Statistical Considerations in Magnetic Resonance Imaging of Brain Function

Brian D. Ripley

Professor of Applied Statistics
University of Oxford

ripley@stats.ox.ac.uk
http://www.stats.ox.ac.uk/∼ripley

Acknowledgements

Most of the computations were done by Jonathan Marchini (EPSRC-funded D.Phil student).

Data, background and advice provided by Peter Styles (MRC Biochemical and Clinical Magnetic Resonance Spectroscopy Unit) and Stephen Smith (Oxford Centre for Functional Magnetic Resonance Imaging of the Brain).

‘Functional’ Imaging

Functional PET and MRI are used for studies of brain function: give a subject a task and see which area(s) of the brain ‘light up’.

Functional studies were done with PET in the late 1980s and early 1990s, now fMRI is becoming possible (needs powerful magnets—that in Oxford is 3 Tesla).

PET has lower resolution, say $3 \times 3 \times 7$ mm voxels at best. So although $128 \times 128 \times 80$ (say) grids might be used, this is done by subsampling. Comparisons are made between PET images in two states (e.g. ‘rest’ and ‘stimulus’) and analysis is made on the difference image. PET images are very noisy, and results are averaged across several subjects.

fMRI has a higher spatial resolution, and temporal resolution of around one second. So stimuli are applied for a period of about 30 secs, images taken around every 3 secs, with several repeats of the stimulus being available for one subject.

The commonly addressed statistical issue is ‘has the brain state changed’, and if so where?

Neurological Change

A longer-term view of function is in the change of tissue state and neurological function after traumatic events such as a stroke or tumour growth and removal. The aim here is to identify tissue as normal, impaired or dead, and to compare images from a patient taken over a period of several months.

In MRI can trade temporal, spatial and spectral resolution. In MR spectroscopy the aim is a more detailed chemical analysis at a fairly low spatial resolution. In principle chemical shift imaging provides a spectroscopic view at each of a limited number of voxels: in practice certain aspects of the chemical composition are concentrated on.

Pilot Study

Our initial work has been exploring ‘T1’ and ‘T2’ images (the conventional MRI measurements) to classify brain tissue automatically, with the aim of developing ideas to be applied to spectroscopic measurements at lower resolutions.

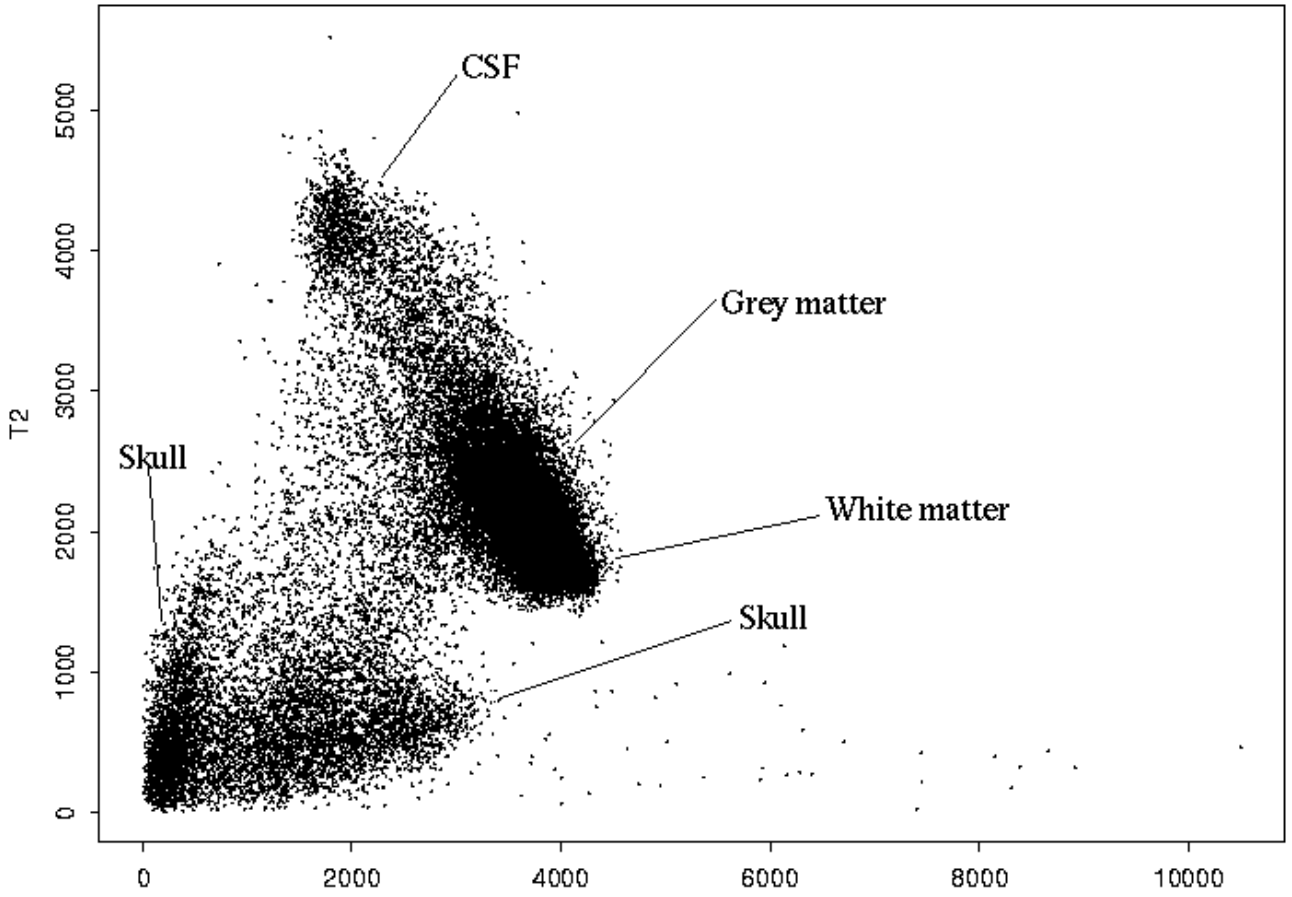
Consider image to be made up of ‘white matter’, ‘grey matter’, ‘CSF’ (cerebro–spinal fluid) and ‘skull’.

Initial aim is reliable automatic segmentation.

Some Data

T1 (left) and T2 (right) MRI sections of a ‘normal’ human brain.

This slice is of 172×208 pixels.



Data from the same image in T1–T2 space.

Imaging Imperfections

The clusters in the T1–T2 plot were surprising diffuse. Known imperfections were:

(a) 'Mixed voxel' / 'partial volume' effects. The tissue within a voxel may not be all of one class.

(b) A 'bias field' in which the mean intensity from a tissue type varies across the image. This effect is thought to vary approximately multiplicatively and to consist of a radial component plus a linear component, the latter varying from day to day.

(c) The 'point spread function'. Because of bandwidth limitations in the Fourier domain in which the image is acquired, the true observed image is convolved with a spatial point spread function of 'sinc' ($\sin x/x$) form. The effect can sometimes be seen at sharp interfaces (most often the skull / tissue interface) as a rippling effect, but is thought to be small.

Bias Fields

There is an extensive literature on bias field correction. One approach uses a stochastic process prior for the bias field, and is thus another re-invention of the ideas known as *kriging* in the geostatistical literature. Based on experience with the difficulty of choosing the degree of smoothing and the lack of resistance to outliers (kriging is based on assumptions of Gaussian processes) we prefer methods with more statistical content and control.

Our basic model is

$$\log Y_{ij} = \mu + \beta_{\text{class}(ij)} + s(i,j) + \epsilon_{ij}$$

for the intensity at voxel (i, j), studied independently for each of the T1 and T2 responses. Here $s(x, y)$ is a spatially smooth function.

Of course, the equation depends on the classification, which will itself depend on the predicted bias field. This circularity is solved by iterative procedure, starting with no bias field.

Estimation

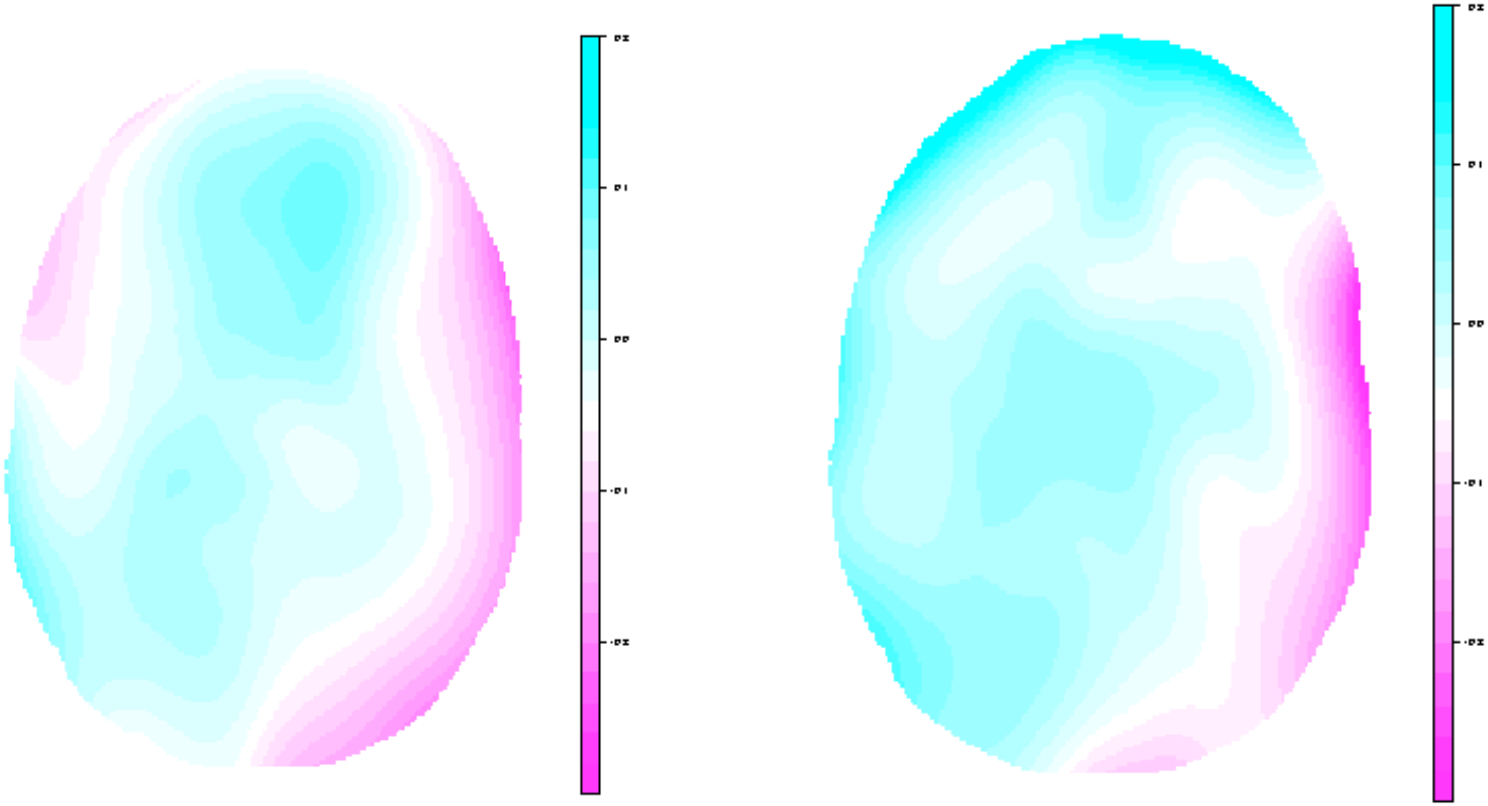
If the classification were known we would use a robust method that fits a long-tailed distribution for ϵ_{ij}, unconstrained terms α_j for each class, and a 'smooth' function s. We cope with unknown class in two ways. In the early stages of the process we only include data points whose classification is nearly certain, and later we use

$$\log Y_{ij} = \mu + \sum_{\text{class } c} \beta_c \, p(c \,|Y_{ij}) + s(i,j) + \epsilon_{ij}$$

that is, we average the class term over the posterior probabilities for the current classification.

For the smooth term s we initially fitted a linear trend plus a spline model in the distance from the central axis of the magnet, but this did not work well, so we switched to *loess*. Loess is based on fitting a linear surface locally plus approximation techniques to avoid doing for the order of 27 000 fits.

Fits of bias fields



Fitted ‘bias fields’ for T1 (left) and T2 (right) images.

The bias fields for these images are not large and change intensity by 5–10%.

Clustering model

Our model is that for the bias-corrected images ($Y_{ij}/\hat{s}(i,j)$) T1 and T2 measurements jointly follow a mixture of six distributions.

The bivariate distributions for the six classes will have elliptical contours. We have experimented with normal and bivariate t distributions: the latter are often desirable as the normal density decays extremely rapidly.

These should correspond to white matter, grey matter, CSF, two types of skull (the distribution of skull seems to split into two ellipses) and ‘outlier’, a diffuse distribution picking up the isolated points.

Our fitting approach is ‘unsupervised’, but the initial locations of the cluster are taken from a reference model whose clusters were labelled by interactive visualization.

Fitting the model

There is an extensive literature on fitting mixtures of densities, and using them as a model for classification. There are two distinct approaches in the literature, and the one we have taken is 'Bayesian' rather than 'maximum likelihood' in character. That is, we estimate the parameters in a finite mixture model

$$p(T1, T2) = \sum_{i=1}^{6} \pi_i \, p_i(T1, T2; \phi_i)$$

and then use the posterior probabilities for a 'soft' classification.

This in contrast to, e.g. Banfield & Raftery (1993), who simultaneously maximize over the parameters *and* the classification of the observations.

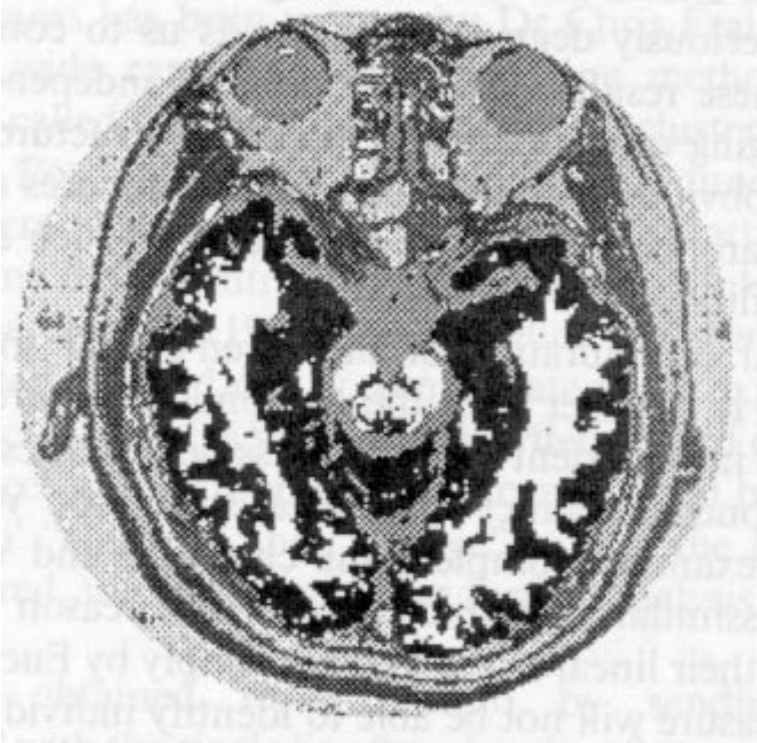
Fitting the parameters (π_i) and (ϕ_i) is currently by maximum-likelihood, although we will investigate using Markov Chain Monte Carlo methods for a full Bayesian scheme with linked hyperpriors.

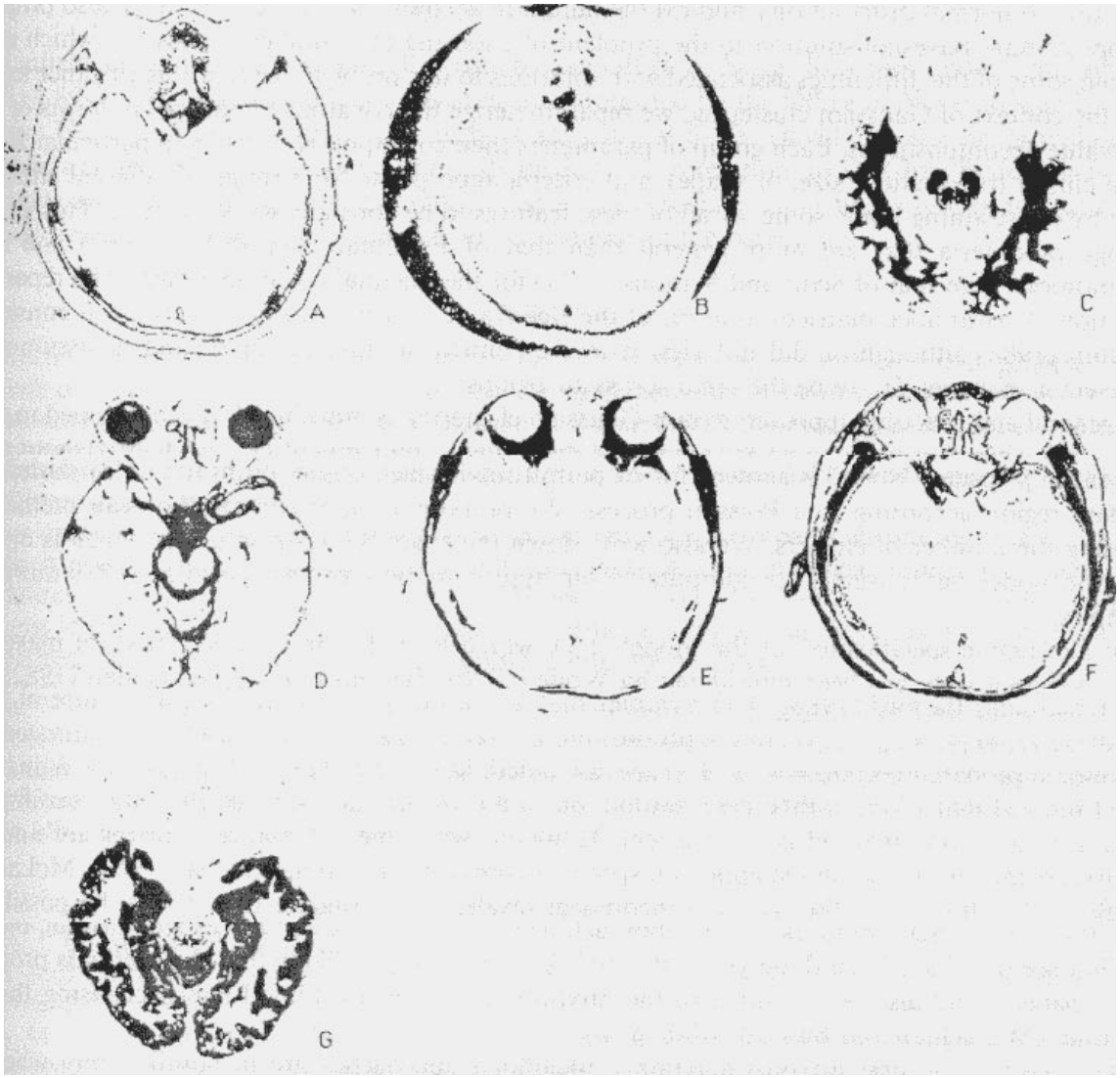
Unsupervised Approach

Banfield & Raftery (1993) *Biometrics*

T1, T2 and PD images of an MRI brain slice with $26\,100$ voxels.

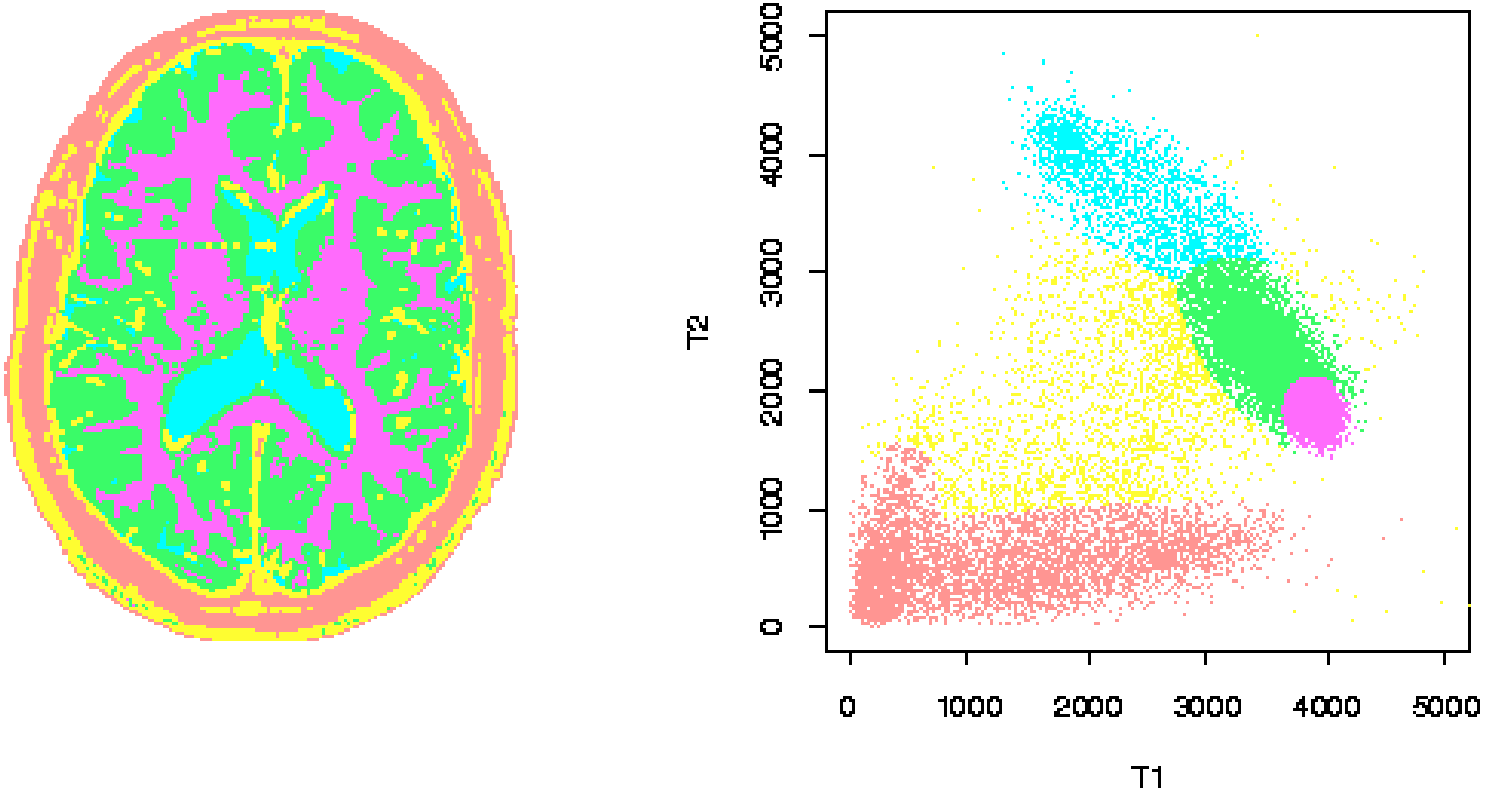
They claim to find seven clusters:



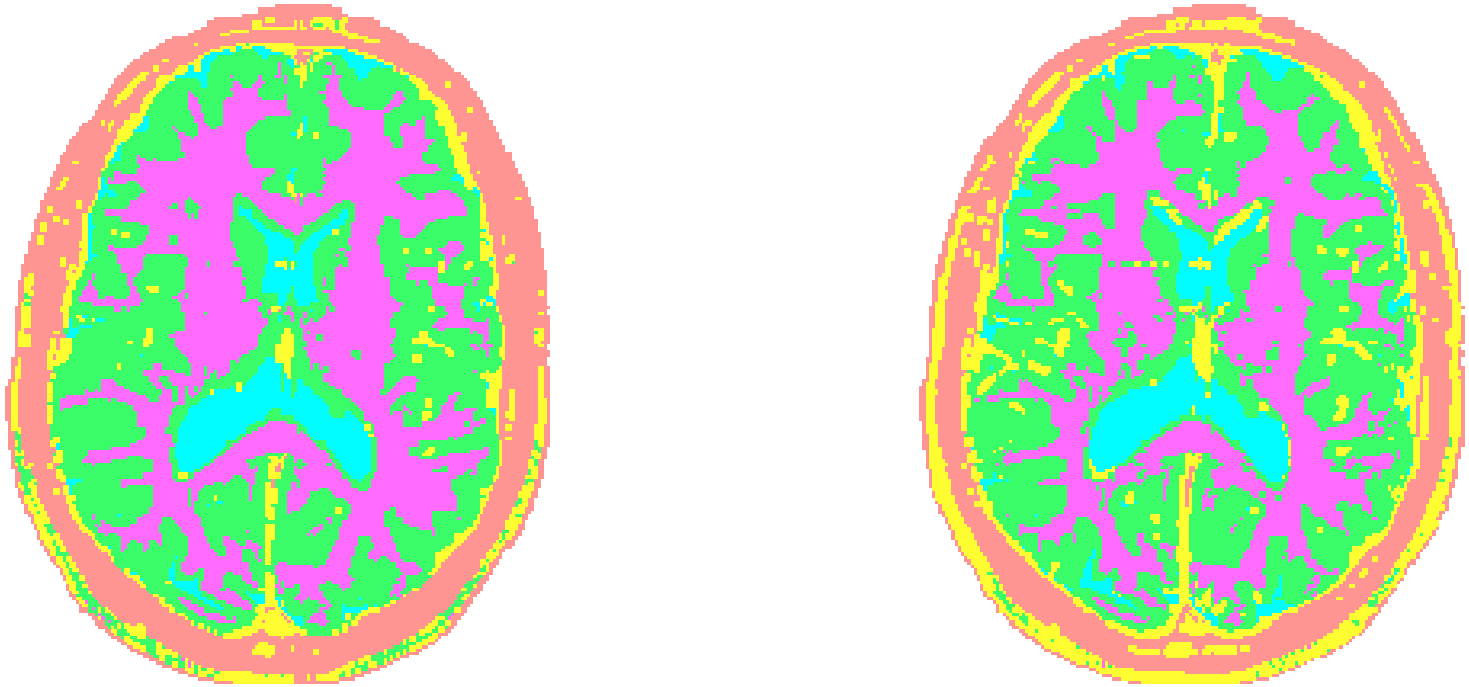


Spatial locations of the 7 ‘clusters’, labelled in retrospect as A: Bone, B: Air: C: White matter, D: Fluids, E: Muscle, F: Fat and G: Grey matter.

Our Results

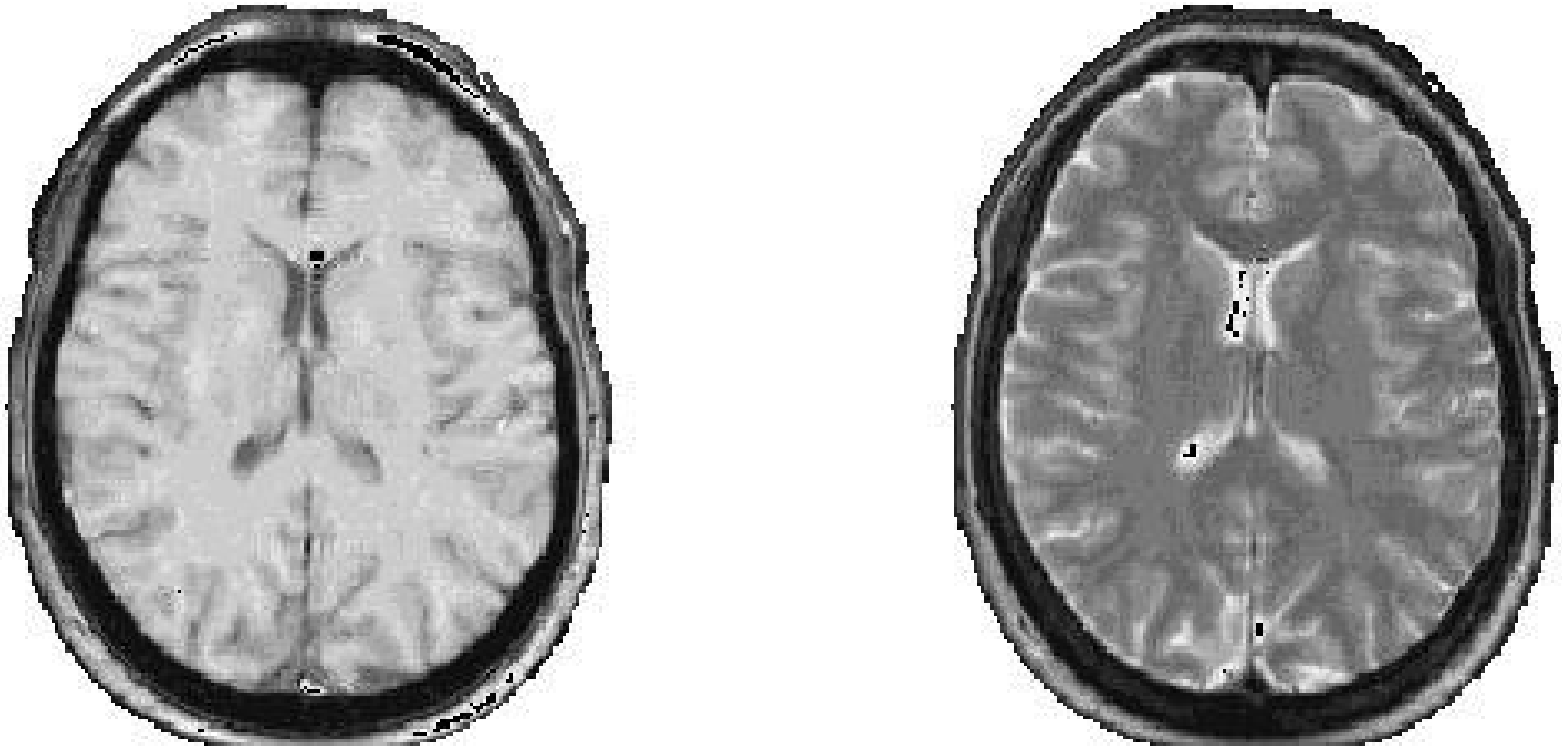


Classified image and T1–T2 plot showing the classification with normally-distributed clusters.

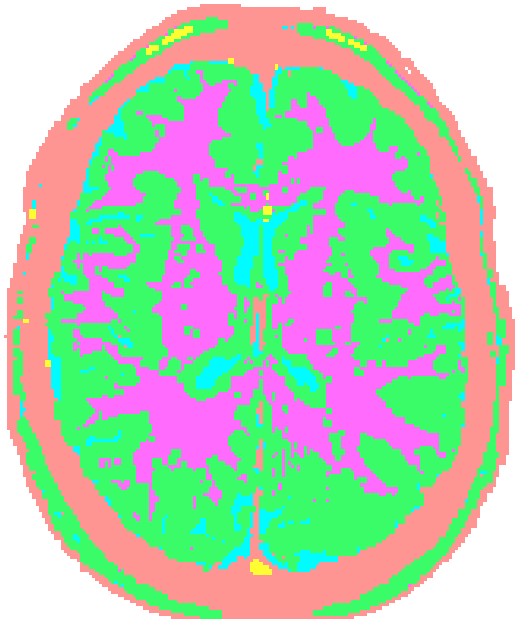


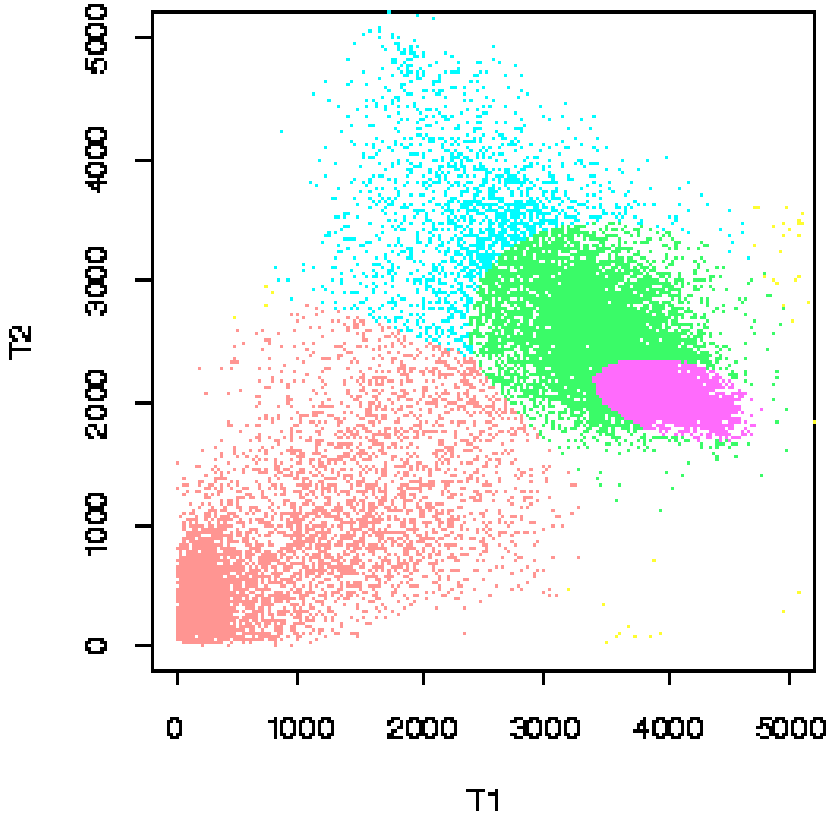
Classification before and after removing the bias field.

A Second Dataset

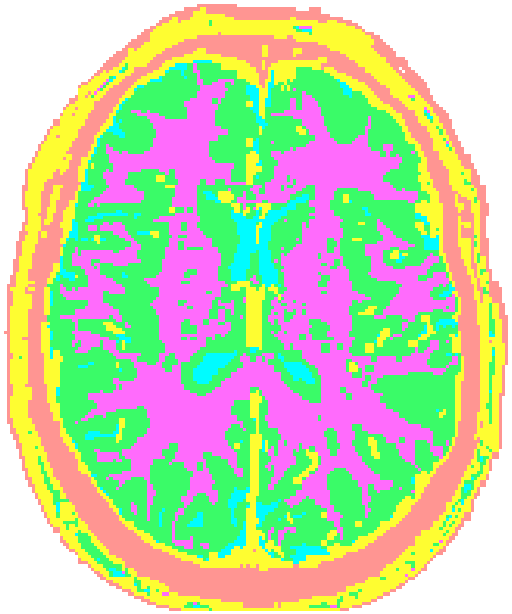


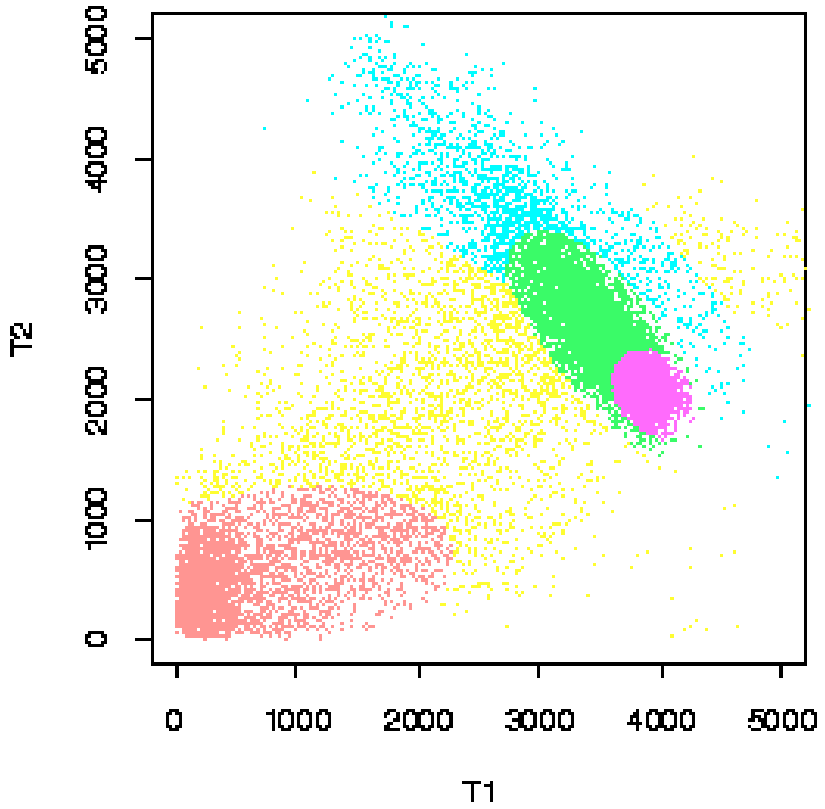
Raw T1 and T2 images.





Second data set: before bias-field correction.



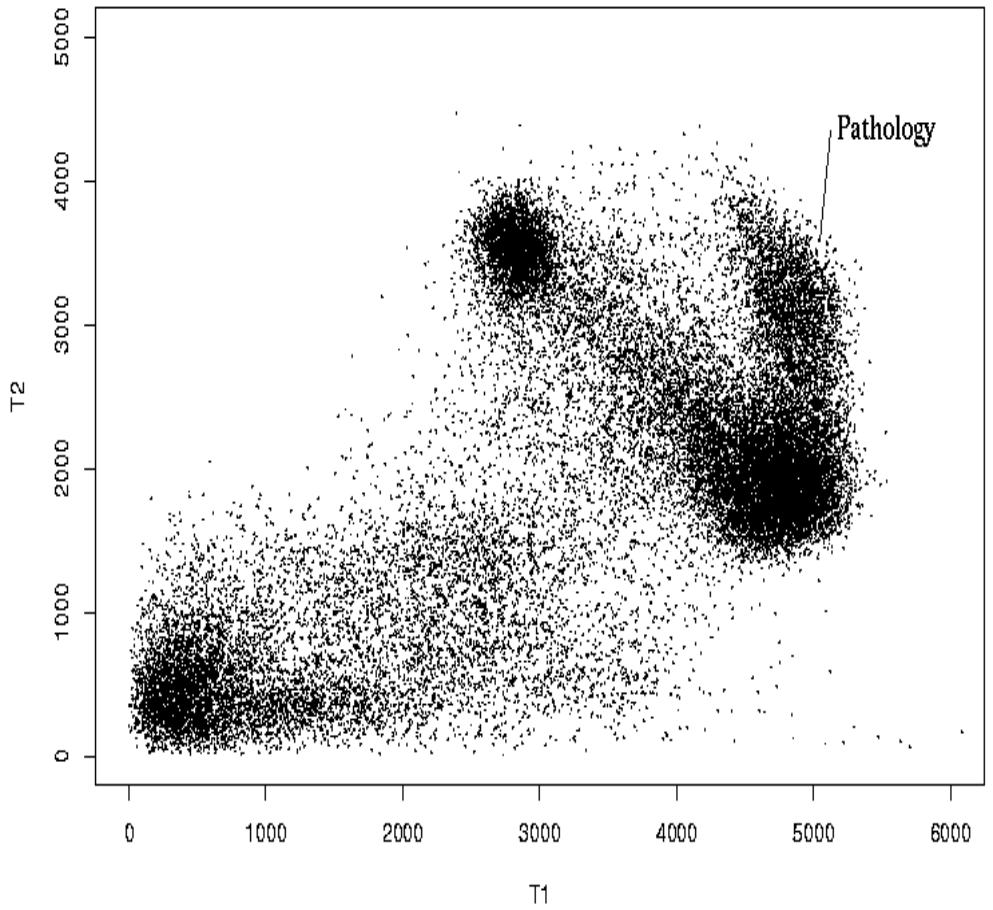


Second data set: after bias-field correction.

Outliers and anomalies

We have found our scheme to be quite robust to variation in imaging conditions and to different ‘normal’ subjects. The ‘background’ class helps greatly in achieving this robustness as it ‘mops up’ the observations which do not agree with the model.

However, outliers can be more extreme:



T1–T2 plot of a brain slice of a brain with a pathology.

This illustrates the dangers of classifying all the points. This is a particularly common mistake when neural networks are used for classification, and we have seen MRI brain scans classified by neural networks where common sense suggested an ‘outlier’ report was the appropriate one.

The procedure presented here almost entirely ignores the spatial nature of the image. For some purposes this would be a severe criticism, as *contextual* classification would be appropriate. However, our interest in these images is not a pretty picture but is indeed in the anomalies, and for that we prefer to stay close to the raw data. The other interest is in producing summary measures that can be compared across time.

Part 2:

Statistics of fMRI Data

SPM

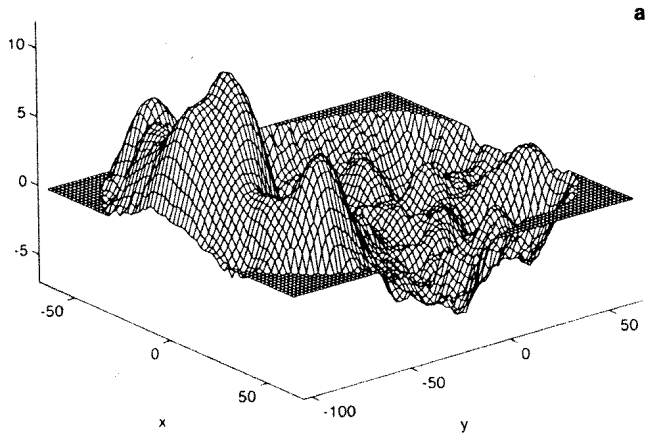
'Statistical Parametric Mapping' is a widely used program and methodology of Friston and co-workers, originating with PET. The idea is to map 't-statistic' images, and to set a threshold for statistical significance.

The t-statistic is in PET of a comparison between states over a number of subjects, voxel by voxel. Thus the numerator is an average over subjects of the difference in response in the two states, and the denominator is an estimate of the standard error of the numerator.

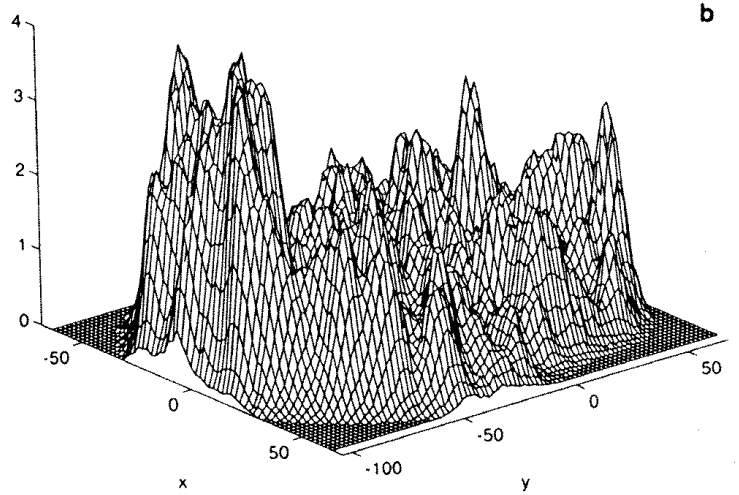
The details differ widely between studies, in particular if a pixel-by-pixel or global estimate of variance is used.

Example PET Statistics Images

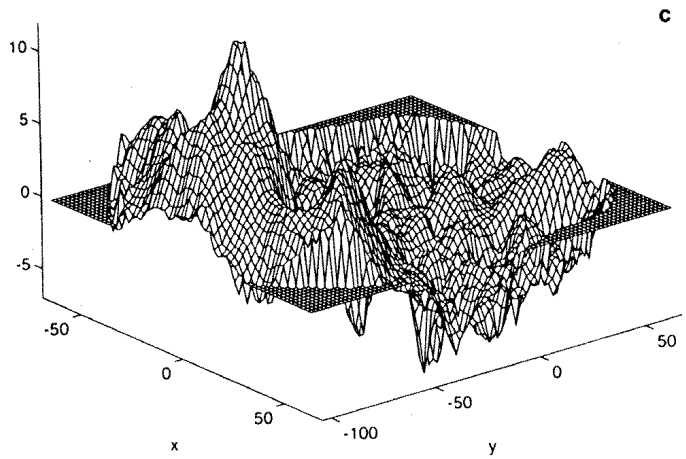
From Holmes *et al* (1996).



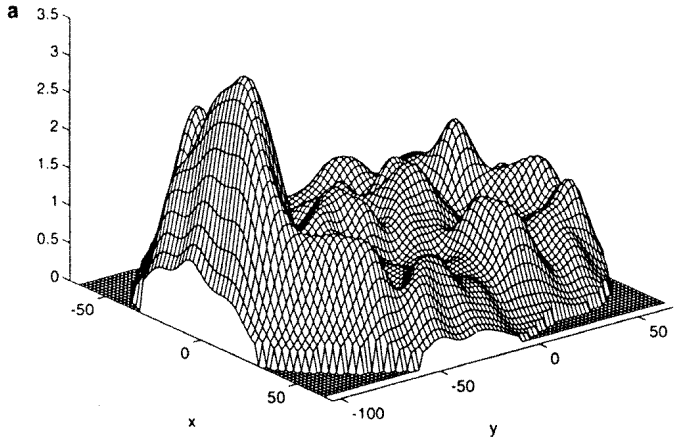
Mean difference image.



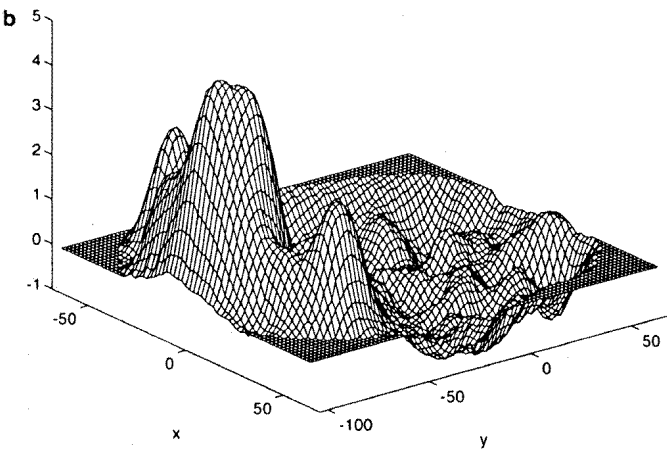
Voxel-wise variance image.



Voxel-wise t–statistic image.



Smoothed variance image.



Resulting t–statistic image.

Multiple comparisons

Finding the voxel(s) with highest SPM values should detect the areas of the brain with most change, but does not say they are significant changes. The t distribution *might* apply at one voxel, but it does not apply to the voxel with the largest response.

Conventional multiple comparison methods (e.g. Bonferroni) will greatly over-compensate as the voxel values are far from independent, so the effective number of observations is far fewer that the number of voxels (which might themselves represent sub-sampling).

Three main approaches:

1. (High) level crossings of Gaussian stochastic processes (Worsley *et al*).
2. Randomization-based analysis (Holmes *et al*).
3. Variability within the time series at a voxel.

Clearly components of variance need to be considered carefully (and have not been).

Issue: do we want the voxels of highest values, or do we want regions of high value?

Euler Characteristics

The Worsley *et al* approach is based on modelling the SPM image X_{ijk} as a Gaussian (later relaxed) stochastic process in continuous space with a Gaussian autocorrelation function (possibly geometrically anisotropic). The autocorrelation function must be estimated from the data, but to some considerable extent is imposed by low-pass filtering.

For such processes there are results (Hasofer, Adler) on the level sets $\{\boldsymbol{x} : X(\boldsymbol{x}) > x_0\}$. These will be made up of components, themselves containing holes. The results are on the expected Euler characteristic (number of sets minus holes) as function of x_0, but for large x_0 there is a negligible probability of a hole, and the number is approximately Poisson distributed. Thus we can choose x_0 such that under the null hypothesis

$$P(X(\boldsymbol{x}) > x_0 \text{ for any } \boldsymbol{x} \in A) \approx 5\%$$

Note that this is based on variability within a single image to address the multiple comparisons point.

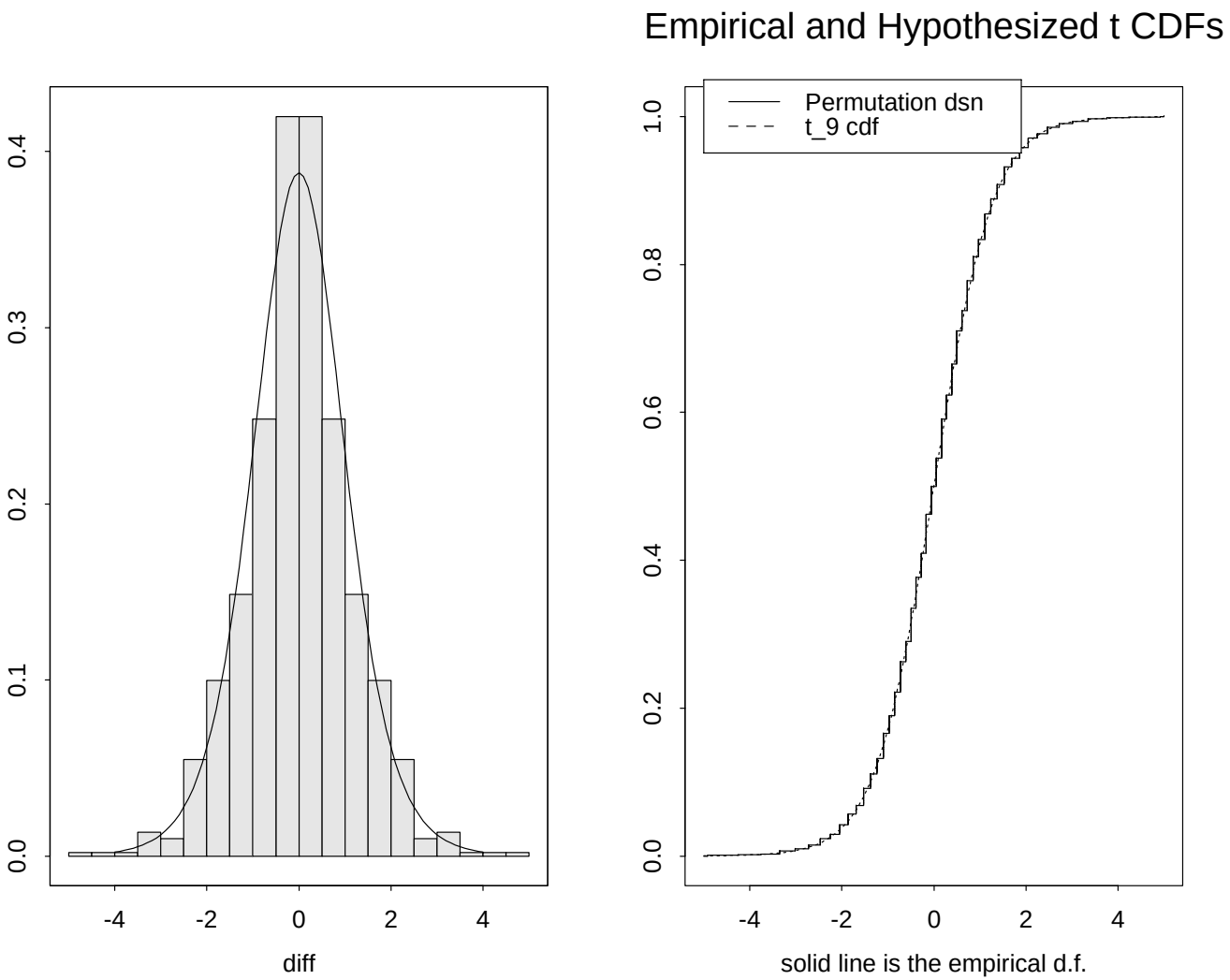
Randomization-based Statistics

Classical statistical inference of designed experiments is based on the uncertainly introduced by the randomization, and not on any natural variability. (Approach associated with Fisher and Yates.)

A typical fPET or fMRI experiment compares two states, say A and B. If there is no difference between the states we can flip the labels within each pair (for each subject in PET, for each repetition $\times$ subject in fMRI). If there are n pairs, there are 2^n possible A–B or B–A labellings. If there is no difference, these all give equally likely values of an observed statistic, so compared observed statistic to the permutation distribution.

Can choose any statistic one can compute fairly easily.

The permutation distribution is often remarkably well approximated by a t distribution. Classic example (Box, Hunter & Hunter, 1977)



Time-Series-based Statistics

The third component of variability is within the time series at each voxel. Suppose there were no difference between A and B. Then we have a stationary autocorrelated time series, and we want to estimate its mean and the standard error of that mean.

This is a well-known problem in the output analysis of (discrete-event) simulations.

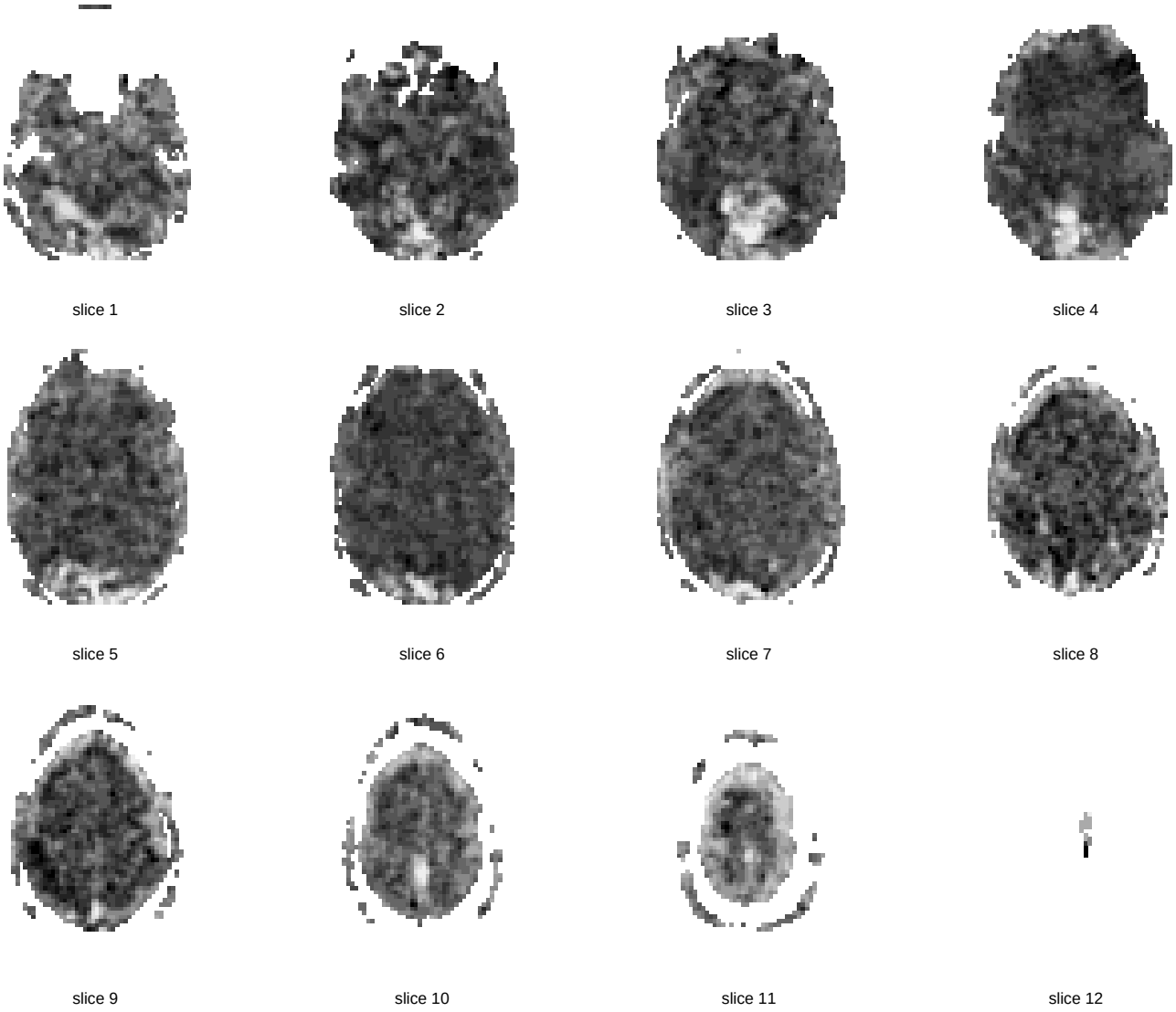
More generally, we want the mean of the A and B phases, and there will be a delayed response (approximately known) giving a cross-over effect. Instead, use a matched filter ($\sin$ wave?) to extract effect, and estimated autocorrelations (like Hannan estimation) or spectral theory to estimate variability. For a $\sin$ wave the theory is particularly easy: the $\log$ absolute value of response has a Gumbel distribution with location depending on the true activation.

fMRI Example

Data on $64 \times 64 \times 14$ grid of voxels. (Illustrations omit top and bottom slices and front and back slices, all of which show considerable activity, probably due to registration effects.)

A series of 100 images at 3 sec intervals: a visual stimulus was applied after 30 secs for 30 secs, and the A–B pattern repeated 5 times.

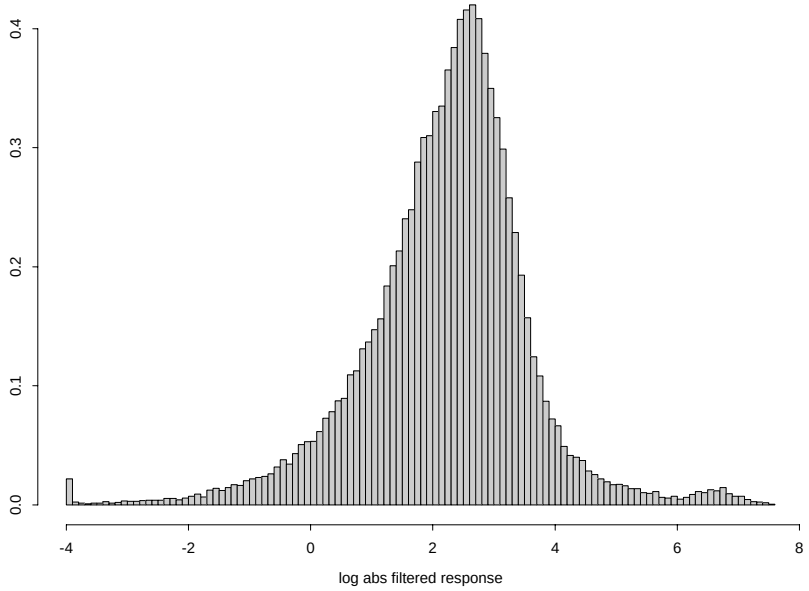
Conventionally the images are filtered in both space and time, both high-pass filtering to remove trends and low-pass filtering to reduce noise (and make the Euler characteristic results valid). The resulting t–statistics images are shown on the next slide. These have variances estimated for each voxel based on the time series at that voxel.



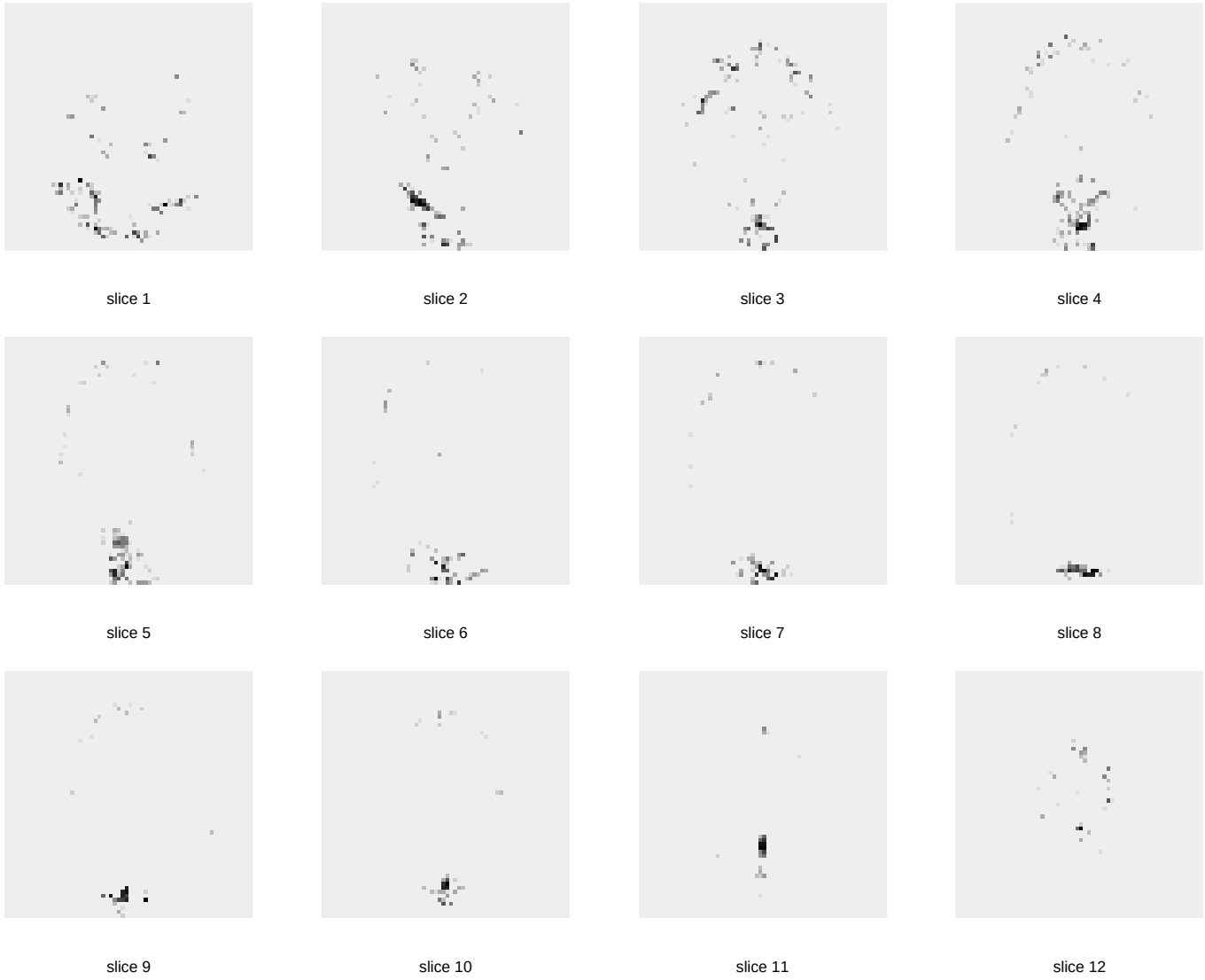
Conventional t–statistic images

Alternative Analyses

We also worked with the raw data, and matched a filter to the expected pattern of response (square wave input, modified by the haemodynamic response). This produced much more extreme deviations from the background variation, and much more compact areas of response.



Histogram of log abs filtered response.



Log abs filtered response, with small values coloured as background.