Chapter Four

Two Stage Testing

In this chapter we consider a two stage approach to assessing multiple subject activation studies. The experimental subjects are divided into two groups, a pilot group and a study group. A statistic image is formed from the pilot group data, from which interesting regions are identified. The study group data is then analysed over these interesting regions, using ROI methods.

A simple two-dimensional simulation is presented, which compares the two-stage approach with some of the voxel-by-voxel approaches described in ch.3.

4.1. Two-Stage Testing

Consider a multiple subject activation study, in which subjects are scanned repeatedly under two mental states which are to be compared. In the absence of any prior hypotheses regarding the brain location engaged differently by the two states under study, we have two avenues for analysis. The first involves a pilot study to identify interesting regions of the brain. These are then examined in a study on new subjects using ROI methods. This is the *two-stage* approach. Hypotheses are generated in stage one, the pilot stage, and assessed in stage two, the study stage. The data used in each stage are independent. The second option proceeds with an analysis that considers the whole intracerebral volume, usually voxel-by-voxel. Most research to date has taken the latter option, forming statistic images using the methods of ch.2, and assessing them with the methods of ch.3.

However, the large multiple comparisons problem of a voxel-by-voxel approach results in methods with low power. A voxel-by-voxel approach may be less powerful than a two-stage approach using the same number of subjects. The small number of comparisons in the study stage of a two-stage approach may make up for the smaller study group and the possibility of spurious region hypotheses being identified in the pilot stage, to produce a test more powerful than a voxel-by-voxel approach with the same number of subjects.

It is this possibility which we shall explore in this chapter.

Two-Stage testing from subject difference images

Consider a simple activation experiment with N subjects, each scanned M times under each of the two conditions, “rest” and “active”. Choose N_1 subjects as the pilot group, with the remaining N_2 (= N-N_1) subjects as the study group.

For simplicity, we shall consider the proportional scaling approach, with paired t-statistics formed from subject difference images as described in §2.3.1. As noted in §2.3.1.2., the paired t-statistic for multiple subject simple activation studies provides a simple statistic that absorbs undesirable effects that otherwise require an extremely complex model, but that the approach is hampered by the low degrees of freedom for variance estimation.

Let Y'_{gijqk} denote the rCBF (rA) measurement at voxel k =1,…,K, of scan j =1,…,M under condition q = 0,1 (0 = “rest”, 1=“active”), on subject i = 1,…,N_g in group g =1,2 (1=“pilot”, 2=“study”), after normalisation for global changes by proportional scaling (§2.1.2.) As usual, we shall refer to voxels by their index, k, to regions by the set of indices of the voxels in that region, and use H_U to mean the omnibus hypothesis over region U, the intersection of the voxel hypotheses for voxels with indices $k \in \mathrm{U}$.

The data for each subject is collapsed into a *subject difference image*, $\Delta_{gi} = (\Delta_{gi1},\ldots,\Delta_{giK})$.

$$\Delta_{gik} = \overline{Y'}_{gi\bullet 1k} - \overline{Y'}_{gi\bullet 0k} \tag{61}$$

We shall consider $\Delta_{gik} \sim N(\mu_k,\sigma_k^2)$. As we shall see, distributional assumptions are not necessary for the pilot group subject difference images, and weaker assumptions are adequate for the study group difference images. The hypothesis of no activation is H_k:μ_k= 0, with alternative hypothesis $\overline{\mathrm{H}}_k$:μ_k> 0. The omnibus hypothesis is H_W, where W = {1,…,K} is the set of (indices of) voxels under consideration.

Stage 1: Region selection from pilot group data

The pilot group *t*-statistic image, $\boldsymbol{T}_1 = (T_{11},\ldots,T_{1K})$, is constructed from the pilot group subject difference images as described in §2.3.1., giving, in the current notation:

$$T_{1k} = \frac{\overline{\Delta}_{1\bullet k}}{\sqrt{S_{1k}^2 / N}} \qquad (62)$$

where $\overline{\Delta}_{1\bullet k} = \frac{1}{N_1}\sum_{i=1}^{N_1}\Delta_{1ik}$ is the *pilot group mean difference* at voxel *k*,(63)

and $S_{1k}^2 = \frac{1}{N_1 - 1}\sum_{i=1}^{N_1}\left(\Delta_{1ik} - \overline{\Delta}_{1\bullet k}\right)^2$ is the variance estimate (64)

Any method for the selection of regions of interest will give a valid test. As well as the pilot study statistic image, additional information can be used to aid region selection. Anatomical information can be used to exclude white matter and ventricular regions from selection. Clinicians prior beliefs can be incorporated, and the interesting regions chosen manually. However, many experimenters prefer automated procedures that eliminate operator judgement from analyses, giving reliable methods that always give the same answer for any particular data. Further, for study by simulation, an automated method of region identification is required.

The method we shall adopt is as follows. Threshold the pilot group *t*-statistic image at a fairly low threshold, and note the clusters of voxels with supra-threshold values. Choose the "most interesting" of these clusters, where the interest of a cluster is measured in terms of the size of the region enclosed between the suprathreshold portion of the statistic image and the threshold plane. We shall call this the *excess weight* of the suprathreshold cluster. For a one-dimensional image, the excess weight is an area (fig.58).

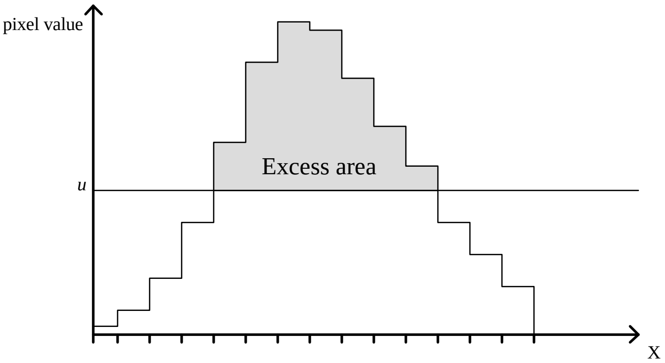

Figure 58
Excess area for a supra-threshold cluster in a one-dimensional image.

In detail, `TSmaxNoR`[55] ROI are chosen as follows: $\boldsymbol{T}_1$ is thresholded at a level $u_1 = t_{N_1\text{-}1,1\text{-}\eta_1}$, the 1-$\eta_1$ point of a Student's *t*-distribution with N_1 -1 degrees of freedom. Suppose there are *R*' supra-threshold clusters of voxels, where cluster *r* consists of voxels $k \in \mathrm{U}'_r$, $r = 1,\ldots R'$. If $R' \leq$ `TSmaxNoR` then we are done: These *R*' clusters define the ROI. If $R' >$ `TSmaxNoR`, then compute the region excess weights as ω_r (eqn.65), and choose as ROI the regions defined by the `TSmaxNoR` clusters with greatest weights.

$$\omega_r = \sum\nolimits_{k \in \mathrm{U}'_r} (T_{1k} - u) \times \texttt{VoxSize}\,, \tag{65}$$

where `VoxSize` is the volume of a voxel.

If a low threshold (corresponding to "large" η_1) is chosen, then the probability of there being no supra-threshold clusters from the pilot group statistic image is very low. In the unlikely event that this does happen, suitable courses of action may be to lower the threshold (increase η_1) until supra-threshold clusters are found, or to proceed with a voxel-by-voxel analysis of the study group data.

Suppose *R* (= min{*R*',`TSmaxNoR`}) regions of interest are identified for examination in the study stage, consisting of voxels with indices $k \in \mathrm{U}_r$, $r = 1,\ldots R$.

Stage 2: ROI analysis of study group

The study stage of the two-stage test is an ROI analysis of the study group data, using the *R* regions of interest identified in the pilot stage. A *t*-test is performed for each region, at a level adjusted for the number of regions under consideration by Bonferroni (or Šidák) correction. The data for a region *t*-test are the means of the subject difference images over that region.

For each study group subject, the mean of the subject difference image over each region is computed, giving data X_{ir}:

$$X_{ir} = \frac{1}{\mathrm{card}(\mathrm{U}_r)} \sum\nolimits_{k \in \mathrm{U}_r} \Delta_{2ik}$$

Here, card(U_r) is the cardinality of set U_r, the number of elements in U_r.
One sample *t*-statistics for each region, T_{2r}, are computed in the usual way:

$$T_{2r} = \frac{\overline{X}_{\bullet r}}{\sqrt{S_r^2 / N_2}}$$

$$\text{where} \quad S_r^2 = \frac{1}{N_2 - 1} \sum_{i=1}^{N_2} \left(X_{ir} - \overline{X}_{\bullet r} \right)^2$$

Assume $X_{ir} \sim N(\mu_r, \sigma_r^2)$. Under the null hypothesis of no (overall) activation in region *r*, $\mathrm{H}_{\mathrm{U}_r}\!:\mu_r = 0$, $T_{2r} \sim t_{N_2\text{-}1}$, a Student's *t*-distribution with N_2-1 degrees of freedom. The alternative hypothesis of activation is one-sided $\overline{\mathrm{H}}_{\mathrm{U}_r}\!:\ \mu_r > 0$. For a level α test by Bonferroni correction, the *R* region *t*-statistics T_{2r} are compared with the 1-α/R point of Student's *t*-distribution with N_2-1 degrees of freedom. Unadjusted *p*-values for each region are given by $1\text{-}\mathrm{F}_{t,N_2\text{-}1}(T_{2r})$, where $\mathrm{F}_{t,df}(\bullet)$ is the CDF of Student's *t*-distribution with *df* degrees of freedom. These give Bonferroni adjusted *p*-values as the smaller of 1 and $R(1\text{-}\mathrm{F}_{t,N_2\text{-}1}(T_{2r}))$.

[55] Words in typewriter font, such as `TSmaxNoR`, are to be read as variable names.

The Bonferroni correction here is unlikely to lead to particularly conservative tests since the number of ROI is small, and since under the null hypotheses $\mathrm{H}_{\mathrm{U}_r}$, the region *t*-statistics are unlikely to be highly correlated for reasonably separated ROI.

4.2. Simulation Methods

To assess the usefulness of this two stage approach, a simulation study was carried out. Sets of subject difference images were simulated, and the two-stage approach compared with some of the voxel-by-voxel approaches previously described in ch.3.

4.2.1. Image space

The image space considered was a two-dimensional arrangement of $K = 64 \times 64$ square pixels. In addition to the usual method of referring to pixels by their index, we shall refer to pixels in *pixel co-ordinates*, referring to a pixel by X and Y displacements in pixels from an origin chosen such that the "bottom left" pixel of the image space has co-ordinates (1,1) (appendix A). The pixels were taken to be of dimensions 2mm×2mm. This gives an image area of 4096 pixels, representing an area of 16384mm^2, which is similar to that of a central slice from a PET image. However, for convenience we shall work in units of pixels, with one "pixel" representing 2mm when indicating a length.

Toroidal image space

To avoid edge effects, a periodic boundary was assumed, equivalent to wrapping the image space round a torus. The top and right of the image space were considered to touch the bottom and left of the image space, respectively. In pixel co-ordinates, pixel (x,64) was considered to be to the immediate left of pixel (x,1) (x =1,…,64), and pixel (64,y) was considered to be immediately below pixel (1,y) (y =1,…,64) (fig.59).

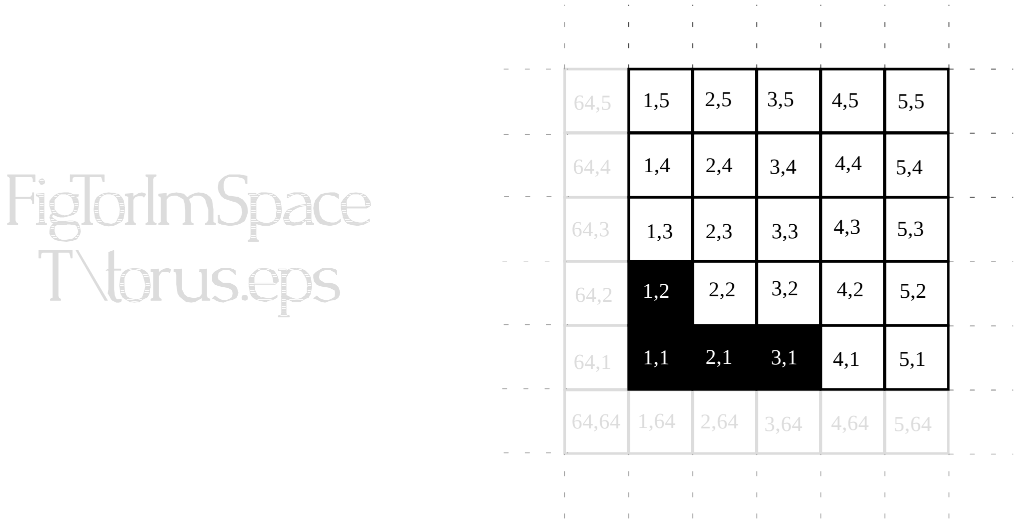

Figure 59
The toroidal image space.
The image space as the pixellated surface of a torus is shown on the left. On the right are shown the neighbours of pixels in the "bottom left" of the 64× 64 image space, when considered as an unwrapped torus. The numbers in the squares are the pixel co-ordinates of the pixel. Four pixels in an "L" shaped cluster are shaded to indicate the orientation of the image space when wrapped onto the torus.

4.2.2. Simulating difference images

Subject difference images

N null subject difference images Δ_{gik} (g = 1,2, $i = 1,\ldots,N_g$, $k = 1,\ldots,K$), were generated by smoothing Gaussian white noise images with a (discretised) Gaussian kernel of FWHM 5 pixels (10mm). The Gaussian white noise images were generated by associating a zero mean, unit variance Gaussian variate with each pixel (fig.60).

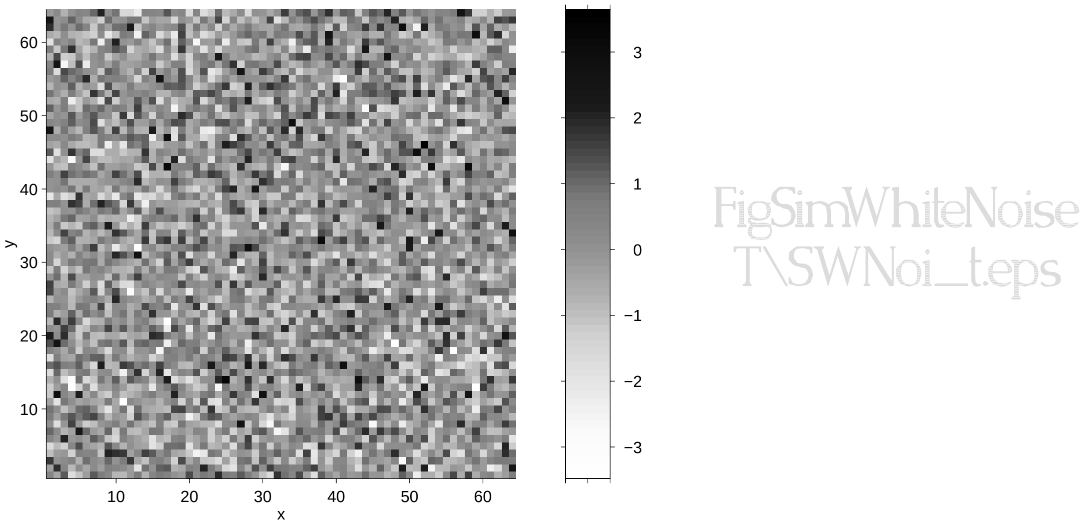

Figure 60
Example white noise image, shown (on the left), with X and Y axes graduated in pixel co-ordinates, and (on the right) folded into a torus.

A Gaussian filter kernel of FWHM 5 pixels was implemented as a moving average filter, with weights computed by evaluating a bivariate Gaussian PDF with zero mean and variance-covariance matrix $\mathbf{\Sigma} = \begin{pmatrix} 5^2 & 0 \\ 0 & 5^2 \end{pmatrix} \frac{1}{8\ln(2)}$, on a regular 17×17 array of points 1 unit (pixel) apart, centred at the origin (fig.61). The smoothing weights were normalised to sum to unity. (See appendix B for details of smoothing and Gaussian kernels.)

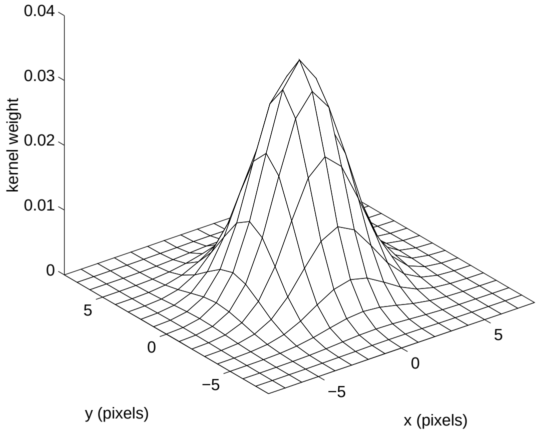

Figure 61
Mesh plot of the (discretised) Gaussian filter kernel.
The vertices of the mesh correspond to the points of the lattice.
The X and Y axes graduated in pixel co-ordinates.

The white noise images were smoothed with the Gaussian moving average filter, which was applied respecting the toroidal structure of the image space. The resulting image was normalised to give simulated difference images with unit variance, by division by the square root of the sum of the squares of the weights of the filter kernel. The sum of the squares of the weights of the filter kernel was found to be close to the theoretical value for continuous two dimensional fields of $1 / (4\pi\sqrt{|\Sigma|})$, as given in appendix C:5.

The null difference images created thus, are strictly stationary discrete Gaussian random fields with zero mean and unit variance. Furthermore, the PRF is equal to the smoothing kernel used, a (discretised) Gaussian kernel of FWHM 10mm. The variance-covariance matrix of partial derivatives of the field is $\boldsymbol{\Lambda}=(2\boldsymbol{\Sigma})^{-1}$ (appendix C:7). An example of a simulated null difference image is shown in fig.62.

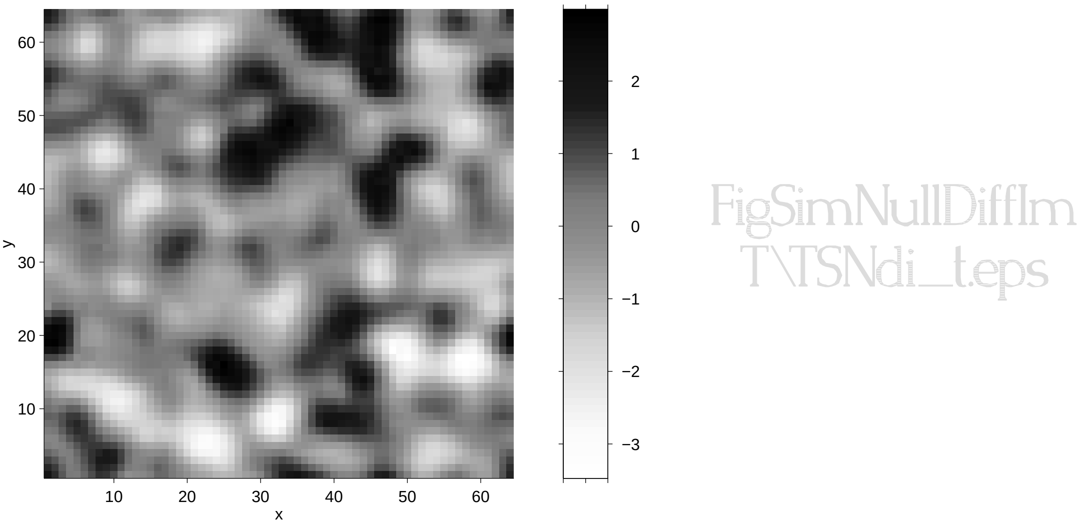

Figure 62
Example simulated null difference image, shown (on the left), with X and Y axes graduated in pixel co-ordinates, and (on the right) folded into a torus.

Signal

To simulate deviations from the omnibus hypothesis, a signal was added to each of the null subject difference images. The signal is a PRF convolved with itself, scaled to have maximum height `SigAmp`, and located at pixel (32,32). This is a fairly common focal signal for use in PET simulation experiments. Since the PRF is a Gaussian kernel of FWHM 5 pixels (equivalent to 10mm), and hence variance-covariance matrix $\boldsymbol{\Sigma} = \begin{pmatrix} 5^2 & 0 \\ 0 & 5^2 \end{pmatrix} \frac{1}{8\ln(2)}$, the signal is a bivariate Gaussian kernel with mean (32,32)', and variance-covariance matrix $2\boldsymbol{\Sigma}$ (appendix C:4), scaled to the appropriate height by multiplication by `SigAmp`$\times 2\pi\sqrt{|2\boldsymbol{\Sigma}|}$. The signal therefore has FWHM of $5\sqrt{2}$ pixels (equivalent to $10\sqrt{2}$ mm).

The signal is discretised by evaluating it on the 64×64 lattice of voxel centres, giving an image of the signal (fig.63), which was added to each of the simulated null subject difference images by adding the values of the two images at each pixel.

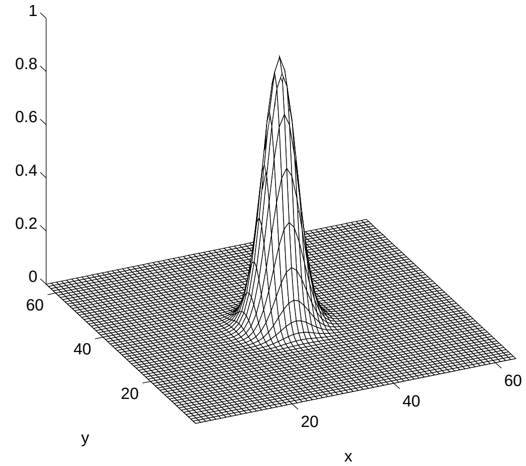

Figure 63
Mesh plot of the signal at unit amplitude.
X and Y axes graduated in pixel co-ordinates.

Correct rejections

Let V be the set of (indices of) the 9 pixels in the 3×3 square centred at (32,32) (fig.64). If a test rejects H_V, by rejecting the null hypotheses for any pixel in V, or by rejecting the hypothesis for a region containing a pixel in V, then the test was considered to have correctly identified the signal. The "true" power of a test at a given `SigAmp` was defined as the probability of rejection of H_V.

31, 33	32, 33	33, 33
31, 32	32, 32	33, 33
31, 31	32, 31	33, 31

Figure 64
3×3 square of voxels at the signal centre.
The numbers in the squares are the pixel co-ordinates of the pixel.

4.2.3. Two-Stage test implementation

Pilot group t-statistic image

The first N_1 simulated difference images were taken as the pilot group difference images, and from these the pilot group t-statistic image, $\boldsymbol{T}_1$ (eqn.62), was formed. (Fig.65)

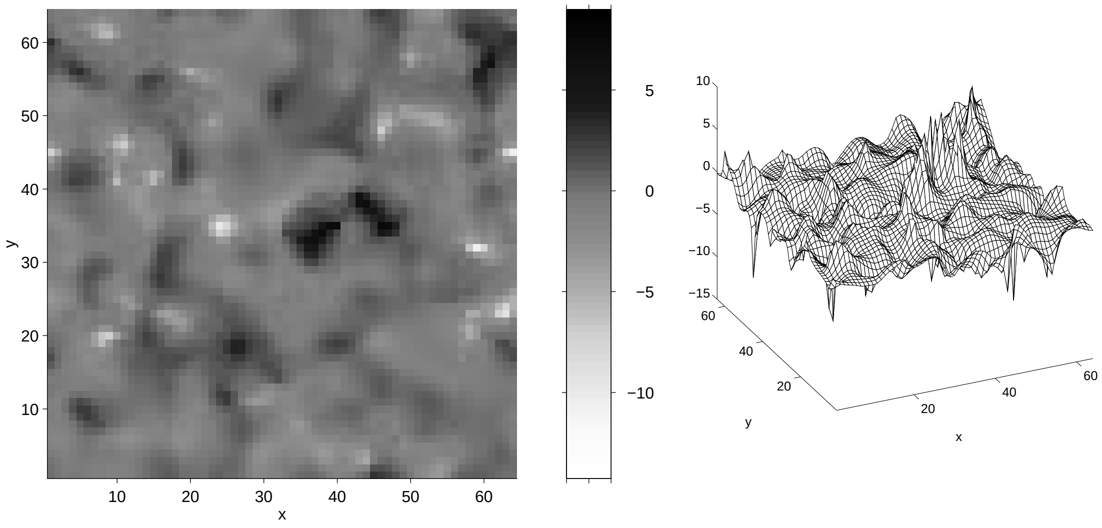

Figure 65
Example pilot group t-statistic image computed from $N_1 = 4$ simulated null subject difference images, shown (on the left), with X and Y axes graduated in pixel co-ordinates, and (on the right) as a mesh plot.

This pilot group t-statistic image is then thresholded at level $u_1 = t_{N_1\text{-}1,1\text{-}\eta_1}$. The pixels with supra-threshold values are the *interesting pixels.*

Interesting cluster identification

The clusters of the interesting pixels were identified as connected subsets of pixels, using a first order neighbourhood scheme. Under such a scheme, two pixels are neighbours if they share a side. Two pixels are *connected* if they can be joined by a path of pixels, in which each pixel in the path is a neighbour of the last. This cluster identification was carried out respecting the toroidal nature of the image space.

This results in a *region map*, an image where the value of each pixel is the number of the cluster it is a member of, or zero if the pixel has sub-threshold pilot *t*-statistic (fig.66)

```
   +------------------------------------------------------------------+
64 | ............................................AAAA................ |
63 | .............................................AAA.........RR..... |
62 | .........................................................RRRRR.. |
61 | .........................................................RRRRRRR |
60 | R.........................................................RRRRRR |
59 | R..........................................................RRRRR |
58 | R..........................................................RRRRR |
57 | ..GG.......................................................RRRRR |
56 | ...GGG....................................................RRRR.. |
55 | ....GG.......SS...........................................RRR... |
54 | .............SS.................Q.........................RRR... |
53 | ...............................QQ.........................RRR... |
52 | ..............................QQQ..........................R.... |
51 | ..............................QQ................................ |
50 | ................................................................ |
49 | ................................................................ |
48 | .......................................PP....................... |
47 | ......................................PPP....................... |
46 | .......................................PPP...................... |
45 | ..................N.......................O..................... |
44 | ..................N............................................. |
43 | ..................N............................................. |
42 | ...MMM............N............................................. |
41 | ...MM........................................................... |
40 | ................................................................ |
39 | .........................................LLLL................... |
38 | .........................................LLLLL.................. |
37 | .........................................LLLLLL................. |
36 | ...................................KKKKK...LLLLLL............... |
35 | ..................................KKKKKK....LLLL................ |
34 | ................................KKKKKKK.....LLLL................ |
33 | .................................KKKKKK.....L................... |
32 | ................J.................KKKK.......................... |
31 | ...............JJ..................KKK.......................... |
30 | ...............JJ..................KK........................... |
29 | ...............JJ............................................... |
28 | ..............JJJ............................................... |
27 | ...............JJ............................................... |
26 | ................................................................ |
25 | ................................................................ |
24 | ................................................................ |
23 | ................................................................ |
22 | ................................................................ |
21 | .........................F...................................... |
20 | .............H..........FFFF.................................... |
19 | .............HH.........FFFFF.........III....................... |
18 | ..............H........FFFFFF.........II......................DD |
17 | D...............EE......FFFF...................................D |
16 | ................................................................ |
15 | ................................................................ |
14 | ................................................................ |
13 | ................................................................ |
12 | .......................CC....................................... |
11 | ........................C....................................... |
10 | ....BB.......................................................... |
 9 | ....BBB......................................................... |
 8 | ......B......................................................... |
 7 | ................................................................ |
 6 | ................................................................ |
 5 | ................................................................ |
 4 | ................................................................ |
 3 | ................................................................ |
 2 | ............................................A................... |
 1 | ............................................AAA................. |
   +------------------------------------------------------------------+
     0000000001111111111222222222233333333334444444444555555555566666
     1234567890123456789012345678901234567890123456789012345678901234
```

Figure 66

Example region map computed from the *t*-statistic image of fig.65, thresholded at $u_1 = -t_{3,0.1} = 1.6377$ (4dp). A region map is an integer image, with pixel values indicating the cluster the pixel belongs to, zero indicating membership of no clusters. For display, zero has been mapped to the character '.', and the integers 1,2,3,… to the characters 'A','B','C',…. Note how clusters A, D & R are defined across the edges of the image space, due to its consideration as an unwrapped torus. The X and Y axes are graduated in pixel co-ordinates.

For each of the *R*' clusters of pixels identified, the excess weight was computed by eqn.65. If more than TSmaxNoR clusters of pixels were identified, then the TSmaxNoR with largest excess weight were chosen to be the *R* ROI for the study stage. Otherwise, all *R*' clusters were taken as the ROI for the study stage. In the unlikely event (for a low threshold) that no interesting pixels were obtained, the whole image was considered as a single region. In practice, something more useful would be done, but this was a convenient conservative approach for the simulation experiment.

Stage 2

The study stage of the two stage test on the simulated subject difference images was implemented as described in §4.1.

4.2.4. Voxel-by-voxel tests

Voxel-by-voxel approaches were applied to the *t*-statistic image, $\boldsymbol{T} = (T_1,\dots,T_K)$, formed from all *N* simulated subject difference images, as described in §2.3.1. It was assumed that $\Delta_{gik} \sim N(\mu_k,\sigma_k^2)$; so under $\mathrm{H}_k{:}\mu_k = 0$, $T_k \sim t_{N\text{-}1}$, a Student's *t* distribution with

N-1 degrees of freedom . As usual, one-sided alternative hypotheses were considered; $\overline{\mathrm{H}}_k$:μ_k> 0.

The voxel-by-voxel approaches considered were a simple Bonferroni approach and Worsley's expected Euler characteristic approach for *t*-fields. In addition, Worsley's expected Euler characteristic approach for Gaussian fields, and Friston's supra-threshold cluster size test were applied to the *t*-statistic image after it had been "Gaussianised".

Bonferroni

The Bonferroni approach was described in §3.2.1. The critical level c_α was computed as the 1-α/K point of the *t*-distribution with *N*-1 degrees of freedom. The *t*-statistic image was thresholded at this level, and pixels with supra-threshold *t*-statistics had their null hypotheses rejected. Rejection of any pixel hypothesis implies rejection of the omnibus hypothesis H_W.

Worsley's expected Euler approach for t-fields

Worsley's expected Euler characteristic method, for a two-dimensional strictly stationary continuous random *t*-field was applied. The method for Gaussian fields was described in §3.3.1., and the expected Euler characteristic of the excursion set of a strictly stationary continuous random *t*-field is given in appendix D:3. For convenience, we shall refer to this test as "Worsley's $T_{\max}$" test.

The equation for the expected Euler characteristic of the excursion set of a strictly stationary continuous random *t*-field thresholded at level *u* was set to α and solved to obtain the critical threshold u_α. The variance-covariance matrix of partial derivatives of the component fields used was the theoretical value for a strictly stationary continuous Gaussian field with zero mean and unit variance formed by convolving a white noise Gaussian process with a Gaussian kernel with variance-covariance matrix $\mathbf{\Sigma}$. The value is $\mathbf{\Lambda} = (2\mathbf{\Sigma})^{-1}$ (appendix C:7). Recall $\mathbf{\Sigma} = \begin{pmatrix} 5^2 & 0 \\ 0 & 5^2 \end{pmatrix} \frac{1}{8\ln(2)}$. As measurements are in pixels, the size of the domain of the field, $\lambda(\Omega)$, is simply the number of pixels, *K*.

The *t*-statistic image was thresholded at this level, and pixels with supra-threshold *t*-statistics had their null hypotheses rejected.

Gaussianised t-statistic image

The *t*-statistic image was transformed to a Gaussian statistic image, by replacing each pixel *t*-statistic with a Gaussian variate with equal probability of being exceeded. (See appendix E for details.) As noted in §3.3.3., although the resulting statistic image has Gaussian marginal distributions under H_W, it is not a discrete Gaussian field. However, many practitioners apply tests for Gaussian random fields to such "Gaussianised" statistic images, and we shall do likewise.

The variances of the partial derivatives of the Gaussianised statistic image were estimated within the image. This was accomplished by taking numerical derivatives at each voxel in the X and Y directions, as a difference in the pixel values, and computing the sample variance of these numerical derivatives over the image space. Since the (square) image space is considered as the unfolded surface of a torus, every pixel has neighbouring pixels in both the X and Y directions. Assuming the covariances of the partial derivatives are zero, this gives an estimate $\hat{\mathbf{\Lambda}}$ of the variance-covariance matrix of partial derivatives $\mathbf{\Lambda}$. Recall §3.3.5. for further details of estimating smoothness.

Worsley's expected Euler approach for Gaussianised statistic image

Worsley's expected Euler characteristic method for a two-dimensional strictly stationary continuous Gaussian random field with (hypothesised) zero mean and unit variance, was applied to the Gaussianised *t*-statistic image as described in §3.3.1. For convenience, we shall refer to this test as "Worsley's $Z_{\max}$" test.

The equation for the expected Euler characteristic of the excursion set of a 2D strictly stationary continuous standard Gaussian random field thresholded at level u (eqn.36), was set to α and solved to obtain the critical threshold u_α. The estimated variance-covariance matrix of partial derivatives for the Gaussianised *t*-statistic image, $\hat{\Lambda}$, was used. Again, the size, $\lambda(\Omega)$, of the domain of the field is simply the number of pixels, K.

The "Gaussianised" *t*-statistic image was thresholded at this level, and pixels with supra-threshold statistics had their null hypotheses rejected. Rejection of any pixel hypothesis implies rejection of the omnibus hypothesis.

Friston's supra-threshold cluster size test on "Gaussianised" statistic image

The supra-threshold cluster size of Friston *et al*. (1994d) was described in §3.5.2. The "Gaussianised" *t*-statistic image was thresholded at $u = -\Phi^{-1}(\eta_F)$, and clusters of supra-threshold pixels identified using the toroidal clustering algorithm used in the pilot stage of the two stage test. For each cluster, the size in pixels was computed. We shall refer to this test as "Friston's $S_{\max}$" test.

Using K as the size of the field, and the estimated variance-covariance matrix of partial derivatives of the "Gaussianised" *t*-statistic, $\hat{\Lambda}$, the critical cluster size s_α for a level α test was computed (eqn.60). Since the size of the domain of the field and the smoothness have been measured in pixels, the critical cluster size is also in pixels.

If a supra-threshold cluster of pixels, with set of indices U, had size greater than the critical size s_α, then the omnibus hypothesis H_U for the pixels in that cluster was rejected. If any regional hypothesis was rejected then the omnibus hypothesis H_W was rejected.

4.3. Results

4.3.1. Simulation parameters

Simulations were carried out with the following parameters:

General		
	Simulations:	`NoSim` = 2 000
	Image dimensions:	$D = 2$ dimensions
	Image space:	Toroidal array of 64 × 64 square pixels
	Number of pixels:	$K = 64 \times 64 = 4096$
	Numbers of subjects:	$N = 12$
	Level for testing:	$\alpha = 0.05$
	Smoothing kernel:	Gaussian kernel, FWHM of 5 pixels (=10mm)
Signal parameters		
	Signal shape:	Gaussian PRF, FWHM $5\sqrt{2}$ pixels, Located at pixel (32,32)
	Signal amplitudes:	{0, 0.9, 1.1, 1.3, 1.4, 1.5, 1.6, 1.7, 1.9, 2.1, 2.3} (The same `NoSim` sets of N null subject difference images were used for each signal amplitude.)
Two-Stage parameters		
	Pilot group size:	$N_1 = 4$
	Study group size:	$N_2 = N - N_1 = 8$
	Pilot stage threshold:	$\eta_1 = 0.1$ $\quad u_1 = t_{N_2\text{-}1,\, 1\text{-}\eta_1} = 1.6377$ (4dp)
	Max number of ROI:	`TSmaxNoR` = 5
Bonferroni parameters		
	Critical threshold:	$c_\alpha = t_{N\text{-}1,\, (1\text{-}\alpha)/K} = 6.9442$ (4dp)
Worsley’s expected Euler test for t-fields		
	Critical threshold:	6.8048 (4dp)
Friston’s supra-threshold cluster size test		
	Threshold:	$\eta_F = 0.01$ $\quad u = -\Phi^{-1}(\eta_F) = 2.3263$

The only parameter that was varied was the signal amplitude, `SigAmp`.

4.3.2. Size of tests

The null simulation, with a signal amplitude of zero, was undertaken to assess the validity and relative sizes of the tests. Individual 95% CIs for the sizes of the various tests for the simulated data are given in table 67. It appears that all the tests are valid, with size at most the desired level $\alpha = 0.05$. Worsley's T_{max} and the Bonferroni approach are rather conservative, with test level well below the desired size.

Two-Stage	(0.0457, 0.0623)
Bonferroni	(0.0088, 0.0172)
Worsley's T_{max}	(0.0122, 0.0218)
Worsley's Z_{max}	(0.0392, 0.0548)
Friston's S_{max}	(0.0346, 0.0494)

Table 67

Individual 95% CIs for the sizes of the tests on the simulated null data, unadjusted for multiple comparisons. Computed from 2 000 simulations, to 4dp, using the normal approximation to the binomial.

4.3.3. Power of two-stage test

Having demonstrated the validity of the tests under study, we move on to compare the power of the two-stage method with the voxel-by-voxel methods.

True power curves

Estimated true power curves for the five tests are given in fig.68. Departures from H_W are parameterised by the amplitude of the added signal. Recall that the true power of the test at a given `SigAmp` (> 0) was defined as the probability of rejection of H_v, where V is the set of the 3×3 pixels at the centre of the signal. Thus, these curves do not show the size of the omnibus test when `SigAmp` = 0. Rather, they show the probability of detection of the centre of a signal which has amplitude zero, and as such aremeaningless. The point `SigAmp` = 0 is therefore omitted from these true power curves.

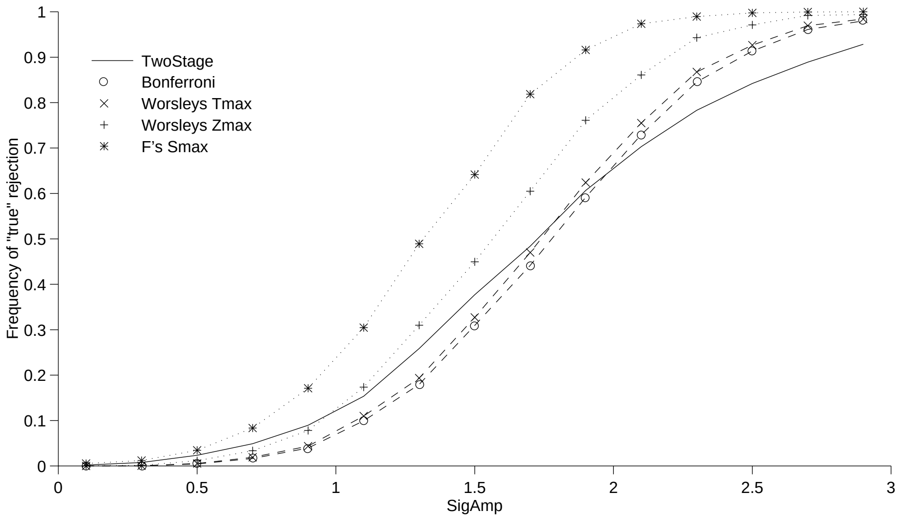

Figure 68
Estimated power curves from the simulations. The power at a given `SigAmp` (> 0) is the probability of rejection of H_V, where V is the set of the 3×3 pixels at the centre of the signal. 2 000 simulated data sets were generated, to which the tests were applied for each `SigAmp`

McNemar's test

The tests were applied to the set of simulated subject difference images at each signal amplitude. Thus, for any two tests on the data at one signal amplitude, the results of the simulation are paired. To compare any two tests, the pairing should be exploited using McNemar's test.

Consider as an example the comparison of the two-stage test and Worsley's T_{max} test over the 2 000 simulations, with `SigAmp` = 1.5. The results of the simulation are summarised in the following 2×2 table:

1 = “reject” H_v		Worsley’s T_{max}		
		0	1	
Two-Stage	0	1014	232	1246
	1	332	422	754
		1346	654	2000

The cases where both tests reach the same conclusion (on the leading diagonal) give no information regarding the relative sizes of the tests. McNemar's test proceeds by considering the cases where the two tests reach differing conclusions.

Let the number of cases where the two tests differ be n, and let q be the number of times the two-stage test rejects when Worsley's $T_{\max}$ does not, the value in cell (0,1) of the table. Then, conditional on n, q is a binomial $Bin(n, \theta)$ variate, for some θ. The hypothesis of equal probability of rejection for the two tests is H:θ =0.5. The one sided alternative hypothesis $\overline{\mathrm{H}}:\theta > 0.5$ assesses whether the two-stage test has greater probability of rejecting the omnibus hypothesis than Worsley's $T_{\max}$, while the two sided hypothesis $\overline{\mathrm{H}}:\theta \neq 0.5$ tests for a difference in the probabilities of rejection of the two tests. In this latter case the test with greater power may be decided by reference to the results table, with negligible probability of type III error, the error of mistaking the direction of a difference.

Results table

To assess the significance of the differences in the power curves of fig.68, one-sided p-values from McNemar's test comparing the true power of the two-stage approach with each of the four pixel-by-pixel methods in turn were computed:

	On t-statistic image computed for all subjects		On "Gaussianised" t-statistic image	
Signal `SigAmp`	Bonferroni	Worsley's $T_{\max}$	Worsley's $Z_{\max}$	Friston's $S_{\max}$
0.1	0.0625	0.0625	0.1875	0.9824
0.3	0.0003	0.0003	0.0038	0.9552
0.5	0.0000	0.0000	0.0011	0.9942
0.7	0.0000	0.0000	0.0028	1.0000
0.9	0.0000	0.0000	0.0622	1.0000
1.1	0.0000	0.0000	0.9849	1.0000
1.3	0.0000	0.0000	1.0000	1.0000
1.5	0.0000	0.0000	1.0000	1.0000
1.7	0.0004	0.1406	1.0000	1.0000
1.9	0.1344	0.9155	1.0000	1.0000
2.1	0.9837	1.0000	1.0000	1.0000
2.3	1.0000	1.0000	1.0000	1.0000
2.5	1.0000	1.0000	1.0000	1.0000
2.7	1.0000	1.0000	1.0000	1.0000
2.9	1.0000	1.0000	1.0000	1.0000
3.1	1.0000	1.0000	1.0000	1.0000

Table 69

One sided p-values from McNemar's test comparing the two-stage approach with each of the pixel-by-pixel approaches at for each set of 2 000 simulated difference images at each `SigAmp`. p-values given to 4dp

These (unadjusted) p-values show the significance of the difference between the power curves of fig.68, at each signal amplitude.

4.4. Conclusions

4.4.1. Discussion

Power

From the simulation results, it appears that the two-stage approach is more powerful for the simulated data than the Bonferroni or Worsley's $T_{\max}$ method for slight signals, but not for more "obvious" stronger signals. The two methods for "Gaussianised" statistic images considered, Worsley's $Z_{\max}$ and Friston's maximum supra-threshold cluster size test, are more powerful than the two-stage procedure.

There are three features of the two-stage approach to consider. Firstly, the first stage, with only $N_1 = 4$ subjects data may not identify the activated region as an interesting ROI for the second stage. Secondly, a two-stage approach has a smaller study group than a single stage method. These drawbacks may be overcome by the third feature, that the number of comparisons in the study stage is small.

The first point would explain why the two-stage outperforms Worsley's $Z_{\max}$ and the Bonferroni approaches for small signal amplitudes but not for larger ones: The first stage "misses" the signal sometimes.

Robustness of two-stage

A key feature of the two-stage approach is its robustness. The only assumption is of normality of the means of the identified ROI in the N_2 study stage subjects, an assumption which seems in little doubt. (Recall the discussion of normality of voxel values of rCBF scans given in §3.3.6.2.)

In contrast, the voxel-by-voxel methods rely on many assumptions, discussed in ch.3, which at best only approximate the truth for real PET data. The effects on the size and power of the tests of departures from the assumptions is only known in a few limited situations.

Of additional concern is the application of methods for Gaussian random fields to "Gaussianised" t-fields. Although the simulation at zero amplitude gave insufficient evidence against a hypothesis of size = α for methods on "Gaussianised" t-statistic images, recall (§3.3.3.) that Worsley (1994b) found his $Z_{\max}$ approach applied to "Gaussianised" three-dimensional t-statistic images to be invalid, with size greater than nominal level α.

The simulation method used here gives ideal conditions for the random field approaches. The simulated t-statistic images generated are strictly stationary discrete Gaussian random fields, with zero mean and unit variance. The "Gaussianised" statistic images are strictly stationary, and have Gaussian marginal distributions with zero mean and unit variance, even if they are not discrete Gaussian fields.

Prospects for two-stage

The two-stage approach has prospects. The prototype two-stage algorithm presented here is robust, and is more sensitive than Worsley's $T_{\max}$ random field method for the simulated data. However, it should be borne in mind that single threshold methods such as Worsley's $T_{\max}$ maintain strong control over FWE at the voxel level, which may be more desirable than a slight increase in sensitivity.

Clearly further examination of the two-stage approach is necessary before it can be adopted (or rejected) as a test for functional mapping experiments.

4.4.2. Further work

As with all simulation experiments, there is a continuum of possible configurations that can be explored. The limited two-dimensional simulation study presented in this chapter gives a rough idea of the potential of the two-stage approach. To gain a better idea, the following improvements could be considered:

Three-dimensions

Firstly, it is desirable to match the parameters of the simulation to the real situation. The most obvious drawback of the current work in this respect is the two-dimensional image space used (the pixellated surface of a torus), when rCBF images are three-dimensional. The step from a single 2D slice to a full 3D volume represents a vast increase in the multiple comparisons problem. In ch.3, Worsley's $Z_{\max}$ approach was seen to be more conservative in three-dimensions than in two. Thus, the two-stage approach in three-dimensions may be more favourable than in two-dimensions.

A simulation with an image space with similar dimensions, shape, and voxel size to the intracerebral area in PET, perhaps along the lines of the simulation of ch.3, would be desirable for its relevance.

Pilot group sizes

The choice of pilot group size (N_1) is critical to the power of the two-stage method. In this chapter, only a single pilot group size of 4 subjects from am experimental group of size 12 was considered. Simulation studies of the two-stage method for a variety of experimental group sizes would be useful for determining the optimum pilot group size.

Region selection method

Another critical aspect of a two-stage algorithm is the method of region selection from the pilot group data. As discussed earlier, any selection method leads to a valid test, provided that the pilot and study stages are independent. That is, the ROI statistics are independent of the pilot group statistic image and other criteria used to select the ROI.

The selection of regions of interest by suprathreshold cluster excess weight, proposed here, is attractive because it combines the size and magnitude of an excursion above the threshold. A criticism of supra-threshold cluster *size* methods is that a very intense focal activation is not considered as "interesting" as an activation which barely exceeds the threshold, but which does so for a larger cluster of voxels. Clearly there is scope for investigation into better methods for region selection.

It has been noted previously that *t*-statistic images with low degrees of freedom are exceptionally "noisy" (§3.3.6.5.). This noise is reflected in the shapes of the identified ROI. (Consider the example region map of fig.66.) Since it is not necessary to know the marginal distributional of the pixel values in the pilot group statistic image, any image processing tool could be used to "clean up" the image, and enhance the region selection method. Smoothing decreases the pixel variance, but leaves signals greater in spatial extent than the filter kernel intact, making them more obvious, albeit at the expense of resolution. Alternatively, since the high frequency spatial noise in the statistic image is inherited from the variance image $\boldsymbol{S}_1^2$ (eqn.64), the variance image itself could be spatially smoothed prior to formation of the pilot group statistic image, an option pursued in ch.6 with the non-parametric approach described there.

Number of ROI

Perhaps as important as the actual method of region selection, is the actual number of regions to be chosen. Too few, and the probability of missing an activated region may be too great. Too many, and the corrections for the multiple comparisons in the study stage may leave the test with low power. In the two-dimensional simulation study presented above, at most `TSmaxNoR` = 5 ROI were chosen. Comparisons of the power of the two-stage method for different numbers of ROI would help choose an optimal value, and assess the sensitivity of the method to changes in this parameter.

Multiple signals, signal shape

Connected with the last point on the number of ROI, is the issue of the number of signals. The simulation study was carried out with a single focal signal. In most simple activation studies there are only one or two sites of activation, but more complicated paradigms may result in many activations. Further, it may be a secondary activation that is of interest, as in the "V5" study. For a 3D PET activation experiment with multiple activations, `TSmaxNoR` = 5 ROI is possibly too small. A worst case scenario might be to consider a signal consisting of two or three Gaussian "humps" of different amplitudes, and examine the power of the two-stage approach for various numbers of ROI. Further, it might be interesting to vary the shape and spatial extent of the signal.

Real data

Finally, it would be interesting to apply the method to real PET data sets.