ACADEMIC PRESS

Available online at www.sciencedirect.com
SCIENCE DIRECT®

NeuroImage

NeuroImage 20 (2003) 2343–2356

www.elsevier.com/locate/ynimg

Validating cluster size inference: random field and permutation methods

Satoru Hayasaka and Thomas E. Nichols*

Department of Biostatistics, The University of Michigan, Ann Arbor, MI 48109, USA

Received 27 May 2003; revised 5 August 2003; accepted 5 August 2003

Abstract

Cluster size tests used in analyses of brain images can have more sensitivity compared to intensity based tests. The random field (RF) theory has been widely used in implementation of such tests, however the behavior of such tests is not well understood, especially when the RF assumptions are in doubt. In this paper, we carried out a simulation study of cluster size tests under varying smoothness, thresholds, and degrees of freedom, comparing RF performance to that of the permutation test, which is known to be exact. For Gaussian images, we find that the RF methods are generally conservative, especially for low smoothness and low threshold. For *t* images, the RF tests are found to be conservative at lower thresholds and do not perform well unless the threshold is high and images are sufficiently smooth. The permutation test performs well for any settings though the discreteness in cluster size must be accounted for. We make specific recommendations on when permutation tests are to be preferred to RF tests.

Introduction

A central interest to neuroscientists is the detection of changes in brain images obtained from PET (positron emission tomography) or MRI (magnetic resonance imaging). Cluster size inference is one of the approaches used in such investigation. A typical cluster size test consists of two steps. First, clusters are defined as sets of contiguous voxels whose intensity exceeds a preselected cluster defining threshold u_c, then the null hypothesis is tested by examining whether or not the spatial extent of these clusters is unusually large by chance alone.

The cluster size test is known to have increased sensitivity compared to tests based on voxel intensity when the signal is spatially extended (Friston et al., 1996; Poline et al., 1997). It is also known that, for signals with small extent, the test becomes more powerful with a high cluster defining threshold, and for signals with large extent, a low threshold increases the power of the test. Friston et al. (1996) suggest using cluster size inference when signals have wider extent than the image smoothness, which is often the case in fMRI studies, and using voxel intensity tests for low resolution images such as those in PET studies. However, the cluster size test has not been validated under various conditions (smoothness, threshold, etc), in particular for *t* images. Furthermore, the sensitivity of this test outside of ideal conditions is not understood either. In this paper we seek to characterize under what condition cluster size tests perform well. In addition, when these tests do not perform well, we examine probable causes in detail.

The idea of cluster size inference was pioneered by Poline and Mazoyer (1993) and Roland et al. (1993); they generate the distribution of cluster sizes from simulated images having the same characteristics, such as spatial autocorrelation, as the observed data. This approach has been further studied in fMRI by Forman et al. (1995) and in PET by Ledberg et al. (1998). The most widely used methods, however, are the ones based on the random field (RF) theory (Friston et al., 1994; Cao and Worsley, 2001).

RF-based cluster size tests are derived from a distribution approximation of cluster sizes based upon various parametric distributions. Like any other parametric tests, several assumptions are required, such as smooth images, a sufficiently high threshold u_c, and the uniform smoothness of images (Worsley et al., 1992, 1996; Petersson et al., 1999).

* Corresponding author. Department of Biostatistics University of Michigan, 1420 Washington Height, Ann Arbor, MI 48109. Fax: +1-734-763-2215.

E-mail address: nichols@umich.edu (T.E. Nichols).

1053-8119/$ – see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2003.08.003

Despite such restrictions, there is only a vague guideline as to how smooth images should be (Petersson et al., 1999). Furthermore, though the choice of threshold is made by investigators according to signals of interest (Poline et al., 1997), there is virtually no consensus on how high the threshold u_c should be for the RF theory to work.

There have been some simulation based validations on Gaussian RF results under reasonable smoothness and threshold. Friston et al. (1994) validated their RF test and found that, for sufficient smoothness and a high threshold, the test performs well. Holmes (1994) carried out simulations with different thresholds and found the RF test to be conservative for low thresholds. However, this conservativeness was not observed in simulations by Poline et al. (1997). Rather, they found that the RF test is anticonservative for low thresholds and becomes conservative for high thresholds. In the same simulations, they also found that the RF test becomes less conservative if images are smoother. One common feature in these validations is that the RF test was validated under ideal conditions, where images are sufficiently smooth and thresholds are reasonably high. In real data analyses, however, investigators prefer to use as little smoothing as necessary to avoid focal signals being blurred and various thresholds u_c to identify signals of their interest. Furthermore, the Gaussian RF results are inappropriate for low degrees of freedom t images. Under such conditions, the behavior of the test has not been well-characterized, especially for t images.

An alternative to the RF test is the permutation test (Holmes et al., 1996; Nichols and Holmes, 2002; Bullmore et al., 1999). Unlike the RF test, it requires almost no assumptions. The sole assumption is null hypothesis exchangeability. Exchangeability holds if permuting the group labels does not alter the distribution of the test statistic. Given exchangeability, the test proceeds by shuffling the data, computing a statistic image, calculating cluster sizes, and recording the size of the largest cluster. In this manner the permutation test generates the null distribution from data itself, and no knowledge of the underlying distribution of image voxels is required. The test is exact for the family-wise error (FWE) rate, which means that the probability of one or more type I errors is the same as the significance level of the test. However, because of a large number of calculations required, the permutation test is more computationally intensive than the RF test. Furthermore, while the test is straightforward for simple designs, multicondition designs or correlated data complicate the test (Bullmore et al., 1996).

In this work, we compare these two approaches and determine which is to be preferred under various conditions. In particular, we simulate Gaussian random fields and t random fields with different degrees of freedom and smoothness, and compare the performance of an RF test relative to the permutation test. We do not use a real data set for validation because the uniform smoothness assumption cannot be verified and is often questionable (Hayasaka and Nichols, 2002). Under nonuniform smoothness, or nonstationarity, there are relatively smooth and rough regions within the image which will alter the distribution of cluster sizes locally, resulting in biased inference (Worsley et al., 1999). Cluster size inference on nonstationary images will be addressed in our future work.

One of the novelties in this study is the validation of these tests on t images, which is done with laborious t image simulations, where a number of independent Gaussian images are simulated to form a t statistic image (there is no algorithm to directly generate smooth t random fields). While some authors (Poline et al., 1997) use Fourier domain simulation to simulate periodic images (where the left edge is continuous with the right edge), we simulate images in the spatial domain to obtain the most realistic results. In addition, we estimate smoothness from the simulated data as done in real data analyses. This estimation process introduces an additional source of variation into the inference. Another notable aspect is that a permutation test is carried out for each realization and its performance is assessed as well.

This paper is structured as follow: Details regarding the tests, as well as simulations are explained in the Methods section. Results from the simulations are presented in the Results section. Finally interpretation of findings from the simulations and conclusions are presented in the Discussion section. Appendices are included which summarize the RF theory in a consistent notation and address important details of the `SPM2` and `fmristat` implementations as well as smoothness estimation. An appendix on permutation theory is also included.

Methods

Model

In a brain image analysis a linear model can be written as

$$Y(v) = X\beta(v) + \sigma(v)\epsilon(v), \tag{1}$$

where $v = (x, y, z) \in \mathbb{R}^3$ is an index for voxels, $Y(v) = \{Y_1(v), Y_2(v), \ldots, Y_n(v)\}'$ is a vector of observed image intensities at voxel v from n scans, X is a known $n \times p$ design matrix, $\beta(v)$ is a p-dimensional vector of unknown parameters, $\sigma(v)$ is an unknown standard deviation at voxel v, and $\epsilon(v) = \{\epsilon_1(v), \epsilon_2(v), \ldots, \epsilon_n(v)\}'$ is a vector of unknown random errors with unit variance. We denote images by omitting the voxel index v (e.g., ϵ_i denotes the error image from the ith scan).

Let $\hat{\beta}(v)$ be an unbiased estimate of $\beta(v)$; then the residuals are

$$e(v) = Y(v) - X\hat{\beta}(v)$$

and an estimate of the residual variance is

$$\hat{\sigma}^2(v) = \frac{1}{\nu} e(v)'e(v),$$

where ν is the error degrees of freedom. If $\epsilon_i(v)$'s are

independent among scans and identically normally distributed, then the statistic image T is defined as

$$T(v) = \frac{\mathbf{c}\hat{\beta}(v)}{\sqrt{\mathbf{c}(X'X)^{-1}\mathbf{c}'}\ \hat{\sigma}(v)}, \quad (2)$$

where $\mathbf{c}$ is a row vector contrast of interest and T is then used to define clusters. Each cluster is formed as a set of contiguous voxels with their T exceeding a fixed cluster defining threshold u_c and sharing at least one common edge. In a 3D data set, this cluster formation method is known as the 18 connectivity scheme. Two voxels are considered as connected when they share a face or an edge, but not a vertex. In a $3 \times 3 \times 3$ voxel cube, all the voxels except the eight corner voxels are considered as connected to the voxel at the center.

Cluster size inference

Let the size of a cluster be S. The true null distribution of S is unknown, but is approximated by various methods such as the RF theory and permutations. The uncorrected P value, or the P value of a single cluster size, is defined as the probability of observing a certain cluster size or larger, and can be calculated from the approximated distribution of S. An uncorrected cluster P value is only appropriate when a cluster can be uniquely defined a priori, independent of size. For example, the nearest cluster to location (x, y, z) is appropriate, but the largest cluster in the occipital pole is inappropriate because it is vague and incorporates size.

Typically multiple clusters could occur at a given threshold, all of which may be of interest, thus creating a multiple comparisons problem among the clusters. To correct for this problem, family-wise error (FWE) rate corrected inferences are used. The FWE is the chance of any type I errors, or false positives, over all the clusters collectively. Such correction yields P values known as corrected P values, P values adjusted for multiple comparisons over all clusters in the volume searched. When more clusters occur under the null hypothesis, either due to low smoothness or a low threshold, then more multiple comparisons need to be accounted for in the corrected P values, reducing the sensitivity.

The FWE correction is implemented by calculating P values based on the null distribution of the largest cluster size $S_{\max}$. The rationale behind using the null distribution of $S_{\max}$ is that the probability of observing $S_{\max}$ larger than s is the same as the probability of at least one or more clusters being greater than s, the event of a family-wise error. Detailed explanation of the FWE correction is found in Appendix A.

Once a cluster is rejected as significant, we can conclude that one or more of the voxels within the cluster is active, though we cannot assert which voxel.

RF test

There are several assumptions of the RF test, so we outline them here. The assumptions of the RF test include:

- Lattice approximation: Images are realizations of a smooth random field sampled at points on a regular lattice. The theory is based on continuous RF, yet our data is discretely sampled on a lattice.
- Smooth images: Images are smooth, that is, their smoothness in terms of FWHM is relatively large compared to the voxel size. This is to support the lattice approximation.
- Stationarity: The smoothness of images is uniform anywhere within the images, or stationary. This ensures that the null distribution of cluster size is homogeneous.
- High threshold: The cluster defining threshold u_c is sufficiently high. The RF theory is based on asymptotic results, that is, the cluster size distribution can be approximated when u_c is raised high (Nosko, 1969; Friston et al., 1994; Cao, 1999).

Two versions of the RF method are considered in this study. The one based on an assumption that S raised to a power is exponentially distributed (Friston et al., 1994), as implemented in the `SPM2` package,[1] and the other based on an assumption that the distribution of S is approximated by the product of a beta and χ^2 random variables (Cao and Worsley, 2001) as implemented in the `fmristat` package.[2] To correct for the FWE rate, the distribution of $S_{\max}$ was used to obtain critical cluster sizes. Details on these methods have been reported in a number of publications. We collect them all in a consistent notation in Appendix A. The distribution approximation of the `SPM` RF test is for Gaussian images, whereas that of the `fmristat` RF test is refined for t images. The `fmristat` package uses the same cluster size distribution approximation as the `SPM` package for Gaussian images, but some calculations are done differently (see Appendix B).

Permutation test

Since being proposed by Holmes et al. (1996), the permutation test for brain image analyses has been further studied (Bullmore et al., 1999; Nichols and Holmes, 2002) and implemented in the `SPM` package as the `SnPM` toolbox.[3] Unlike the RF test mentioned before, this test does not require any distributional assumption, and produces valid P values even when the distribution of the image voxel is unknown. One of the few assumptions and the rationale for this test is the exchangeability assumption; that is, under the null hypothesis, scan labels can be permuted without altering the joint distribution of cluster sizes. This approach is suitable for second level PET or fMRI analyses based on summary statistics (Holmes and Friston, 1999). Note that to apply the permutation test directly on BOLD fMRI time

[1] Wellcome Department of Imaging Neuroscience, University College London. http://www.fil.ion.ucl.ac.uk/spm.

[2] Keith J. Worsley. http://www.math.mcgill.ca/keith/fmristat.

[3] Andrew Holmes and Tom Nichols. http://www.fil.ion.ucl.ac.uk/spm/snpm.

series, the temporal autocorrelation must be accounted for (Bullmore et al., 1996, 2001). In our test, we focus on the distribution of $S_{\max}$ in order to correct the FWE rate. Stationarity is not assumed in the permutation test, though variable smoothness will result in non-uniform sensitivity.

The implementation of this test is explained in detail in Nichols and Holmes (2002). In this study, the permutation test is used as implemented in the SnPM toolbox.

Simulation

We carry out Gaussian and *t* image simulations to validate two RF tests, as implemented in the SPM2 package and in the fmristat package, and the permutation test as implemented in the SnPM toolbox. For each simulation, rejection rates are recorded and their 95% confidence intervals (CIs) are calculated by normal approximation of a binomial proportion

$$\hat{p} \pm 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n_p}},$$

where $\hat{p}$ is the observed rejection rate and n_p is the number of realizations. The width of the CIs depends on values of $\hat{p}$. For example, at $\hat{p} = 0.05$ and $n_p = 3{,}000$, the width of the CI is $\hat{p} \pm 0.008$. Because the significance level of all the tests is set to 0.05, on average, 95% of all the CIs should cover 0.05. We essentially examine many CIs, so as many as 5% of them may not cover 0.05 by chance alone. However, it is impossible to compute the expected number of CIs not covering 0.05 by chance alone, because the results are correlated due to the same white noise image used in each realization of the simulations.

If the simulated rejection rate is smaller than 0.05, then the test is conservative but is still considered valid. On the other hand, if the rejection rate is greater than 0.05, then the test is anticonservative, or liberal, and is no longer considered as valid.

Gaussian image simulation

We generate each smooth Gaussian image in three steps. First, a single $104 \times 104 \times 104$ white noise image is generated for each realization, which is then smoothed with a 3D Gaussian kernel with different full-width at half-maximum (FWHM) (1.5, 3, 6, and 12 voxels). Finally the outer 36 voxels from the smoothed images are truncated in order to avoid nonuniform smoothness at the edge, yielding a $32 \times 32 \times 32$ image. The resulting image is then thresholded with thresholds u_c's with upper tail Gaussian probabilities of 0.01, 0.001, and 0.0001.

Three thousand realizations of Gaussian images are generated. The two RF tests are applied to the simulated data at 0.05 significance level. The permutation test is not applied because there is only one Gaussian image generated in each realization, which yields nothing to permute. In the RF test, the known smoothing kernel width is used instead of estimating smoothness from a single image in each realization, because there are no residuals from which to estimate smoothness.

t image simulation

We generate each *t* image by calculating a *t*-statistic image (2) from a set of Gaussian images. In our simulation, for each realization, a set of 10, 20, or 30 $32 \times 32 \times 32$ Gaussian images are generated by the method described above, with smoothing kernel FWHM 0 (no smoothing), 1.5, 3, 6, and 12 voxels. Then a *t* image is calculated based on a model, either a one-sample *t* test or a two-sample *t* test with equal sample sizes. The degrees of freedom for the *t* image is 9, 19, or 29 for the one-sample test, or 8, 18, or 28 for the two-sample test corresponding to group sizes of 5 and 5, 10 and 10, and 15 and 15. Our use of a two-sample *t* statistic image was motivated by our collaborators' data of comparing controls and schizophrenics (Taylor et al., 2002), and the results should be similar to that of a one-sample test with the same degrees of freedom. The generated *t* image is thresholded at the quantiles of a *t*-random variable with appropriate degrees of freedom with the upper tail probabilities of 0.01, 0.001, and 0.0001, and clusters are defined.

The image smoothness is estimated from each realized data set. Details regarding smoothness estimation are found in Appendix D.

For each sample size, 3000 sets of Gaussian images are simulated to generate *t* images, and both SPM and fmristat RF tests and the permutation test with 100 permutations are applied at 0.05 significance level.

Quality of Gaussian images simulated

Gaussian images, both for the Gaussian simulation and the *t* simulation in this study, are generated by convolving a white noise image with a Gaussian smoothing kernel (Worsley et al., 1992; Worsley, 1996). However, with decreasing smoothness, the Gaussian kernel is more coarsely sampled and it is unclear whether the nature of the dependence is affected by this. To investigate the quality of Gaussian images simulated, we carry out two additional simulations. In the first simulation, images of size $96 \times 96 \times 96$ voxels (after truncation), smoothed with a kernel of FWHM nine voxels, are generated. Then they are down-sampled at every third voxel so that the resulting image should be $32 \times 32 \times 32$ with FWHM 3 voxels (down-sized simulation). The other simulation is done in the same manner as the down-sized simulation, except that images are not down-sampled. Thus the simulated image size and its smoothness are three times that of the first simulation (oversized simulation). For each simulation, 3,000 realizations of two-sample (5 and 5) *t* images are generated, and a comparison is made on the 95th percentiles of the peak intensity and the largest cluster size at 0.001 threshold, among the down-sized and over-sized simulations, as well as the conventional method.

Note that in this comparison, cluster sizes are measured

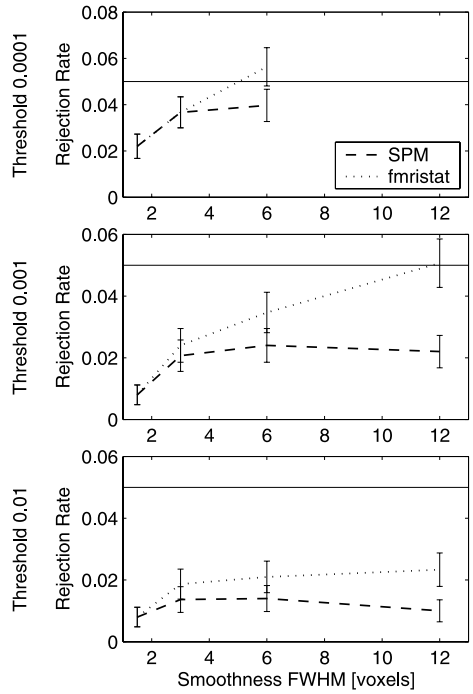

Fig. 1. Results from the Gaussian simulation. Rejection rates of the RF tests when thresholded at upper-tail probabilities 0.01, 0.001, and 0.0001 (from bottom to top), along with their 95% confidence intervals. Fine solid lines indicate the desired type I error rate (0.05) of the test.

in terms of RESELs (RESolution ELements), volume measured in units of smoothness:

$$\text{RESELs} = \frac{\#\text{ voxels}}{\text{FWHM}^3}.$$

We use RESELs, instead of voxels, because the search volumes in the three simulations in terms of RESELs are the same even though the search volumes in terms of voxels are different.

Robustness to smoothness outliers

The robustness of the permutation test against a violation in the exchangeability assumption is examined. In particular, in a two-sample *t* test setting, we investigate by simulations the effect of a single image with different smoothness (smoothness outlier), and also the effect of a systematic smoothness difference between two groups (smoothness difference).

For the smoothness outlier simulation, 19 images with the same smoothness (FWHM 0, 1.5, 6, or 12 voxels) and one image with a different smoothness (FWHM 12, 6, 1.5, or 0 voxels, respectively) are generated for each realization and a *t* image for a two-sample test is calculated. For the smoothness difference simulation, two groups of 10 images having different smoothness, FWHM 6 or 12 voxels for one group and FWHM 1.5 or 0 voxels for the other group, respectively, are generated for each realization and a *t* image for a two-sample test is calculated.

For both simulations, 3000 realizations are generated and the `SPM` and `fmristat` RF tests and the permutation test with 100 permutations are applied.

Computing environment

Each simulation is divided into segments of 200 to 1000 realizations to be run on several different computers separately, and the results are merged once all the segments are done. The random number generator was reset in each segment using the seed generated from a computer's internal clock. The fastest computer used in this study was a Dell PC with dual 2.4 GHz Xeon processors and 2 GB of RAM, on a Linux platform, with MATLAB version 6.5 (MathWorks Inc., Natick, MA). It took this computer 12 days to compute the *t* image simulation with df = 28, using both processors. Note that these simulation times are due to random number generation, smoothing, and repeated permutation tests. A single permutation test with 100 permutations on a given realization took 20–30 s on average.

Results

Gaussian image simulation

Results from the Gaussian simulation are shown in Fig. 1. The plots show that the RF tests are conservative in most settings. The tests are especially conservative when the threshold is low, u_c corresponding to $\alpha = 0.01$ or 0.001. It is also found that the tests are conservative for low smoothness, and for high smoothness, the `fmristat` RF test becomes less conservative, while the significance level does not change dramatically for the `SPM` RF test. Both tests are unable to calculate a critical cluster size as a real number at 0.0001 threshold with smoothness 12 voxels FWHM. As explained in Eq. (10) in Appendix A, the critical cluster size cannot be calculated for certain combinations of u_c and smoothness.

t image simulation

Because our one-sample and two-sample simulations produced similar results, we only present the results from the two-sample simulation.

RF test

With a widely used threshold of 0.01, the RF tests seem generally conservative, especially for low smoothness. Fig. 2 shows the rejection rates of the RF tests from *t* image simulation. The rejection rates do not approach to 0.05 unless the threshold is extremely high (0.0001) and images are smooth. In some cases, at a high threshold and low smoothness, the RF tests are extremely anticonservative. For low thresholds, rejection rates decrease with increasing df. While it is unusual for performance to worsen with

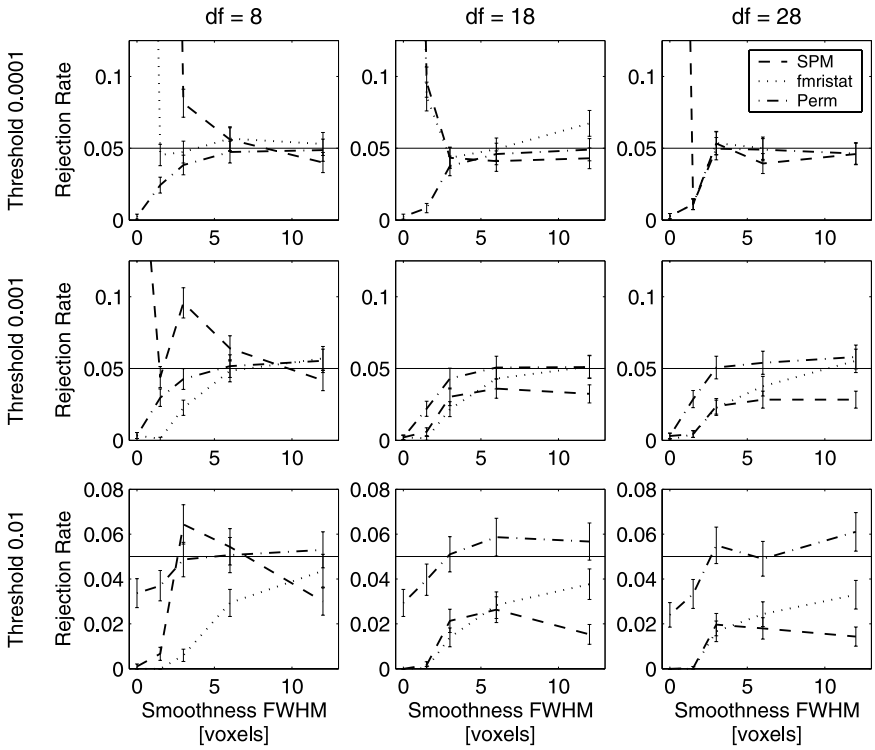

Fig. 2. Results from the *t* image simulation. Rejection rates of the RF tests and the permutation test for different sample sizes (5 and 5, 10 and 10, and 15 and 15, from left to right) when thresholded at upper-tail probabilities 0.01, 0.001, and 0.0001 (from bottom to top), along with their 95% confidence intervals. Fine solid lines indicate the desired type I error rate (0.05) of the test.

increasing df, the high df results do appear to converge to the Gaussian results (see Figs. 1 and 2, right column).

It is found that the `fmristat` RF test is more conservative in low smoothness and less conservative in high smoothness, compared to the `SPM` approach.

Permutation test

The permutation test in general works well for sufficiently smooth images at any threshold or any df (see Fig. 2). However, for low smoothness (FWHM < 3 voxels), the test is generally conservative, and it worsens as images become less smooth. This conservativeness is related to the discreteness in the $S_{\max}$ distribution, and in terms of P values, the method remains accurate. We return to this in detail in the Discussion section below.

Quality of Gaussian images simulated

Table 1 displays the results (95th percentiles) from the down-sized and oversized simulations, along with the results from the conventional simulation method with appropriate parameters. The percentiles are the FWE-controlling intensity and cluster size thresholds. It can be seen that the conventional simulation and the down-sized simulation produce very similar results, indicating that discretization of the Gaussian kernel had little impact, at least for 3 voxel FWHM smoothness. Thus we believe that our simulation of *t* images was appropriately done.

In contrast, the oversized simulation with equivalent RESEL volume has an appreciably larger intensity threshold and a 10% larger cluster size threshold. Such discrepancies indicate that $96 \times 96 \times 96$ images are a better approximation of a smooth random field, compared to $32 \times 32 \times 32$ images, and suggest that the lattice approximation is poor even for 3 voxel FWHM smoothness (for df $= 8$).

Robustness to smoothness outliers

Table 2 shows the rejection rates of the two RF tests and the permutation test when a smoothness outlier is present for a two-sample (10 and 10) *t* simulation thresholded at 0.01. While the results from the permutation test is some-

Table 1
Comparison of the 95th percentiles of the peak intensity and the largest cluster at 0.001 threshold from the conventional method, the down-sized simulation, and the oversized simulation for a two-sample t_8 image

Image size	95th percentile peak intensity	95th percentile cluster size [RESELs]
$32 \times 32 \times 32$ (conventional method)	11.4727	0.6138
$32 \times 32 \times 32$ (down-sized)	11.7513	0.6601
$96 \times 96 \times 96$ (oversized)	16.5730	0.7425

Note. All have the same RESEL volume but discrepancies between 32^3 and 96^3 volumes suggest that the lattice approximation does not hold for 3 voxel FWHM.

Table 2
Familywise rejection rates of the two RF tests (`SPM` and `fmristat`) and the permutation test when a smoothness outlier is present for a two-sample t_{18} simulation thresholded at 0.01 (the no outlier case, 3 and 3 is provided for reference)

Smoothness	0	1.5	3	6	12
Outlier smoothness	12	6	3	1.5	0
Smoothness estimate	1.21	1.69	3.04	4.66	4.77
Rejection rates					
`SPM`	0.000	0.005	0.025	0.120	0.330
`fmristat`	0.000	0.001	0.021	0.112	0.323
Permutation	0.030	0.039	0.051	0.052	0.047

Note. Smoothness estimates are also shown, which are highly underestimated for smooth images, which explain the anticonservativeness in the RF tests.

what close to the case of no smoothness outliers, the RF test results become highly anticonservative, especially for high smoothness images with a rough outlier, possibly due to underestimation of smoothness.

Table 3 shows the rejection rates when there is a systematic smoothness difference between two groups of 10 images in a two-sample test setting, thresholded at 0.01. Such smoothness difference influences the permutation test to be slightly anticonservative. However, compared to the RF tests which are highly anticonservative, the permutation is more robust when its null hypothesis exchangeability assumption is violated.

Discussion

We have simulated Gaussian images and *t* images and have applied three different cluster size tests to the simulated images, two RF tests, and the permutation test. Their performances at different thresholds, smoothness, and dfs are recorded, which enable us to assess the specificity and robustness of these tests.

Comparison with other Gaussian simulation results

There have been some simulation-based validations of the RF test on Gaussian images, comparable to our Gaussian image simulation. Friston et al. (1994) validated the RF test on 10,000 simulated images with size $32 \times 32 \times 64$ with FWHM 5.7 voxels thresholded at 2.8 (upper-tail probability 0.0026). In this simulation, the cluster size distribution from the RF test is very close to the simulated cluster size distribution. These results are different from ours, where the RF test is found to be conservative. Some possible explanations for these discrepancies include: their use of estimated FWHM, their larger search volume, and their possible nonuniform smoothness at edges of simulated images.

Holmes (1994) simulated 10,000 images of $65 \times 87 \times 26$ masked in the shape of a brain (72,410 voxels), smoothed with a nonisotropic Gaussian filter of FWHM $5 \times 5 \times 2.5$ voxels and thresholded at upper-tail probabilities $P = 0.01$, 0.001, and 0.0001. His filter is also nonstationary, in that he truncates and renormalizes the kernel when it contacts the mask. The RF test is done based on both known kernel FWHM and estimated FWHM. To be consistent with our results, we focus on the one with known FWHM. The results from Holmes' simulation are somewhat consistent with our results, except at threshold $P = 0.01$ where the results are less conservative compared to ours. Some possible explanations for this discrepancy include a brain-shaped search volume which reduces the chance of clusters touching the boundary (as a brain being more spherical than a box) and being truncated, and a larger search volume.

Poline et al. (1997) simulated 3000 Gaussian images of size $64 \times 64 \times 32$ with smoothing kernels FWHM $4.7 \times 4.7 \times 3.9$, $7.05 \times 7.05 \times 5.9$, and $9.4 \times 9.4 \times 7.85$ voxels, thresholded at 2.0, 2.5, 3.0, and 3.5 (upper-tail probabilities 0.023, 0.006, 0.0013, and 0.00023, respectively). Their results indicate that the higher the smoothness, the less conservative the test becomes, which is consistent with our results in Gaussian simulation. However, contrary to our simulation, for lower thresholds (2.0 and 2.5), they find that the test is actually anticonservative, and as the threshold is raised to 3.0, the test becomes conservative, approaching to the true significance level at threshold 3.5. A possible explanation for this discrepancy is the fact that they simulate images in a periodic manner, so that clusters are not truncated by the edge.

In general other authors have found that the RF test performs better at high thresholds in smooth images, which is consistent with our results.

t simulation results

RF theory

The RF cluster size tests rely on a number of approximations. In an RF cluster size test, the expected value (or the mean) of S is obtained from the expected values of the suprathreshold volume N and the number of clusters L, based on the relationship

$$\mathbf{E}[S] = \frac{\mathbf{E}[N]}{\mathbf{E}[L]}.$$

Table 3
Familywise rejection rates and smoothness estimates from the smoothness difference simulation (the no difference case, 3 and 3 is provided for reference)

Smoothness FWHM			
Group 1	3	6	12
Group 2	3	1.5	0
Smoothness estimate	3.04	2.24	1.65
Rejection rates			
`SPM`	0.025	0.571	0.612
`fmristat`	0.021	0.486	0.571
Permutation	0.051	0.077	0.066

Note. Two groups of 10 Gaussian images with different smoothness are used to generate a t_{18} image and cluster size tests are applied at the 0.01 threshold.

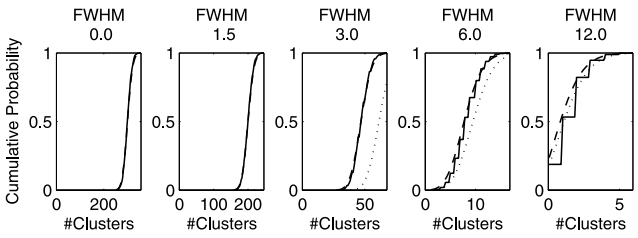

Fig. 3. The distribution of the number of clusters at 0.01 threshold for two-sample t_{18} images (solid lines). The Poisson distribution having the same mean as that of the observed distribution (dashed lines) approximates the observed distribution quite well. However, the Poisson distribution with the mean based on the RF theory (dotted lines, off the plots for 0 and 1.5 FWHM) does not approach the observed distribution unless images are very smooth.

Details on the derivation of the expected values above are presented in Appendix A. Because each voxel in a statistic image is a t or Z statistic, the distribution of N can be easily approximated. We examine the estimation of the other quantities, namely $\mathbf{E}[L]$ and $\mathbf{E}[S]$, to better understand the shortcomings of the RF tests.

The distribution of L is approximated by a Poisson distribution in the RF theory. As seen in Fig. 3, the observed distribution (solid lines) of L can be well-approximated by a Poisson distribution having the same mean (dashed lines). However, in practice, the mean of this distribution is unknown, thus estimated based on the RF theory which uses topological features of the suprathreshold volume (Worsley et al., 1996). When the RF theory estimate is used, the mean is grossly overestimated, and the resulting approximated distribution (dotted lines) deviates from the observed. The left panel in Fig. 4 shows the bias in the estimated $\mathbf{E}[L]$, which is substantial for low smoothness. A possible explanation for this overestimation is that the RF theory expects subvoxel clusters (i.e., clusters whose volume is less than a voxel) to occur which cannot be observed in a real statistic image. Such subvoxel clusters could occur more in low smoothness where lattice approximation is crude, resulting in a substantial overestimation seen in Fig. 4.

The distribution of S in an RF test is approximated either by Eq. (7) for `SPM` or Eq. (8) for `fmristat`. Fig. 5 shows the observed cluster size distribution and approximated cluster size distributions used in `SPM`. Each plot shows the cumulative probability, which can be interpreted as a plot of percentiles. The point where the cumulative probability is 0.95 is the 95th percentile, or the uncorrected 0.05 critical cluster size. The bottom row shows magnified cumulative probability plots around 0.95. Even when having the same mean as the observed distribution (solid lines), the theoretical distribution (dashed lines) does not approximate the observed distribution well unless images are very smooth. When the mean of the theoretical distribution is derived solely using the RF theory (dotted lines), this deviation from the observed worsens for high smoothness, yet corrects for conservativeness for low smoothness. For low smoothness, the observed distribution is discrete, with the majority of cluster sizes being 1 or 2 voxels, whereas the theoretical distribution is continuous. Therefore the RF test can only be either extremely conservative or anticonservative, depending on where the majority of such small clusters lie relative to the theoretical critical cluster size.

Because $\mathbf{E}[L]$ is overestimated, one might expect underestimation of $\mathbf{E}[S]$. However, such underestimation only occurs for low smoothness (see Fig. 4 right panel). For high smoothness, $\mathbf{E}[S]$ is actually overestimated possibly because the bias in $\mathbf{E}[L]$ is small and at the same time some parts of clusters are truncated by the boundary of the search volume, yielding smaller clusters than expected by the RF theory (see Fig. 6 for an illustration). As it can be seen in the plots of cluster truncation rates in Fig. 7, clusters are more likely to be truncated at low thresholds and in smooth images. Though the `SPM` method incorporates small volume correction using the unified approach (Worsley et al., 1996), the RF theory does not account for such cluster truncation, resulting in overestimation of $\mathbf{E}[S]$ and ultimately conservativeness of RF tests. It may seem that such cluster truncation is due to the small image size in our simulations; however, we have verified the performance of the RF tests on Gaussian, t_8, and t_{18} images of size $48 \times 48 \times 48$ voxels and found that the performance of the RF tests is similar to that of $32 \times 32 \times 32$ voxel images. Cluster truncations could occur at any image sizes, and neither the RF nor the permutation test is equipped to correct for such truncations. The permutation test is able to produce correct P values at low thresholds, but it is possible that the permutation test may have reduced sensitivity at the boundaries because of such cluster truncations. This problem of cluster truncation is a shortcoming for any stationary cluster size test.

In summary, the RF tests should not be used at low smoothness where the lattice approximation assumption fails and the cluster size distribution approximation is inaccurate. Even if images are sufficiently smooth, say FWHM ≥ 3 voxels, then clusters being truncated at the edge could lead to conservativeness, which is particularly of concern at low thresholds.

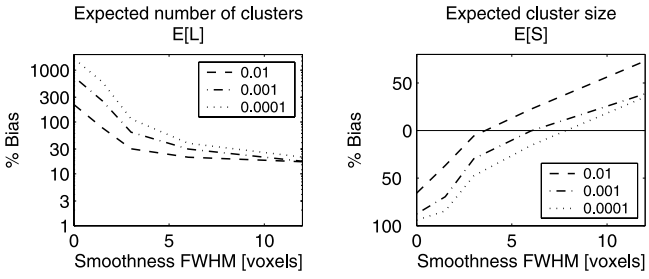

Fig. 4. Bias in estimating the expected number of clusters (left) and the expected cluster size (right) by the RF theory compared to the observed values for two-sample t_{18} images thresholded at different thresholds.

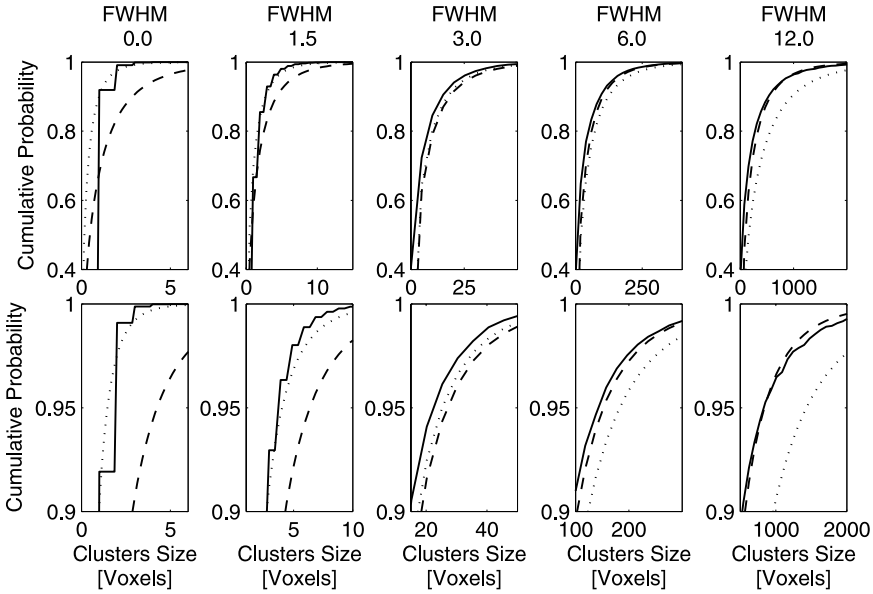

Fig. 5. The distribution of cluster sizes at 0.01 threshold for two-sample t_{18} images (solid lines). The shape of the distribution based on the theory (SPM) does not approximate the observed distribution well unless images are smooth, even when the theoretical distribution is set to have the same mean as that of the observed distribution (dashed lines). The RF theory (dotted lines) is biased relative to the theoretical distribution with the observed mean, but happens to be close to the observed distribution for low smoothness. The top row shows the overall shape of the distributions from the 40th to 100th percentiles, while the bottom row shows the shape of the distributions around the 95th percentiles, the uncorrected critical cluster sizes.

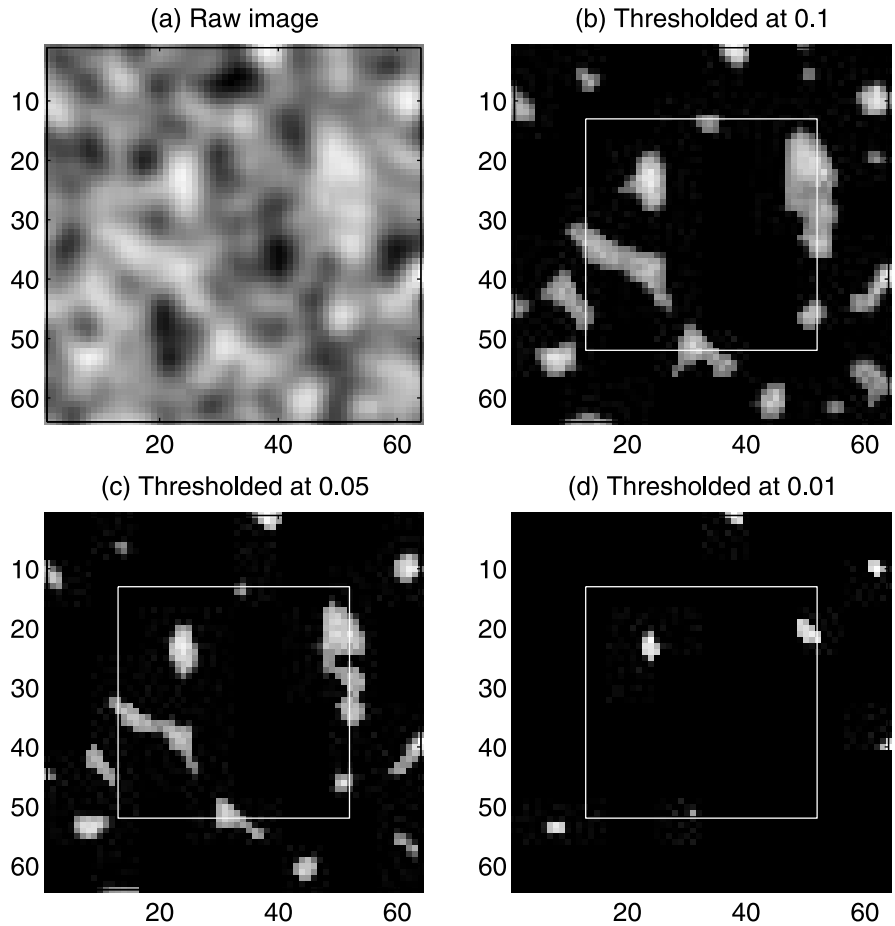

Fig. 6. A 2D illustration of clusters truncated by the edges. The raw image (a) is thresholded at different thresholds: 0.1 (b), 0.05 (c), and 0.01 (d). The edges (white box) truncate large parts of some clusters in (b) and (c), but not much in (d).

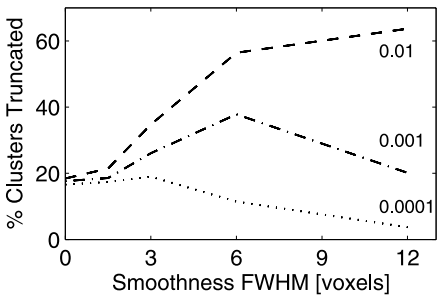

Fig. 7. The proportion the clusters touching the boundary and being truncated by the boundary for different thresholds on two-sample t_{18} images. Cluster truncation is most frequent for low thresholds and high smoothness.

Permutation test

For sufficient smoothness (FWHM $\geq$ 3 voxels), the permutation test seems to perform well for any thresholds and dfs. Fig. 8 shows the observed $S_{\max}$ distribution (solid lines), along with approximated $S_{\max}$ distributions from the three tests examined, for two-sample (10 and 10) *t* images at 0.01 threshold. From the figure, it can be seen that the distribution from a single permutation test (dash–dot lines) is close to the observed distribution for any smoothness despite a small number (100) of permutations, while the SPM RF test (dashed lines) and the `fmristat` RF test (dotted lines) are conservative for some smoothness.

The conservativeness of the permutation test rejection rates under low smoothness is due to the discreteness in the $S_{\max}$ distribution. Because of this discreteness, the 95th percentile in the observed distribution cannot be uniquely defined. Thus it is impossible to attain the rejection rate of 0.05, even when the approximated distribution is close to the observed distribution. Nevertheless, the permutation test produces accurate *P* values even under such circumstances; for example, with no smoothing, for a cluster of size 5 voxels in a *t* image with df = 28 at 0.01 threshold, *P* values are 0.045 for the truth, where as the average *P* values from 3,000 realizations are 0.048 for the permutation, and 0.756 for SPM.

In summary, the permutation test performs well in all settings considered. Even when discreteness of cluster size distribution is an issue for low smoothness, the test yields accurate *P* values.

Conclusions

In this study we carried out simulations to validate two cluster size inference methods, the RF test and the permutation test, in Gaussian images and *t* images. It was found that the RF tests do not perform well in some settings when theoretical approximations are not accurate. On the other hand, the permutation test works well for any threshold smoothness, and df, and showed great robustness when assumptions are violated. Thus, when possible, the permutation test should be used. If the permutation test cannot be used or the RF test is chosen, then the smoothness and the threshold should be chosen wisely.

As a practical guideline for users of cluster size tests, we only recommend using the RF test for very smooth images

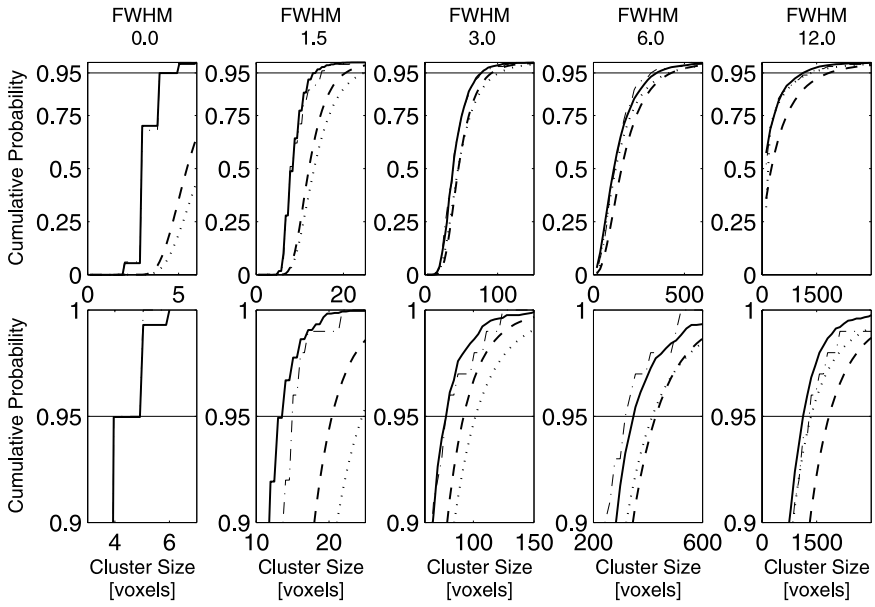

Fig. 8. The observed distributions of the maximum cluster size for the two-sample t_{18} images thresholded at 0.01 (solid lines), along with its approximations based on the SPM RF test (dashed lines), the `fmristat` RF test (dotted lines), and a single permutation test with 100 permutations (dash–dot lines). The top row shows the overall shape of the distributions from the 0th to 100th percentiles, while the bottom row shows the shape of the distributions around the 95th percentiles, the 0.05 FWE corrected critical cluster sizes (the permutation test overlaps the FWHM 0.0 case).

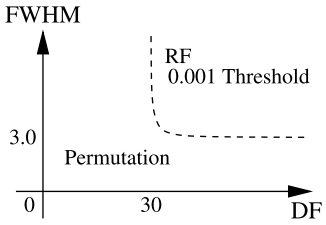

Fig. 9. Recommended usage of the RF and permutation cluster size tests. For high df and high smoothness, the RF test with 0.001 threshold is recommended. Otherwise the permutation test is more reliable.

with high df only. Though the permutation test is exact, as df increases, the computational burden associated of the permutation test increases as well, and the test may not be very practical. For large df, 0.001 threshold is typically used and at this threshold, though conservative, the RF test is stable for smooth images (FWHM > 3) as seen in the t_{28} and Gaussian simulations, thus perhaps more desirable. When df is small, 0.01 threshold is often used, and at this threshold, the RF test is overly conservative; thus, the permutation test is preferred. For low smoothness, often the RF test is unstable, or either extremely conservative or anticonservative, depending on the threshold, thus, we recommend using the permutation test for any df for images with low smoothness. Because the choice of threshold does not influence the permutation test much, 0.01 may be useful as it picks up more clusters. Fig. 9 summarizes the above recommendations.

In this study we did not simulate signals, so we are unable to make detailed comments on the power of these tests, except that a conservative test will generally be less sensitive than an exact test.

Acknowledgments

We thank Dr. Keith Worsley for useful comments and advice. We would also like to thank the fMRI Laboratory at the University of Michigan for use of their computing resources.

Appendix A: RF cluster size test

Under the RF theory, the distribution of the cluster size S is obtained based on the distribution of the suprathreshold volume N, the number of clusters in the search volume L, and the identity

$$\mathbf{E}[S] = \frac{\mathbf{E}[N]}{\mathbf{E}[L]}. \tag{3}$$

The expected value of N is

$$\mathbf{E}[N] = V(1 - F(u_c)), \tag{4}$$

where V is the search volume and $F(\cdot)$ is the cumulative distribution function (cdf) of an appropriate random variable. If the image is assumed to be a Gaussian random field, then $F(\cdot)$ is the cdf of a Gaussian random variable, and if the image is considered as a t-random field with degrees of freedom ν, then $F(\cdot)$ is the cdf of a t-random variable with v degrees of freedom. If the image is D dimensional (typically $D = 3$), then the expected value of L is

$$\mathbf{E}[L] = \sum_{d=0}^{D} R_d \rho_d(u_c), \tag{5}$$

where R_ds are d dimensional RESEL counts and ρ_ds are d-dimensional resel densities. R_ds and ρ_ds depend on the underlying random field (see Worsley et al. 1995, 1996).

It is known that, for a Gaussian RF with a large u_c,$S^{2/D}$ is approximately distributed as an exponential random variable (Nosko, 1969; Friston et al., 1994) with mean $1/\psi$, where

$$\psi = \left[\frac{\Gamma(D/2 + 1)\mathbf{E}[L]}{\mathbf{E}[N]}\right]^{2/D}. \tag{6}$$

The distribution of S can then be approximated by

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

Strictly, Eq. (7) is valid only for Gaussian images. A more refined approximation of the cluster size distribution in t RF was found by Cao (1999) and Cao and Worsley (2001); they find S is approximately distributed as

$$S \sim cB^{1/2}\left(\frac{U_0^D}{\Pi_{b=1}^{D} U_b}\right)^{1/2}, \tag{8}$$

where B is a Beta random variable with parameters $(1, (\nu - D)/2)$, U_0 is a χ^2 random variable with degrees of freedom $\nu + 1 - D$, and U_bs ($b = 1, 2, \ldots, D$) are independent χ^2 random variables with degrees of freedom $\nu + 2 - b$. c is a constant chosen so that Eq. (3) is satisfied.

Once the distribution of each cluster is found, either from Eq. (7) or (8), then the critical cluster size is to be found, adjusted for a desired family-wise error rate (FWER), or the probability of false rejections controlling for multiple comparisons among clusters. In this case, clusters are assumed to be independent (Adler, 1980; Friston et al., 1994), and the number of clusters whose size exceeds s, say Ls, can be approximated by a Poisson distribution with the mean $\mathbf{E}[L] \cdot Pr(S > s)$. Using this result, the FWER can be found as the probability of at least one cluster exceeding s, or 1 minus the probability that no cluster exceeding s, which is

$$Pr(L_s \geq 1) \simeq 1 - \exp(-\mathbf{E}[L] \cdot Pr(S > s)).$$

Note that the probability of at least one cluster exceeding s is equivalent to the probability that the largest cluster exceeding s. Thus, in the test, the largest cluster size $S_{\max}$ is

used as the test statistic, instead of all the cluster sizes, and its distribution is expressed as

$$Pr(S_{\max} > s) \simeq 1 - \exp(-\mathbf{E}[L] \cdot Pr(S > s)), \quad (9)$$

which is the FWE corrected P value of a cluster of size s. The critical cluster size is obtained as the cluster size at which Eq. (9) yields the desired significance level.

If the distribution of S is assumed to be Eq. (7), then the FWER adjusted critical cluster size k_α at a desired significance level α can be easily calculated by

$$k_\alpha \simeq \left[\frac{\ln\left(\frac{-\mathbf{E}[L]}{\ln(1-\alpha)} \right)}{\psi} \right]^{D/2}. \quad (10)$$

Note that if the ratio $(-\mathbf{E}[L])/(\ln(1-\alpha))$ is less than 1, then the natural log of the ratio becomes negative and k_α cannot be calculated. Such instances could occur when the expected number of clusters $\mathbf{E}[L]$ is too small due to an unrealistic combination of threshold u_c and the smoothness, or when α is extremely large.

Appendix B: *t* RF cluster size test implementations

In different software programs, theories described in Appendix A are implemented differently. In the `SPM2` package, the cluster size distribution for t images is assumed to be in the form of Eq. (7), whereas in the `fmristat` package, the distribution is assumed to be Eq. (8). Another difference between these packages is in the calculation of $\mathbf{E}[L]$. Though $\mathbf{E}[L]$ should be calculated with D different dimensional terms, in practice, if the search volume is large, the lower dimensional terms are often negligible. In the `SPM2` package, $\mathbf{E}[L]$ in Eq. (9) is calculated using all the dimensional terms, while the $\mathbf{E}[L]$ in the ψ parameter in Eq. (6) is calculated only with the highest dimensional term $R_d\rho_d(u_c)$, thus the ψ parameter becomes

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

In the `fmristat` package, $\mathbf{E}[L]$ is always calculated with the highest dimensional term $R_d\rho_d(u_c)$ only.

For Gaussian images, both `SPM2` and `fmristat` approximate the cluster size distribution with Eq. (7), though as mentioned above, they calculate $\mathbf{E}[L]$ differently.

Appendix C: Theory of the permutation test

In this appendix we outline the theory of the permutation test, using a two-sample test setting as an example. More detail and the case of one-sample and more complicated cases are considered in Pesarin (2001).

Consider a two-sample t test where there is one observation per subject. For groups of size n_1 and n_2, $n = n_1 + n_2$, the data can be represented in vector $Y = \{Y_1, \ldots, Y_{n1}, Y_{n1+1}, \ldots, Y_n\}'$. Let the distribution of group j be F_j, so that $Y_i \sim F_1$ for $i = 1, \ldots, n_1$, and $Y_i \sim F_2$ for $i = n_1 + 1, \ldots, n$. Under the null hypothesis of no group effect $F_1 = F_2 \equiv F$. Consider a univariate statistic of interest $T(Y)$ which summarizes evidence for a group difference. For example, T could measure the mean difference between the groups, or compute a two-sample t statistic.

Let S be an $n \times n$ permutation matrix, a matrix of zeros and ones such that SY shuffles the order of the elements of Y. Let $\mathcal{S} = \{S_k\}$ be the set of all permutation matrices corresponding to $N = \binom{n}{n_1}$ possible unique group assignments. Let $S_1 = I$, the identity corresponding to the unpermuted data.

Given the null hypothesis, the group labels are irrelevant and, given exchangeability, every possible permutation of the data S_kY, $k = 1, \ldots, N$, has the same distribution. Thus, $T(S_kY)$ has the same distribution for all possible k. Given particular observed data set y, the permutation distribution is defined by

$$\mathcal{P} = \{T(S_k y) : k = 1, \ldots, N\}. \quad (11)$$

Because of exchangeability, the distribution of $\mathcal{P}$ is uniform, with probability $1/N$ that $T(S_ky)$ takes on any particular value in $\mathcal{P}$. The P value is the probability of obtaining a statistic as large or larger than $T(y)$; conditioning on all possible permutations of the data, we have

$$Pr\{T(Y) \geq T(y) | \mathcal{H}_0, \mathcal{P}\} = \frac{1}{N} \sum_{k=1}^{N} \mathrm{I}_{\{T(S_ky) \geq T(y)\}}. \quad (12)$$

An α-level threshold can be found as the statistic value corresponding to the largest P value less than or equal to α.

While the permutation test is conditional on the data observed (and all $\mathcal{H}_0$-equivalent data sets), if the standard assumptions of random sampling from a population are made, the test is also unconditionally valid (see Pesarin (2001), pp. 61–63). Thus, for the data set at hand the validity of the permutation test relies on almost no assumptions (exchangeability only). However, if the investigator wishes to assume that their data are randomly drawn from a common population under the null hypothesis, as done parametrically, the test has the same validity and interpretation as any parametric test.

Appendix D: Smoothness estimation

In neuroimage analyses, there exist different ways to estimate image smoothness. Widely used methods are Kiebel et al. (1999) and Forman et al. (1995). Jenkinson (2000) explains both methods in detail. In our simulation, we use the approach by Kiebel et al., the one used in the `SPM2` package.

The smoothness of images are estimated in terms of FWHM from the variance–covariance matrix of partial derivatives of residual images. The variance–covariance matrix of spatial partial derivatives of a random field G is defined as

$$\Lambda = \mathrm{Var}\left(\frac{\partial G}{\partial(x, y, z)}\right) = \begin{pmatrix} \mathrm{Var}\left(\frac{\partial G}{\partial x}\right) & \mathrm{Cov}\left(\frac{\partial G}{\partial x}, \frac{\partial G}{\partial y}\right) & \mathrm{Cov}\left(\frac{\partial G}{\partial x}, \frac{\partial G}{\partial z}\right) \\ \mathrm{Cov}\left(\frac{\partial G}{\partial y}, \frac{\partial G}{\partial x}\right) & \mathrm{Var}\left(\frac{\partial G}{\partial y}\right) & \mathrm{Cov}\left(\frac{\partial G}{\partial y}, \frac{\partial G}{\partial z}\right) \\ \mathrm{Cov}\left(\frac{\partial G}{\partial z}, \frac{\partial G}{\partial x}\right) & \mathrm{Cov}\left(\frac{\partial G}{\partial z}, \frac{\partial G}{\partial y}\right) & \mathrm{Var}\left(\frac{\partial G}{\partial z}\right) \end{pmatrix} = \begin{pmatrix} \lambda_{xx} & \lambda_{xy} & \lambda_{xz} \\ \lambda_{yx} & \lambda_{yy} & \lambda_{yz} \\ \lambda_{zx} & \lambda_{zy} & \lambda_{zz} \end{pmatrix}. \quad (13)$$

In real data, the Λ matrix is estimated based on standardized residual images u, which is defined at each voxel v as

$$u(v) = \frac{e(v)}{\left(\frac{1}{\nu} e(v)' e(v)\right)^{1/2}} = \frac{e(v)}{\hat{\sigma}(v)},$$

where ν is the error degrees of freedom. Partial derivatives of u is calculated by taking the difference between $u(v)$ and adjacent voxels in x, y, and z directions and dividing it by the voxel dimension. Denote this by

$$\Delta u(v) = \left(\frac{\Delta u(v)}{\Delta x}, \frac{\Delta u(v)}{\Delta y}, \frac{\Delta u(v)}{\Delta z}\right).$$

Then an estimate of $|\Lambda|$, $|\hat{\Lambda}|$, is given by

$$|\hat{\Lambda}| = \frac{1}{V} \sum_{v} \left| \frac{1}{\nu} \Delta u(v)' \Delta u(v) \right|, \quad (14)$$

where V is the number of voxels. This expression can be seen as an averaging of determinants over space, and $|1/\nu\ \Delta u(v)' \Delta u(v)|$ as an averaging of matrices over observations. Note that Eq. (14) differs slightly from Kiebel et al., 1999 because we write it in terms of standardized residuals and not normalized residuals ($u(v)/\sqrt{\nu}$).

FWHM is expressed in terms of $|\Lambda|$ by

$$\mathrm{FWHM} = (4 \ln 2)^{1/2} |\Lambda|^{-1/2D} \quad (15)$$

(Worsley, 2002). Unfortunately the obvious estimate of FWHM, replacing $|\Lambda|^{-1/2D}$ with $|\hat{\Lambda}|^{-1/2D}$ results in a biased estimator. Worsley (2002) shows that $|\hat{\Lambda}|^{-1/2D}$, needs to be divided by a bias correction which is a function of the degrees of freedom. In our case, the estimate of FWHM is to be used in the RF test in the form $1/\mathrm{FWHM}^D$, instead of FWHM as it is (Worsley et al., 1996). It turns out the correction factor for $|\hat{\Lambda}|$ in $1/\mathrm{FWHM}^D$ is 1 (Worsley, 2002), so $\widehat{\mathrm{FWHM}}$ can be obtained from Eq. (15) with $\hat{\Lambda}$ substituted for Λ:

$$\widehat{\mathrm{FWHM}} = (4 \ln 2)^{1/2} |\hat{\Lambda}|^{-1/2D}.$$

For the calculation of Eq. (14), the `SPM2` package assumes the off-diagonal elements of $\hat{\Lambda}$ to be zero and calculates $\widehat{\mathrm{FWHM}}$ accordingly. In this simulation, we calculated the $\widehat{\mathrm{FWHM}}$ in that manner. However, if the off-diagonals are assumed to be zero, then the bias correction factor is no longer 1. The `multistat.m` function in the `fmristat` package calculates an appropriate bias correction factor according to the df and the dimensionality of the search space.

References

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

Bullmore, E., Brammer, M., Williams, S.C.R., Rabe-Hesketh, S., Janot, N., David, A., Mellers, J., Howard, R., Sham, P., 1996. Statistical methods of estimation and inference for functional MR image analysis. Magn. Reson. Med. 35, 261–277.

Bullmore, E.T., Suckling, J., Overmeyer, S., Rabe-Hesketh, S., Taylor, E., Brammer, M.J., 1999. Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. IEEE Trans. Med. Imaging 18, 32–42.

Bullmore, E., Long, C., Suckling, J., 2001. Colored noise and computational inference in neurophysiological (fMRI) time series analysis: resampling methods in time and wavelet domains. Hum. Brain Mapp. 12, 61–78.

Cao, J., 1999. The size of the connected components of excursion sets of χ^2, t, and F fields. Adv. Appl. Probability 31, 579–595.

Cao, J., Worsley, K.J., 2001. Applications of random fields in human brain mapping, in: Moore, M. (Ed.), Spatial Statistics: Methodological Aspects and Applications, Springer Lecture Notes in Statistics, Vol. 159. Springer, New York, pp. 169–182.

Forman, S.D., Cohen, J.D., Fitzgerald, J.D., Eddy, W.F., Mintun, M.A., Noll, D.C., 1995. Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. Magn. Reson. Med. 33, 636–647.

Friston, K.J., Worsley, K.J., Frackowiak, R.S.J., Mazziotta, J.C., Evans, A.C., 1994. Assessing the significance of focal activations using their spatial extent. Hum. Brain Mapp. 1, 210–220.

Friston, K.J., Holmes, A., Poline, J.-B., Price, C.J., Frith, C.D., 1996. Detecting activations in PET and fMRI: levels of inference and power. NeuroImage 4, 223–235.

Hayasaka, S., Nichols, T E., 2002. A resel-based cluster size permutation test for nonstationary images Presented at the 8th International Conference on Functional Mapping of the Human Brain, June 2–6, 2002, Sendai, Japan. Available on CD-ROM in NeuroImage 16 (2), 1062–1063.

Holmes, A.P., 1994. Statistical Issues in Functional Brain Mapping. Ph.D. thesis, University of Glasgow.

Holmes, A.P., Friston, K.J., 1999. Generalizability, random effects, and population inference. Proceedings of the 4th International Conference on Functional Mapping of the Human Brain, June 7–12, 1998, Montréal, Canada. NeuroImage 7, S754.

Holmes, A.P., Blair, R.C., Watson, J.D.G., Ford, I., 1996. Nonparametric analysis of statistic images from functional mapping experiments. J. Cereb. Blood Flow Metab. 16 (1), 7–22.

Jenkinson M., 2000. Estimation of Smoothness from the Residual Field. Technical Report, Oxford Centre for Functional Magnetic Resonance Imaging of the Brain (FMRIB).

Kiebel, S.J., Poline, J.-B., Friston, K.J., Holmes, A.P., Worsley, K.J., 1999. Robust smoothness estimation in statistical parametric maps using standardized residuals from the general linear model. NeuroImage 10, 756–766.

Ledberg, A., Åkerman, S., Roland, P.R., 1998. Estimation of the probability of 3D clusters in functional brain images. NeuroImage 8, 113–128.

Nichols, T.E., Holmes, A.P., 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum. Brain Mapp. 15, 1–25.

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

Pesarin, F., 2001. Multivariate Permutation Tests. Wiley, New York.

Petersson, K.M., Nichols, T.E., Poline, J.-B., Holmes, A.P., 1999. Statistical limitations in functional neuroimaging II. Signal detection and statistical inference. Philos. Trans. R. Soc. (Lond) Series B 354, 1261–1281.

Poline, J.-B., Mazoyer, B.M., 1993. Analysis of individual positron emission tomography activation maps by detection of high signal-to-noise-ratio pixel clusters. J. Cereb. Blood Flow Metab. 13, 425–437.

Poline, J.-B., Worsley, K.J., Evans, A.C., Friston, K.J., 1997. Combining spatial extent and peak intensity to test for activations in functional imaging. NeuroImage 5, 83–96.

Roland, P.E., Levin, B., Kawashima, R., Åkerman, S., 1993. Three-dimensional analysis of clustered voxels in 15-*O*-butanol brain activation images. Hum. Brain Mapp. 1, 3–19.

Taylor, S.F., Liberzon, I., Decker, L.R., Koeppe, R.A., 2002. A functional anatomic study of emotion in schizophrenia. Schizophrenia Res. 58, 159–172.

Worsley, K.J., 1996. The geometry of random images. CHANCE 9, 27–40.

Worsley, K.J., 2002. Non-stationary FWHM and its effect on statistical inference of fMRI data Presented at the 8th International Conference on Functional Mapping of the Human Brain, June 2-6, 2002, Sendai, Japan. Available on CD-ROM in NeuroImage 16 (2), 779–780.

Worsley, K.J., Evans, A.C., Marrett, S., Neelin, P., 1992. Three-dimensional statistical analysis for CBF activation studies in human brain. J. Cereb. Blood Flow Metab. 12, 900–918.

Worsley, K.J., Marrett, S., Neeline, P., Evans, A.C., 1995. A unified statistical approach for determining significant signals in location and scale space images of cerebral activation, in: Myers, R., Cunningham, V.J., Bailey, D.L., Jones, T. (Eds.), Quantification of Brain Function Using PET, Academic Press, San Diego, pp. 327–333.

Worsley, K.J., Marrett, S., Neelin, P., Vandal, A.C., Friston, K.J., Evans, A C., 1996. A unified statistical approach for determining significant signals in Images of cerebral activation. Hum. Brain Mapp. 4, 58–73.

Worsley, K.J., Andermann, M., Koulis, T., MacDonald, D., Evans, A.C., 1999. Detecting changes in nonisotropic images. Hum. Brain Mapp. 8, 98–101.