to FPR=1. This is merely because the "threshold" being varied to generate the ROC curve for the cluster-based method is minimum cluster size; a minimum cluster size of 1 voxel corresponds to the highest FPR possible, and therefore an ROC curve cannot be meaningfully extended to the right of this value (though for the purposes of related AUC values, the curve is extended horizontally from its end point to FPR=1).

The top row shows clearly the relative instability of "cluster-B"; furthermore, this method has poor performance overall. "Voxel" performed nearly as well as TFCE for the more focal signal (test signal 5), but considerably worse than TFCE for the more extended signal (test signal 7). "cluster-A" performed similarly to TFCE for lower SNR, but considerably worse at higher SNR. Overall, TFCE performs the best of all methods tested.

Figure 7 shows the AUC values as boxplots over all 7 test signals and all 3 SNR levels (see caption for detailed explanation). These results suggest that TFCE provides better ROC AUC figures than all the other methods in virtually all cases. Furthermore, regarding the difference between the cluster-based AUC values and the TFCE values, the fact that the cluster-based 25th percentile is lower than the median and the 75th percentile is higher, illustrates the much greater variability (i.e., instability of results across different noise images) in cluster-based thresholding than in TFCE.

5 Real data example

We now give an illustration using real data. We used data published as part of a VBM-style analysis of early-onset schizophrenia [Douaud et al., 2007]. Structural MR (T1-weighted) images from 25 adolescent-onset schizophrenic patients were compared with images from 25 healthy age- and gender-matched adolescents. Tools from FSL [Smith et al., 2004] were used to pre-process the data according to the "optimised-VBM" approach [Good et al., 2001]. For each subject the Jacobian-modulated grey-matter segmentation image, registered into MNI152 standard-space, was smoothed according to the different methods as described below; the reduction in grey-matter in the schizophrenic group (compared with the controls) was tested for significance using permutation testing, controlling family-wise error-rate (i.e., fully correcting for multiple comparisons over space) for all thresholding methods.

For all methods tested we used the parameter settings reported as optimal in our previous evaluations: for TFCE, smoothing σ=1.5 voxels, H=2, E=0.5; for cluster-based thresholding, smoothing σ=1.5 voxels, cluster-forming-threshold=2.5; for voxel-based thresholding, σ=3 voxels. The working resolution of the data in standard space is 2x2x2mm. The results are shown in Figure 8.

If one accepts the group differences reported by the methods as "true", then it is clear that the voxel-based testing is extremely insensitive to the group difference; almost no significant voxels are found. (At lower, more commonly-used, smoothing levels, even these few significant voxels disappear.) TFCE finds major areas of difference in left and right Heschl's gyrus / parietal operculum and in the supplementary motor area (SMA), as well as a few smaller areas of difference. The cluster-based thresholding does not find the smaller areas, does not find SMA, and the areas that are found have lower statistical significance (p=0.04 corrected) than the equivalent peaks found by TFCE (both $p<0.01$ corrected).

Not only is TFCE more sensitive than cluster-based thresholding in this example, but it also retains information about relative significance of effect *within* the reported areas of group difference; it provides information regarding local maxima in the final significance map, which is not possible with cluster-based thresholding. In the coronal slice, it is clear that even though there is

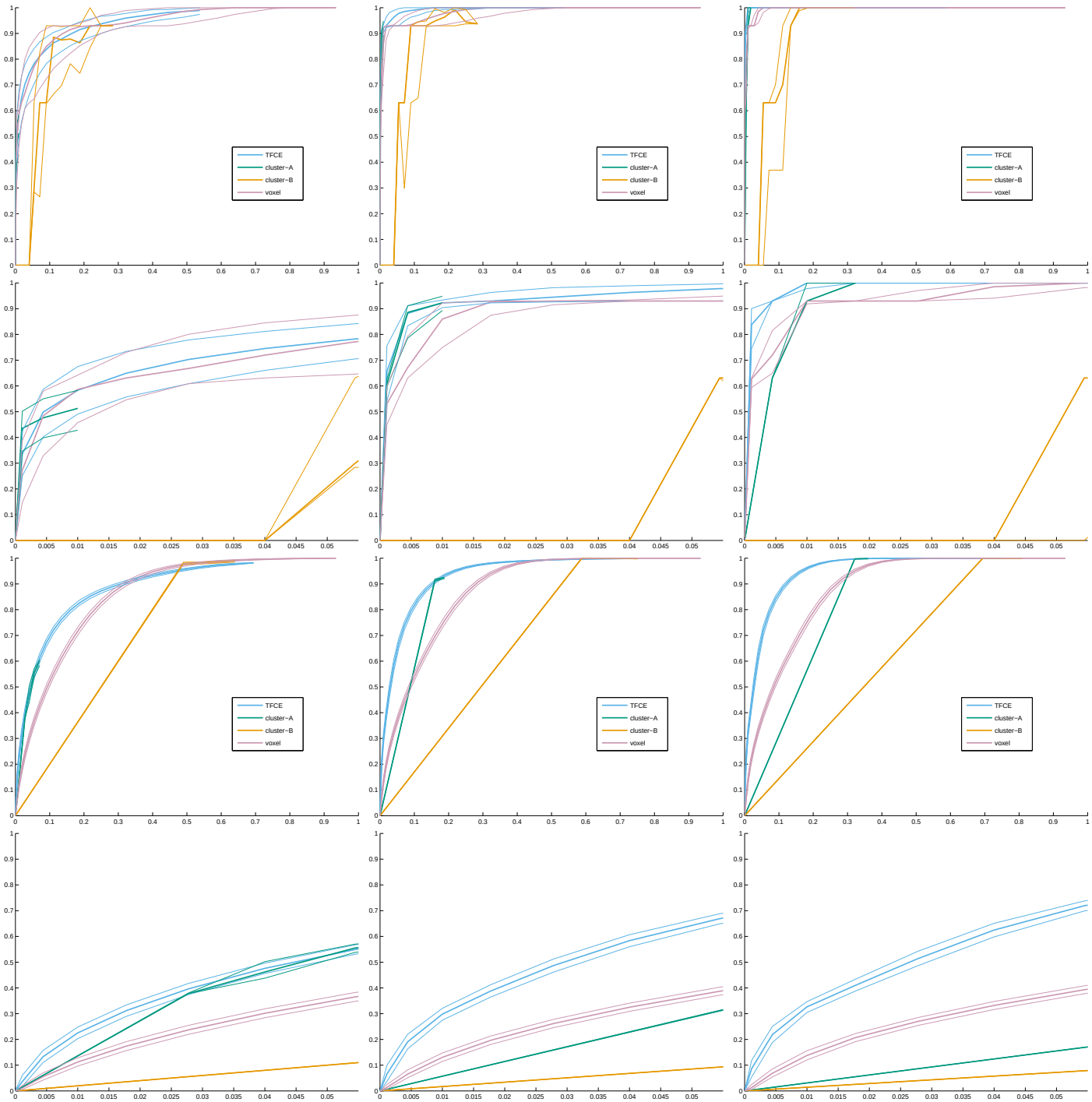

Figure 6: “Standard” ROC curves. Top half: test signal 5 (Penny’s blobs). Bottom half: test signal 7 (real FMRI data). The 3 columns are (left to right): SNR 1, 2 and 5. Rows 2 and 4 show expanded detail from rows 1 and 3, showing the FPR range 0:0.05. For each processing method, the median and IQR (over 1000 signal+noise images) ROC plots are shown.

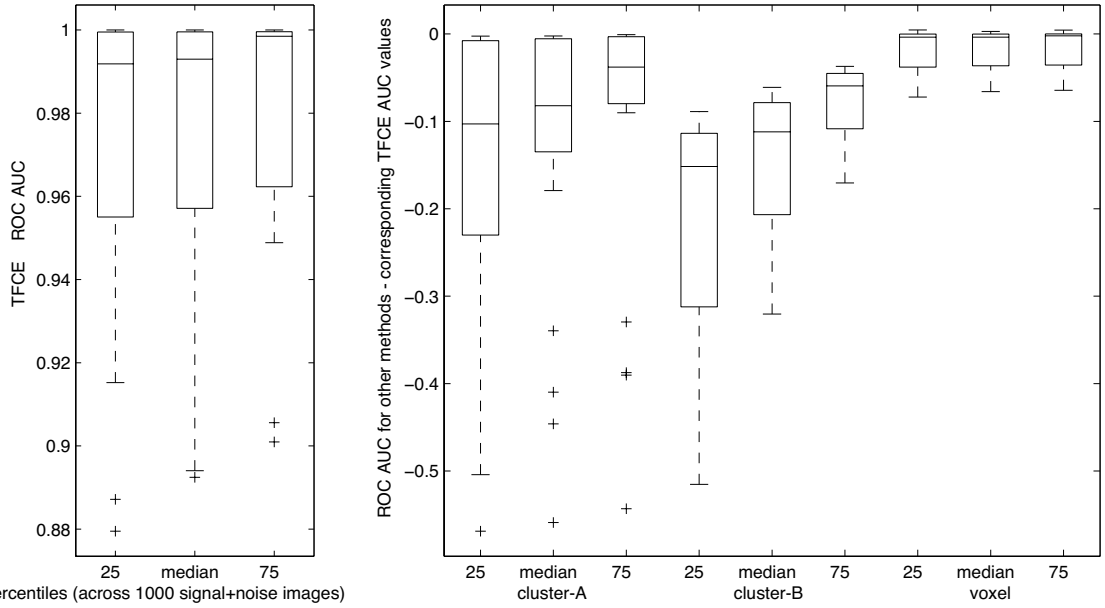

Figure 7: "Standard" ROC AUC plots, with boxplots over all 7 test signals and all 3 SNR levels. Left: boxplots of the TFCE ROC AUC values for 3 percentile values (25, 50, 75) over the 1000 signal+noise images. These percentiles correspond to the 3 curves shown in the full ROC plots displayed in Figure 6. Right: difference between the other methods' AUC values and the TFCE values, calculated separately for each test signal and each SNR level (to clarify method differences), and then displayed as a boxplot.

one large "cluster" on the left side of the brain that includes both Heschl's gyrus and parietal operculum, TFCE is showing strongly distinct peaks for the two areas.

Note that the SMA *was* found in the previously-published (cluster-based) analysis of this data, with slightly increased smoothing of σ=1.75 voxels, and with the lower cluster-forming threshold of t>1.7; one of the advantages of TFCE, as seen here, is that these areas of significant difference were found using the default TFCE parameter settings, whereas the SMA was only seen using cluster-based thresholding when the cluster-forming threshold was tuned for the study/data in question (a group difference was not expected to be very strong in this study of what is thought to be a neurodevelopmental disorder, so the cluster-based thresholding was adjusted for finding subtle, more diffuse regions of anatomical differences).

Finally, in Figure 9, we show the TFCE-enhanced t-statistic images and the thresholded significance maps for a range of E and H values. Where E is large, or greater than H, the enhanced maps are dominated by very extended, relatively uninformative areas of significant group difference, due to the subtle, more diffuse increase in grey-matter "density" in controls. At the other extreme, where E is small, or much less than H, sensitivity to any effect is very low, presumably as the spatial extent of the signal is being almost completely ignored. At the chosen values of E=0.5 H=2, we see the results as presented above, namely a significant group difference that is both sensitively and interpretably found.

6 Discussion

In this paper we have presented a new, simple, approach for defining a cluster-like voxel-wise statistic in a way that we feel is more natural and stable than the commonly-used approach of an initial cluster-forming hard thresholding. TFCE enhances cluster-like features in a statistical image without having to define clusters as binary units. Through the use of permutation testing it is straightforward to control the FPR of the TFCE output image, including controlling for multiple comparisons across voxels. We have found that TFCE gives greater sensitivity in general, than commonly-used methods, to finding true signal in the presence of noise, whether the signal is low and diffuse or focal and strong.

Because TFCE generates a voxel-wise output image, it maintains information about spatial detail *within* extended areas of signal. For example, local maxima in the TFCE output image can easily be identified, and separated from each other if a "cluster" contains more than one maximum. Furthermore, the local maxima *locations* will be identical to those in the original statistical image. This means that TFCE provices rich and interpretable output, retaining much more spatial information than traditional cluster-based approaches.

One might ask, when considering interpretability and spatial specificity of TFCE-based inference: "So. We have a voxel v in the

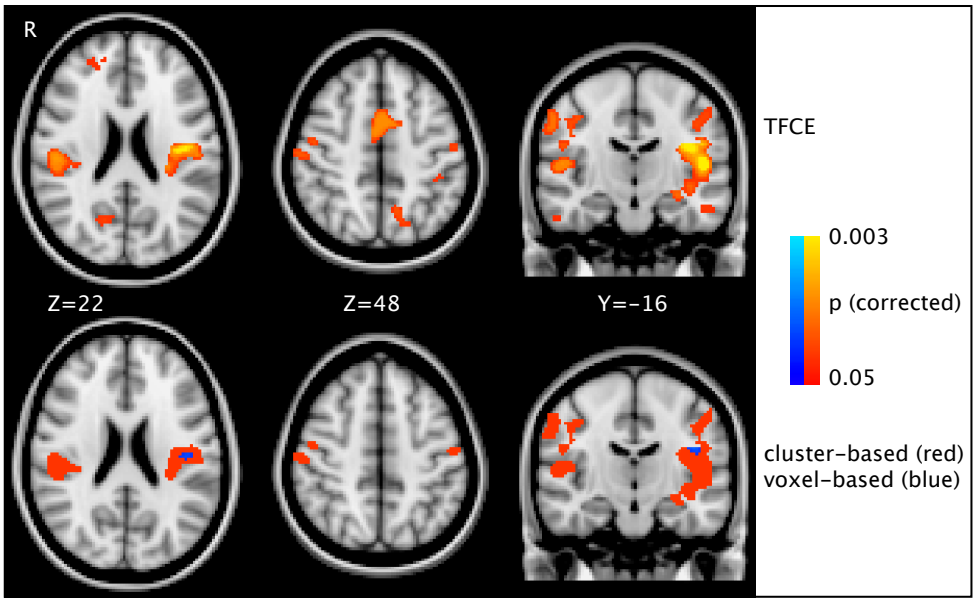

Figure 8: VBM-style analysis showing reduction of grey-matter in adolescent-onset schizophrenic patients, compared with controls. 3 different thresholding methods were compared: TFCE, cluster-based and voxel-based. Thresholding for all methods was at $p<0.05$, corrected for multiple comparisons across space using permutation testing.

TFCE output image that is statistically significant–which regions of the original data caused the result at v to reach significance?" (v is likely to be a local TFCE maximum, but there is no need to limit it to be such). A clear answer to this can be obtained by defining the smallest area of the local neighbourhood (i.e., finding the largest h_0) whose contribution to the TFCE score at v just about results in the TFCE score at v being statistically significant. This region can then be interpreted as the region in the original data which contributed to the result at the voxel of interest.

Our proposed approach seeks to avoid the need for any hand-tuning of the extent of data pre-smoothing, or the TFCE parameters. We believe that our testing shows that we have achieved this with the settings: smoothing of σ=1.5 voxels, H=2, E=0.5. However, a limitation of these fixed values is the assumption in our testing of no spatial correlation in the noise. This is an area for future development; it may be that it is important to optimise the smoothing extent taking into account the intrinsic data smoothness already present. However, a further factor, likely to be even more important, relates to a separate reason for data smoothing, and not covered at all in our development of TFCE to date: one reason for smoothing data in multi-subject studies (whether functional or structural) is to ameliorate the effects of misregistration across subjects. Unless an effect of interest is well-aligned across subjects, voxelwise analysis cannot hope to find it. Under the assumption of imperfect registration, data smoothing can, to some extent, help generate spatial overlap of effect across subjects. Where this is the case in a given dataset, a single fixed smoothing extent (e.g., of σ=1.5 voxels) is unlikely to automatically be the optimum, and further thought is needed.

We have not so far considered attempting to fuse the TFCE measure with concepts of "spatial-smoothness scale-space". One might consider integrating the TFCE score not just over the cluster-forming height/extent in the way proposed, but also integrating over a second-dimension - data smoothing extent. Although we have tried to find an optimal data smoothing extent, for the reasons discussed above, there may still be value in searching/integrating over a range of smoothings, as suggested in scale-space and wavelet literature going back more than a decade [Worsley et al., 1996a, Lindeberg et al., 1999, Coulon et al., 2000, Fadili and Bullmore, 2004, Flandin and Penny, 2007, Van De Ville et al., 2007]. Further investigations in these directions will be the subject of future work.

As well as work that attempts to apply smoothing over a range of scales, there has also been much work using adaptive smoothing of signal and/or noise in the data, for example [Woolrich et al., 2004, Harrison et al., 2007] (the latter being based on generic adaptive/nonlinear data smoothing going back as far as [Perona and Malik, 1990, Smith and Brady, 1997]). Such applications in neuroimaging generally have an explicit generative model for signal and/or noise, which can be inverted to optimise the local smoothing. The ambition of these techniques is similar to our approach but consider transformations of the data as opposed to the statistical image. Two advantages of our approach are its relative simplicity (allowing for the use of existing pre-processing and modelling), and its relatively flexible spatial model (allowing for a greater set of signal shapes, for example, than that modelled in [Hartvig and Jensen, 2000]).

With respect to the values of H and E; we believe that we have found (and derived theoretically in the Appendices) settings that work well over a wide range of signal shapes and SNR values. It is of course the case that for any given dataset, tweaking of

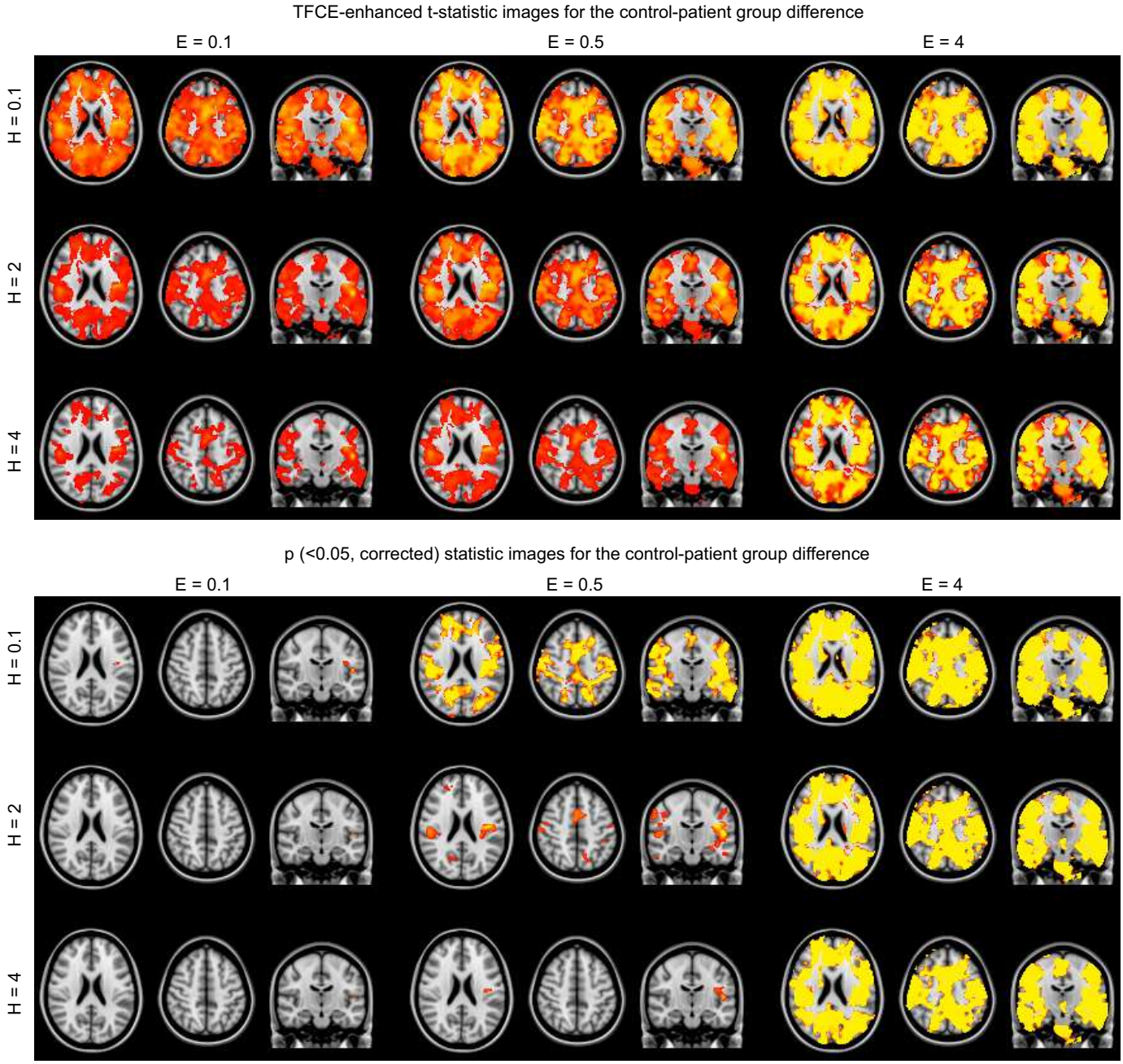

Figure 9: Control-patient group difference using TFCE with a range of settings for E and H. Top: TFCE-enhanced t-statistic images (colour overlay shows where TFCE-enhanced maps are greater than zero, and all are rescaled to have their image-wide maximum the same). Bottom: thresholded significance maps (same colour scale as used in Figure 8).

the parameters may generate "bigger blobs" and "smaller p-values" than the default recommended settings, but it is our hope that in general this will not be necessary, or even attempted; we would hope that there is much less to be gained from such tweaking than, for example, in the optimising of the cluster-forming threshold in traditional cluster-based thresholding. Such tweaking can be statistically dangerous, as it is hard to persuade the experimenter to honestly correct for "multiple comparisons" across different thresholdings!

TFCE has been implemented as an option in the "randomise" permutation-based inference tool in FSL 4.0 (including full permutation-based correction for multiple comparisons across space), and is therefore already publically available for general use.

7 Acknowledgments

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, to Dimitri van de Ville for his help with WSPM, to Gwenaëlle Douaud and Tony James for providing the schizophrenia data, and to Mark Woolrich, Adrian Groves, Karla Miller and Gwenaëlle Douaud for useful discussions.

A TFCE shape characteristics illustrated

In this appendix we further characterize the TFCE transformation. A strength of the TFCE statistic image that it is a monotonically-increasing function, within the neighbourhood of a local maximum, of the original statistic image. Of interest, however, is the form of this transformation, and how it may alter or distort signals. In Figure 10 we illustrate the TFCE output when the input image is a 3D Gaussian of half-width 10 voxels and maximum image intensity 5.

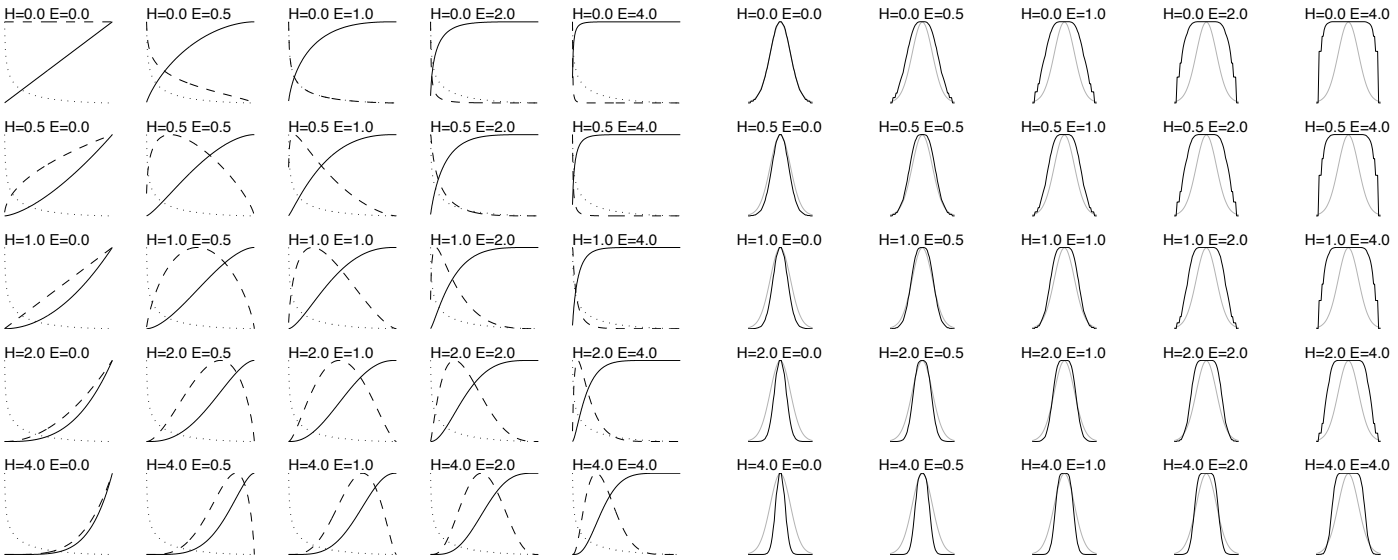

Figure 10: TFCE characteristics when the input image is a pure 3D Gaussian. Left: illustration of the "internal workings" of TFCE, when fed the same 3D Gaussian input signal, for various values of H and E. The dotted line shows $e(h)$, i.e., the supporting-section extent (volume) as a function of height; this is independent of E and H. The dashed line shows the resulting TFCE score for each supporting section, as a function of h. The solid line shows the final TFCE output value, as a function of the input intensity (i.e., is the integration of the dashed line). In all cases the x-axis is in units of h/h_{max} and the y-axis is normalised so that all plots have the same maxima. Right: dark-grey plots show the 1D cross-sectional profile through the TFCE output image for different values of H and E (light-grey shows the input profile, which is the same in all cases).

Figure 10 (left) shows this TFCE output (solid line) as a function of the original statistic value h_p. The dashed line shows the derivative of this line, i.e., indicates the relative contribution of "supporting sections" to the TFCE score, as a function of h. The figure shows that if E is selected too large, then there is too much sensitivity to small values, and that changes near the peak will have no impact. Choosing H too large gives weight only to the highest value, and will produce a statistic that behaves like a voxel-wise statistic. A moderate choice of H and E gives a statistic that is sensitive to all levels of the signal; we are drawn, for example, to H=2.0, E=0.5, as this gives some weight to moderate values of h, with somewhat greater emphasis to the largest values.

In Figure 10 (right) we show the TFCE outputs as 1D intensity profiles through the centre of the image. The different dark-grey curves show the TFCE output when using different values for H and E; the light-grey profile is the same in all cases and shows the input Gaussian profile. The output is re-normalised to the same peak intensity as the input, to allow comparisons of TFCE output shape only. It is clear that when $E \ll H$, TFCE is only sensitive to strong input peaks, and largely ignores spatial extent, hence the desired behaviour of sensitivity to spatially-extended signal is not achieved. At the other extreme, when $E \gg H$, the response to spatially-extended signal is almost independent of its height, not achieving the goal of greater sensitivity to higher-intensity signal.

B Motivation of TFCE statistic I: Weighted cluster norm

In this and the next section we present two different motivations for the TFCE measure and justification for particular choices of parameters H and E. Throughout we use the following notation: Let T_i ($i = 1, ..., V$) be the original statistic image with V voxels, and let $e_i(h)$ be the extent of the cluster containing voxel i, where clusters are defined with cluster-forming threshold h; if $T_i < h$ then $e_i(h)$ is zero. For a given statistic image and choice of dh, there are K cluster-forming thresholds considered $\{h_k\} = \{h_0, h + dh, h + 2dh, ..., h_{max}\}$, and so each voxel has a K-vector associated with it $\{e_i(h_k)\}_k$.

In this section we consider a general approach to summarizing the K-vector of extent information at each voxel with a p-norm, while in the next section we show how Fisher's P-value combining method and Random Field Theory can arrive at a similar final result.

A p-norm is a natural summary of vector magnitude. If $e_i(h)$ were considered for continuous h, the p-norm would be

$$[\int_{h=h_0}^{\infty} [e(h)]^p dh]^{1/p}$$

Discretizing, and generalizing to include a weighting function, we have

$$\left[\sum_k w(h_k)[e(h_k)]^p dh\right]^{1/p}.$$

Dropping the outer-most power, setting $p = E$ and choosing $w(h) = h^H$ produces the TFCE statistic.

B.1 Choice of H parameter based on voxel-wise h P-value

While both H and E are free parameters, one can use the distribution of the input statistic image to motivate a choice for H independent of E. In particular, h will typically have the interpretation as a z-score threshold (or closely related, as in a Student's t statistic). While $H = 1$ may seem like a natural chioce, z-scores are a non-linear measure of evidence against the null. For example, a 1 unit change in z-score from 1 to 2 corresponds to a 7-fold change in P-value, while a 1 unit change from 2 to 3 corresponds to a 17-fold change in P-value.

We are instead drawn to $-\log$ P-values, where a 1 unit change consistently represents an order of magnitude change in the P-value, the null-hypothesis chance of observing data as or more extreme than that actually observed. We now derive an approximation for $-\log$ P-values to find a value of H.

Assuming that the original statistic image value T is Gaussian, the P-value for statistic value h is

$$P(T \geq h) = 1 - \Phi(h) = \int_h^{\infty} \phi(h) dt \tag{2}$$

where Φ is the cumulative distribution function and ϕ is the probability distribution function of the standard Gaussian. Mill's ratio, $(1 - \Phi(h))/\phi(h)$, has a tight upper bound of $1/h$ for $h > 0$ [Gordon, 1941] and can be used to give the the following approximation

$$\begin{aligned} -\log P(T \geq h) &= -\log(1 - \Phi(h)) && (3)\\ &\approx -\log \phi(h)/h && (4)\\ &= \frac{1}{2}\log(2\pi) + h^2/2 + \log(h) && (5) \end{aligned}$$

Neglecting constants and lower order terms, this suggests that $-\log P$-values are approximately proportional to the square of the statistic value. This provides support for the view that both (a) $w(h) = h^H$ is a reasonable form for a weighted p-norm, and that (b) $H = 2$ is a reasonable choice for this parameter.

C Motivation of TFCE statistic II: Fisher's P-value combining statistic

Returning to the vector of cluster information for each voxel, $\{e_i(h_k)\}_k$, we consider another way to combine the elements of the vector.

C.1 Fisher's P-value combining method

Fisher's P-value combining meta-analysis method [Fisher, 1948] combines K independent tests. For P-values P_1, P_2, ..., P_K under the complete null hypothesis

$$-2\sum_{k=1}^{K}\log_e P_k \sim \chi^2_K. \tag{6}$$

While there are many meta analysis methods (see, e.g., [Lazar et al., 2002]), this test generally performs well when many of the K tests exhibit a signal [Pesarin, 2002] (as opposed to when only a single test exhibits an effect, in which case tests based on the minimum P-value are better).

While dependence violates the use of the χ^2 distributional result, it suggests that summing $-\log$ P-values is a sensible approach to combining inferences between tests. Specifically, we could expect statistics of the form

$$\sum_k -\log P_{e(h_k)}$$

(where $P_{e(h)}$ is the P-value of a cluster found with cluster forming threshold h), to perform well.

For the TFCE method, this suggests that H and E should be selected so that $h^H e(h)^E$ approximates $-\log$ P-values of clusters found with different thresholds.

In subsequent sections we use random field theory to derive an approximate $-\log$ P-value for clusters found with threshold h.

C.2 Random field theory

We now use Random Field Theory to obtain an approximate result for uncorrected P-values for a cluster of size $e(h)$, which produces a TFCE statistic with $H = 2$ and $E = 2/3$, close to our empirically determined $H = 2$ and $E = 0.5$. Random Field Theory uses results for continuous random processes to find P-values. These results were introduced to neuroimaging first by [Friston et al., 1994], though we mostly follow the notation of the summary given in Appendix A of [Hayasaka and Nichols, 2003]. We first find the mean cluster size under the null, the distribution about that mean, and finally $-\log$ P-values for an observed cluster size.

Let T be a voxel value in the statistic image (suppressing the i subscript) and e be the size of a randomly selected cluster; let N be the total suprathreshold volume (equivalently, the sum of all of the cluster sizes); let L be number of clusters observed. Assuming a large search region relative to the smoothness, and thus independence of number of clusters and cluster size, the mean cluster size under the null is

$$\mathsf{E}(e) = \frac{\mathsf{E}(N)}{\mathsf{E}(L)}. \tag{7}$$

The numerator is easily obtained

$$\mathsf{E}(N) = V\mathsf{P}(T > h) \tag{8}$$

and specifically for a Gaussian image $\mathsf{E}(N) = V(1-\Phi(h))$ where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of a Gaussian.

The expected number of clusters is $\mathsf{E}(L)$ approximated by the expected Euler Characteristic. The Euler Characteristic (EC) of a D-dimensional random field thresholded at h is written χ_h, and is the number of clusters minus the number of holes ($D \geq 2$) plus the number of handles ($D \geq 3$). For sufficiently high h the probability of a hole or handle is small, and so the EC offers a good approximation of the number of clusters.

Worsley et al's general results [Worsley et al., 1996b] give a closed-form expression for $\mathsf{E}(\chi_h)$ as a sum of D terms:

$$\mathsf{E}(\chi_h) = \sum_{d=0}^{D} R_d \rho_d(h), \tag{9}$$

where R_d is the RESEL count and ρ_d is the EC density. The RESEL count is a length, area or volume, depending on d, and it is the product of the spatial measure and a roughness measure $|\Lambda|^{1/2}$, where Λ is the $d \times d$ variance-covariance matrix of the partial derivatives of the data. Usually, only the $d = D$ term is appreciable, so for the 3D case we have $\mathsf{E}(\chi_h) \approx R_3(h^2 - 1)\exp(-h^2/2)(2\pi)^{-4/2}$. Thus the expected cluster size for a 3D Gaussian image with cluster-forming threshold h is

$$\begin{aligned}
\mathsf{E}(e) &= \frac{V(1-\Phi(h))}{V|\Lambda|^{1/2}(h^2-1)e^{-h^2/2}(2\pi)^{-2}} && (10)\\
&\approx \frac{\phi(h)/h}{|\Lambda|^{1/2}(h^2-1)e^{-h^2/2}(2\pi)^{-2}} && (11)\\
&= \frac{\phi(h)}{|\Lambda|^{1/2}h(h^2-1)\phi(h)(2\pi)^{-5/2}} && (12)\\
&\propto (h^3-h)^{-1} && (13)
\end{aligned}$$

Finally, using the result that cluster size to the $2/D$ power follows an exponential distribution [Nosko, 1969], we have

$$e^{2/D} \sim \text{Exp}\left(\left[\frac{\mathsf{E}(e)}{\Gamma(D/2+1)}\right]^{-2/D}\right) \quad (14)$$

where $\text{Exp}(\lambda)$ is the exponential distribution with mean $1/\lambda$, and $\Gamma(\cdot)$ is the gamma function.

Fortunately, an exponential distribution has a very simple form for P-values and $-\log P$-values: The CDF of an exponential is $1 - \exp(-t\lambda)$, the P-value function is thus $\exp(-t\lambda)$ and $-\log P = t\lambda$.

So, $-\log P_s$, the $-\log$ P-value for an observed 3-D cluster size s, is

$$\begin{aligned}
-\log P_s &= -\log \mathsf{P}(S > s) && (15)\\
&= -\log \mathsf{P}(S^{2/3} > s^{2/3}) && (16)\\
&= s^{2/3}\left[\frac{\mathsf{E}(S)}{\Gamma(3/2+1)}\right]^{-2/3} && (17)\\
&\propto \left[\frac{s}{(h^3-h)^{-1}}\right]^{2/3} && (18)\\
&\approx s^{2/3}h^2 && (19)
\end{aligned}$$

Thus we have just found that a Fisher's combining P-value method, combining cluster sizes for different h values, is approximated by TFCE with $H = 2$ and $E = 2/3$.

References

[Bullmore et al., 1999] Bullmore, E., Suckling, J., Overmeyer, S., Rabe-Hesketh, S., Taylor, E., and Brammer, M. (1999). Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural mr images of the brain. *IEEE Transactions on Medical Imaging*, 18(1):32–42.

[Bunch et al., 1978] Bunch, P., Hamilton, J., Sanderson, G., and Simmons, A. (1978). A free response approach to the measurement and characterization of radiographic-observer performance. *Journal of Applied Photographic Engineering*, 4:166–172.

[Chakraborty and Winter, 1990] Chakraborty, D. and Winter, L. (1990). A free response approach to the measurement and characterization of radiographic-observer performance. *Radiology*, 174:873–881.

[Coulon et al., 2000] Coulon, O., Mangin, J.-F., Poline, J.-B., Zilbovicius, M., Roumenov, D., Samson, Y., Frouin, V., and Bloch, I. (2000). Structural group analysis of functional activation maps. *NeuroImage*, 11:767–782.

[Douaud et al., 2007] Douaud, G., Smith, S., Jenkinson, M., Behrens, T., Johansen-Berg, H., Vickers, J., James, S., Voets, N., Watkins, K., Matthews, P., and James, A. (2007). Anatomically-related grey and white matter abnormalities in adolescent-onset schizophrenia. *Brain*, 130:2375–2386.

[Fadili and Bullmore, 2004] Fadili, M. and Bullmore, E. (2004). A comparative evaluation of wavelet-based methods for hypothesis testing of brain activation maps. *NeuroImage*, 3:1112–1128.

[Fisher, 1948] Fisher, R. (1948). Combining independent tests of significance. *American Statistician*, 2:30.

[Flandin and Penny, 2007] Flandin, G. and Penny, W. (2007). Bayesian fMRI data analysis with sparse spatial basis function priors. *NeuroImage*, 34:1108–1125.

[Friston et al., 1996] Friston, K., Holmes, A., Poline, J.-B., Price, C., and Frith, C. (1996). Detecting activations in PET and fMRI: Levels of inference and power. *NeuroImage*, 40:223–235.

[Friston et al., 1994] Friston, K., Worsley, K., Frackowiak, R., Mazziotta, J., and Evans, A. (1994). Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*, 1:210–220.

[Good et al., 2001] Good, C., Johnsrude, I., Ashburner, J., Henson, R., Friston, K., and Frackowiak, R. (2001). A voxel-based morphometric study of ageing in 465 normal adult human brains. *NeuroImage*, 14(1):21–36.

[Gordon, 1941] Gordon, R. (1941). Values of mill's ratio of area to bounding ordinate of the normal probability integral for large values of the argument. *Annals of Mathematical Statistics*, 12:364–366.

[Harrison et al., 2007] Harrison, L., Penny, W., Ashburner, J., Trujillo-Barret, N., and Friston, K. (2007). Diffusion-based spatial priors for imaging. *NeuroImage*, 38:677–695.

[Hartvig and Jensen, 2000] Hartvig, N. and Jensen, J. (2000). Spatial mixture modelling of fMRI data. *Human Brain Mapping*, 11(4):233–248.

[Hayasaka and Nichols, 2003] Hayasaka, S. and Nichols, T. E. (2003). Validating cluster size inference: Random field and permutation methods. *NeuroImage*, 20(4):2343–2356.

[Lazar et al., 2002] Lazar, N. A., Luna, B., Sweeney, J. A., and Eddy, W. F. (2002). Combining brains: A survey of methods for statistical pooling of information. *NeuroImage*, pages 538–550.

[Lindeberg et al., 1999] Lindeberg, T., Lidberg, P., and Roland, P. (1999). Analysis of brain activation patterns using a 3-D scale-space primal sketch. *Human Brain Mapping*, 7:166–194.

[Miller et al., 2006] Miller, K., Wiggins, G., and Wiggins, C. (2006). Isotropic, high-resolution FMRI at 7T using 3D stack-of-segmented EPI. In *Proc. Int. Soc. of Magnetic Resonance in Medicine*.

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

[Perona and Malik, 1990] Perona, P. and Malik, J. (1990). Scale-space and edge detection using anisotropic diffusion. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 12(7):629–639.

[Pesarin, 2002] Pesarin, F. (2002). *Multivariate Permutation Tests: With Applications in Biostatistics*. Wiley.

[Smith and Brady, 1997] Smith, S. and Brady, J. (1997). SUSAN - a new approach to low level image processing. *International Journal of Computer Vision*, 23(1):45–78.

[Smith et al., 2004] Smith, S., Jenkinson, M., Woolrich, M., Beckmann, C., Behrens, T., Johansen-Berg, H., Bannister, P., De Luca, M., Drobnjak, I., Flitney, D., Niazy, R., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J., and Matthews, P. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, 23(S1):208–219.

[Van De Ville et al., 2007] Van De Ville, D., Seghier, M., Lazeyras, F., Blu, T., and Unser, M. (2007). WSPM: Wavelet-based statistical parametric mapping. *NeuroImage*, 37:1205–1217.

[Woolrich et al., 2004] Woolrich, M., Jenkinson, M., Brady, J., and Smith, S. (2004). Fully Bayesian spatio-temporal modelling of FMRI data. *IEEE Trans. on Medical Imaging*, 23(2):213–231.

[Worsley et al., 1996a] Worsley, K., Marrett, S., Neelin, P., and Evans, A. (1996a). Searching scale space for activation in PET images. *Human Brain Mapping*, 4:74–90.

[Worsley et al., 1996b] Worsley, K. J., Marrett, S., Neelin, P., Vandal, A. C., Friston, K. J., and Evans, A. C. (1996b). A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping*, 4:58–73.