Some Statistical Contributions

to

Image Analysis and fMRI

Brian D. Ripley

Professor of Applied Statistics
University of Oxford

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

Outline

Part 1: Stochastic Models in Image Analysis

Part 2: Statistical Analysis of fMRI Data

Acknowledgements

Part 1 was suggested by the *Statistical Image Analysis and Processing Study Group* committee. Mainly based on my examples, some with Rafael Molina and Alistair Sutherland.

Part 2 is joint work with 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).

Part 1:

Stochastic Models in Image Analysis

Overview

Statistics has two senses:

- State–istics, the collection and summarization of data.
- Handling uncertainty

Although the first is important in image processing and analysis, we concentrate on the second.

Formalizing Uncertainty

We think of the image formation process in two steps:

- From image description S to true image X.
- From true image X to observed image Z.

We can usually describe these fairly accurately: the first is the image *representation* and the second is determined by the physics of imaging. The observation process will often involve considerable uncertainty, from noise processes or the uncertainty inherent in e.g. Poisson statistics.

We ought then to be able to describe $P(Z \mid S)$ up to a few parameters: many examples will follow. So we assume we know

$$P(Z \mid S; \theta)$$

Representations

- X directly, e.g. in optical astronomy or MRI the true response.
- A pixel-based classification of X, that is each pixel is assigned to one of a small number of classes. Used in remote sensing of, say, land use or forest quality.
- A representation of boundaries of objects (in continuous space).
- A scene description. E.g.
  ‘there are three pedestrians in this image, and the one highlighted is behaving suspiciously’,

  or
  ‘this mammogram contains a 2mm tumour at the coordinates given by the cross-hairs, and some fibrous tissue highlighted in blue’

Over time, interest has moved down this list.

Prior Information

We want to invert the process: having seen Z we want to understand S. This inverse is not well-determined: many aspects of S are not distinguishable from noise. This has lead to ideas of *regularization*, the infamous *Maximum Entropy* principle, and much else.

We prefer to be much less prescriptive. The inverse can be controlled in two ways:

- The mapping from S to X, the representation.
- Asserting prior opinions about S, via probability models.

Note that both express opinions about the space of true images Z, but the first is more rigid (and possibly more fruitful).

Prior Information and Data

Given one or more observed images Z, we obtain a *likelihood* for the image description S and any parameters θ by

$$\ell(S, \theta; Z) \propto P(Z \mid S; \theta)$$

We can combine the prior information and the information from observations by Bayes' formula (lemma, theorem, $\ldots$). So the information we have about the image representation after the observation(s) is

$$P(S \mid Z) \propto P(Z \mid S)P(S) \propto \ell(S, \theta; Z)P(S)$$

Posterior Information

The $P(S \mid Z)$ is known as the *posterior*: we only need to know it up to a normalizing constant as it sums/integrates to one. In general it is a probability distribution over a very large space (possibly a finite discrete space). How do we summarize it?

- Show the most probable S (the MAP estimate)
- Show some typical S
- Show the most probable values of aspects of interest in S (MPM).

Suppose we want to 'restore' a pixel-based image. Then the first option will show one image, the second a number of images which may help us judge our uncertainty in the restoration, and the third might plot an image showing the most likely land use (say) at each pixel. But if we are really interested in the area of wheat in the image, we should aim to find a distribution of that, and summarize it. So our goal may be specific functions $\phi(S)$.

Traps

This process is much misunderstood: indeed there is a whole school of ‘Bayesian’ physicists who fall into most of the following traps.

1. $P(S \mid Z)$ is a density. Unlike MLEs, MAP estimates do not transform: the MAP of $\phi(S)$ is not (usually) $\phi(\widehat{S})$. And often what this is a density with respect to is rather arbitrary (grey-levels on linear, log, gamma-corrected scales)?
2. The mode (which is the MAP estimate) may be a very poor summary of a distribution.
3. Some aspects of the answer will be what you specified in the prior.
4. There are no ‘uninformative’ priors: a prior can be ‘uninformative’ for S but informative for $\phi(S)$.

Remote Sensing of Land Use

We will consider $S = X$ as a grid of pixels with either a finite discrete set of true types (colours).

Simplest case is two colours. For K colours we have the Potts-Strauss model

$$\begin{aligned} &P(X_s = c \mid X_t, t \neq s) \\ &\qquad \propto \exp\ \beta\#\{\text{nhbrs of } s \text{ with colour } c\} \end{aligned}$$

or, as a joint density

$$\begin{aligned} &P(X_1 = x_1, \ldots, X_n = x_n) \\ &\qquad \propto \exp \beta\#\{\text{nhbr pairs of the same colour}\} \end{aligned}$$

More generally, Markov random fields.

[Ideas here mainly from Don and Stu Geman, also Besag, $\ldots$.]

Now suppose we observe $Z_s \sim f_{S_s}(z_s)$ independently from pixel to pixel. Then the posterior distribution is also a Potts-Strauss model;for $K = 2$ we have

$$\log P(S = s \mid Z) = \text{const} + \sum_{i=1}^{n} \lambda_i s_i + \sum_{i,j \text{ nhbrs}} \beta[s_i s_j + (1 - s_i)(1 - s_j)]$$

for $e^{\lambda_i} = f_1(Z_s)/f_0(Z_s)$.

In the case $K = 2$ (only) this can be maximized by network-flow algorithm (Greig, Porteous & Seheult, 1989).

In other cases we use *simulated annealing*, which simulates from

$$P(S \mid Z)^{1/T}$$

for $T \searrow 0$ as the sweep number increases.

Simulation-based inference, for MPM via counting frequencies.

Optical Astronomy

‘Cleaning’ images on a, say, 2048×2048 CCD detector. The response is a count of photons (after a photon multiplier), typically 500 to $20\,000$ counts.

Noise arises from two sources. Thermal noise in the detector has a constant variance over the detector, and in addition there is Poisson noise arising from the discrete nature of photons.

Blurring is caused by the motion of the earth’s atmosphere.

the point-spread function h can be modelled as

$$h(r) = (\beta/\pi R^2)\bigl[1 + (r/R)^2\bigr]^{-\beta}$$

where r is the distance from the source to the receiving pixel and β is about 3. In our studies $R \approx 3.5$, so a point source is effectively spread over a few hundred pixels. The form of h can be checked, since most images contain true point sources (stars).

[With Rafael Molina (Granada) and colleagues.]

Spatially smooth priors

The conditional autoregression (CAR) is defined by a symmetric matrix C such that $I - C$ is positive definite, and has $S \sim N\big(\mu, \kappa(1 - C)^{-1}\big)$ so that

$$E(S_i \mid S_j, j \neq i) = \mu_i + \sum_j C_{ij}(S_j - \mu_j)$$

The matrix C is chosen to reflect spatial proximity. Let N be the neighbour incidence matrix ($N_{ij} = 1$ if i and j are neighbours, 0 otherwise). Then $C = \phi N$ with ϕ just less than $1/r$, r the number of neighbours, will give rise to spatially smooth realization of S.

Applied to $Y = \log(X + p)$, $p \approx 100$. $P(Z \,|\, X = S)$ is given by blurring by the PSF, and then using independent Poisson counts of photons.

Tomography

Tomography is a class of indirect observation techniques: nothing is different in principle, although the model $P(Z \mid X)$ is far more complex.

In all the examples here, the uncertainly come from Poisson-distributed counts. Invasive techniques.

Positron Emission Tomography

Where the event is a pair of photons being detected on a line.

Shepp & Vardi (1982) *IEEE Trans. Med. Imaging* considered maximum likelihood estimation by the EM algorithm. Up to 10 million events.

SPECT

being *single photon emission computerized tomography*. Photons are emitted as a space-time Poisson process. Lower counts, have to deal with absorption in tissue.

The prior information is rather different: medical images tend to have large areas of fairly constant intensity with sharp (and often complex) boundaries. Geman & McClure (1987) introduced a prior that was like a CAR process for small differences between neighbours, but had a much larger penalty for large differences.

$$P(X) = \sum_{s \text{ nhbr of } t} \beta\phi(X_s - X_t)$$

where, for example, $\phi(x) = -1/[1 + (x/\delta)^2]$.

Continuum Models

A radically different approach is to make S a model of the content of the image, and have a complex $S \to X$ mapping.

- HANDS (Chow, Grenander, Keenan, 1989; GCK, 1991)
- Galaxy shapes (Ripley & Sutherland, 1990)
- Nematodes (Ripley, 1992)
- Mitochondria (Grenander & Miller, 1994)
- Faces (Phillips & Smith)
- ‘Snakes’, tracking (Andrew Blake’s group, Gary Jacob, $\ldots$).

All of these have a simple S in continuous space, usually one or more closed contours. The model for S is usually multivariate normal, and certainly simple.

Simulation-based inference again.
Usually of samples from posterior density.

Galaxies

Galaxies are classified by their visual shape into

Elliptical E0–E11 according to the degree of ellipticity.

Spiral with (SB) or without (S) bars, with no arms left (0), and by the tightness of the arms (a, b, or c). Also, if there are arms, by their length (r, rs, or s). Thus an example classification is SBa(rs).

Irregular Really an *outlier* class.

A prior model

We built a prior for a sketch S of a spiral galaxy as

$$\text{disk} \quad + \quad \text{bars} \quad + \quad \text{arms}$$

This sketch is subject to global transformations for

scale uniform over a limited range

location uniform over the image

orientation uniform rotation, reflection with probability 0.5

squeeze as the galaxy is at a uniform angle to the celestial sphere.

Tracking

We now bring time into the analysis. The representation S is a contour, usually represented by the control points of a Bezier curve or a spline. Thus there is a low-dimensional state for the curve, and a prior distribution over that state favouring a smooth curve (but the representation excludes rough curves).

The state evolves according to a simple Gaussian autoregressive process, of order greater than one to allow ‘momentum’ and Kalman filtering is used to find a posterior distribution of the state.

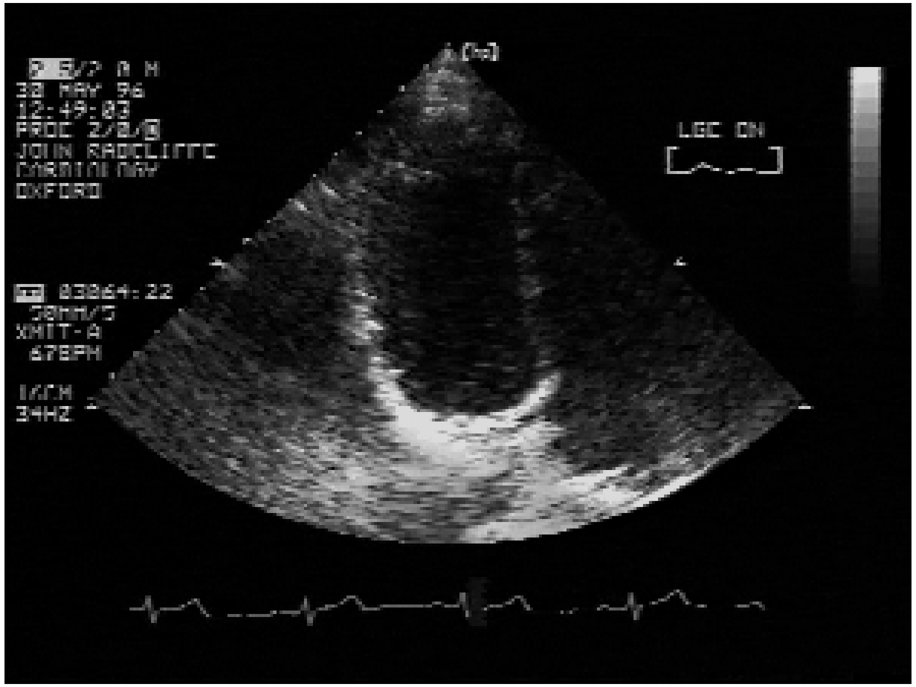
Lots of computational dodges, also MCMC to allow non-Gaussian models, and occasional large excursions (when the tracker loses the object).

Examples (Andrew Blake’s group):

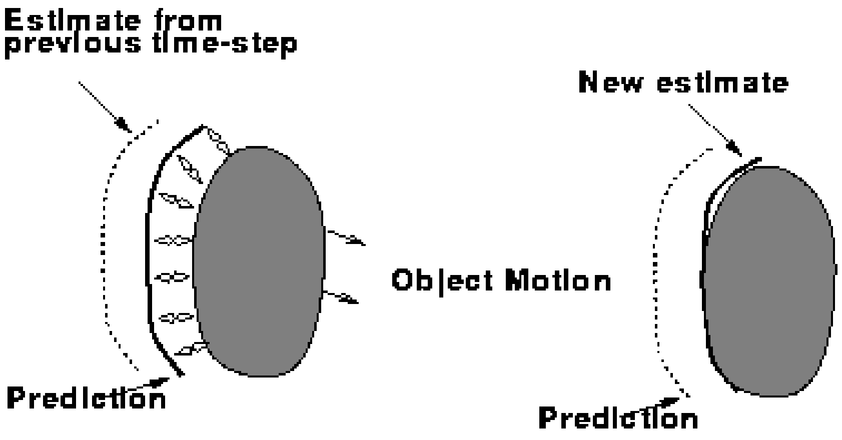
leaves fist cursor marker-free gait tracker

Beating Hearts

Work of Gary Jacob (Medical Imaging / Statistics, Oxford).



A typical echogram taken from the Apical two chamber view showing the left ventricle of the heart.



‘On any given image frame, a dynamical model makes a prediction as to the position of the left ventricular boundary on the next frame. Following the prediction step, measurements are taken by casting normals from the contour to seek desirable image features.The contour estimate is then updated using the Kalman gain to combine the prediction and the ‘best’ image feature measurement.’

Part 2:

Statistics of fMRI Data

‘Functional’ Imaging

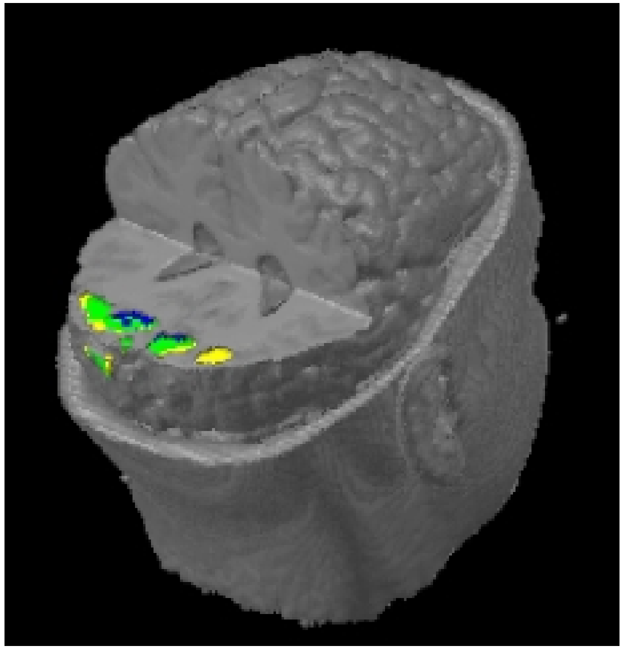
Functional PET and MRI are used for studies of brain function: give a subject a task and see which area(s) of the brain ‘light up’.

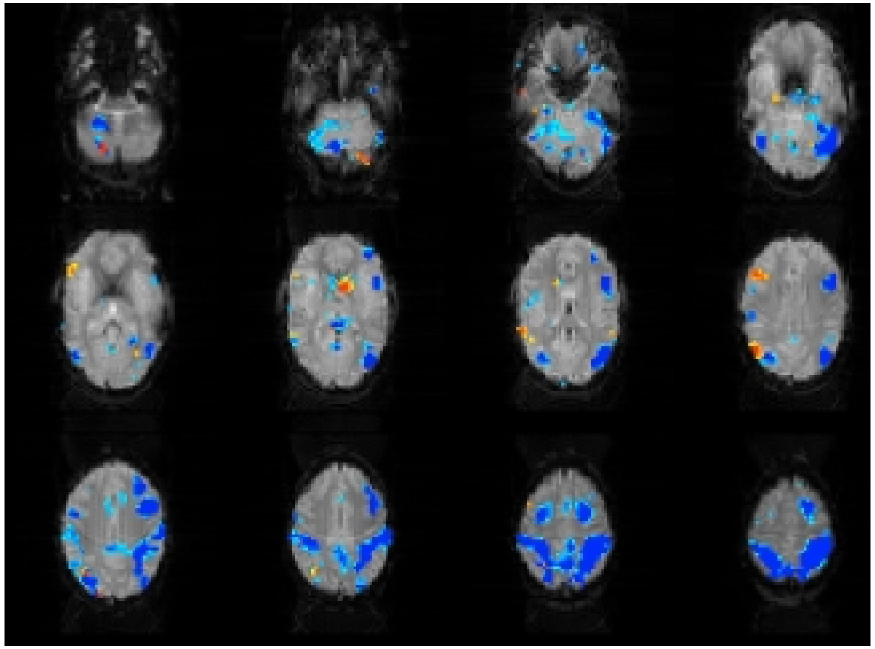
Functional studies were done with PET in the late 1980s and early 1990s, now fMRI is becoming possible (needs powerful magnets—that in Oxford is 3 Tesla). Down to $1 \times 1 \times 3$ mm voxels.

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 most commonly 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.

[Current and future work]

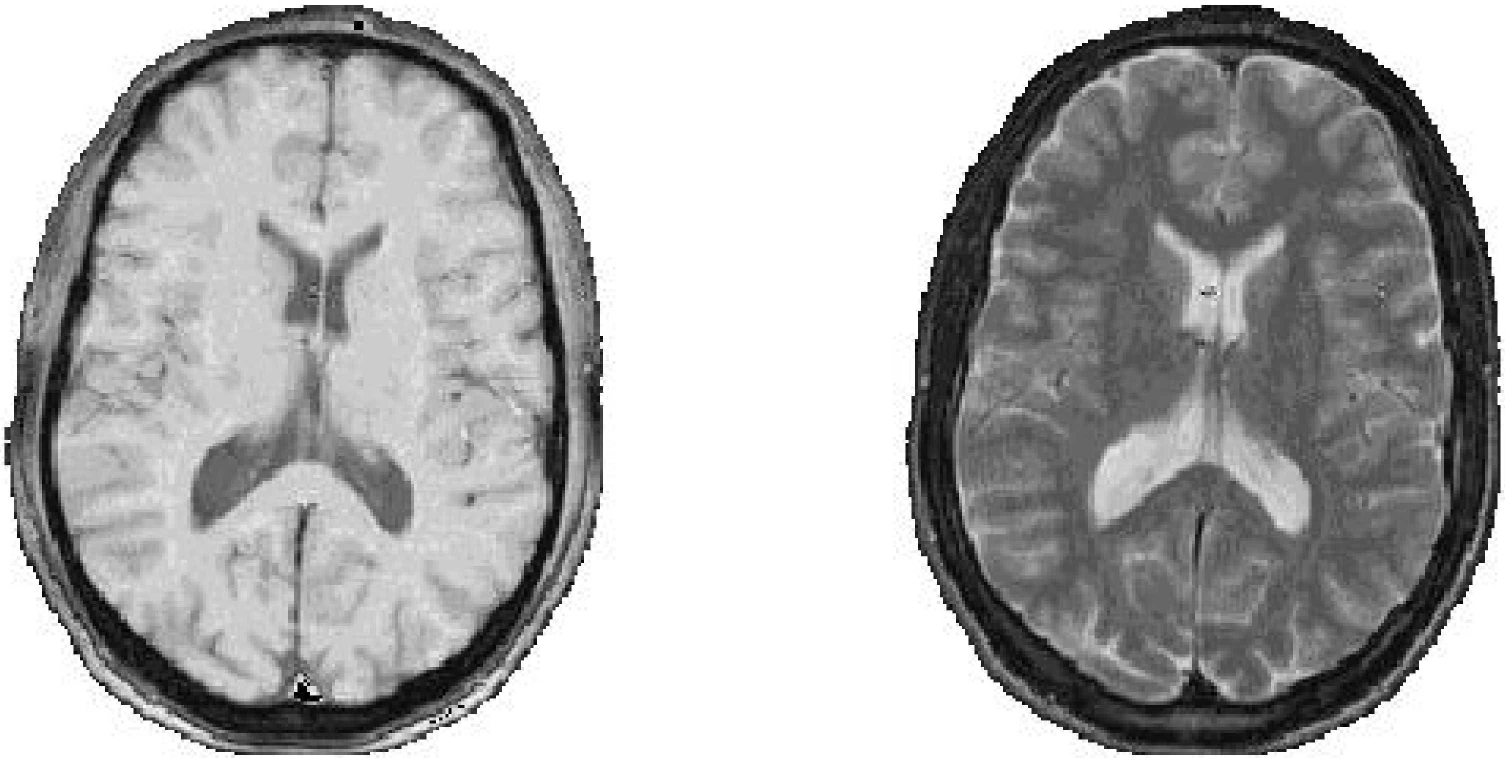
Pilot Study

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. Also useful for confining fMRI activation to just brain tissue (and not, say, eye muscles).

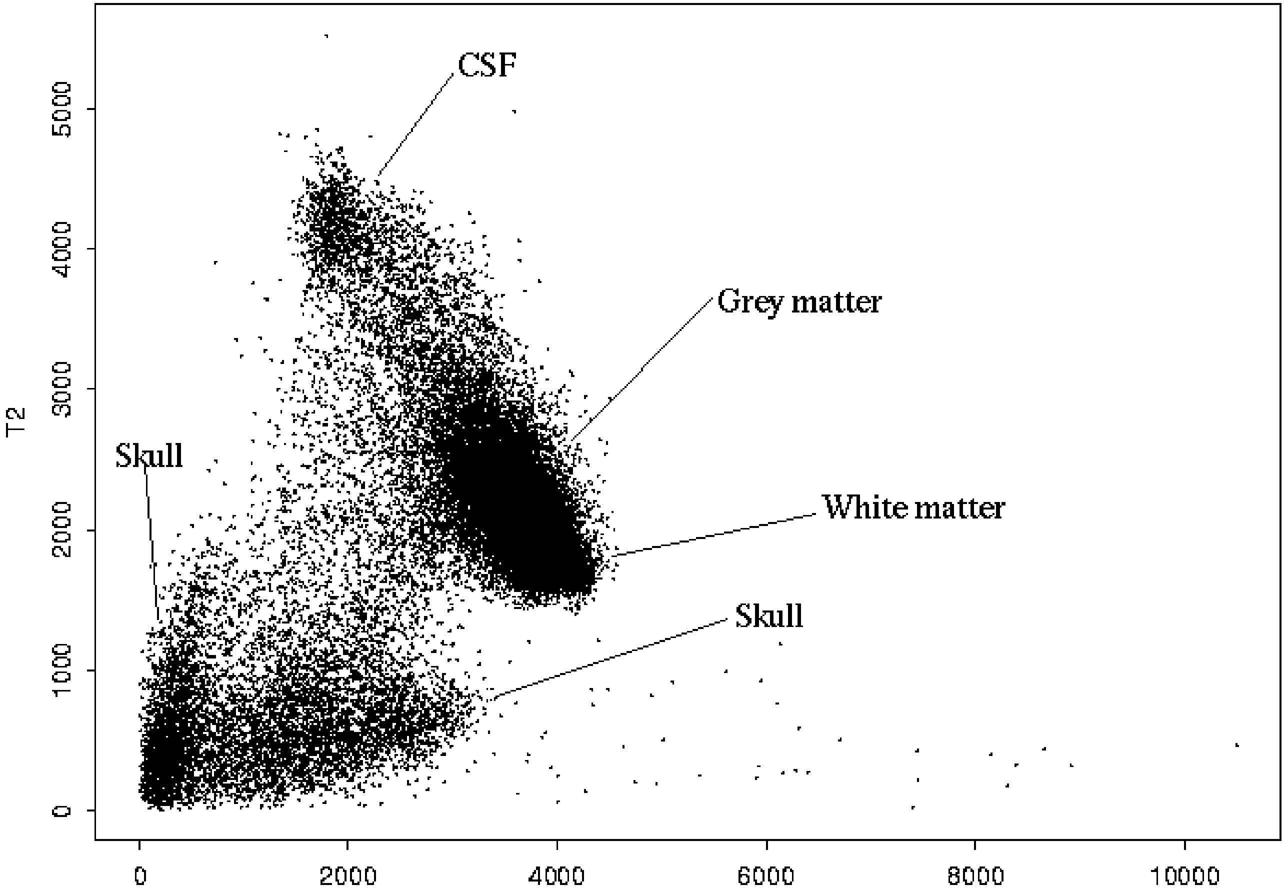
Some Data



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

This slice is of 172×208 pixels.

Imaging resolution was 1 x 1 x 5 mm.



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; mainly caused by inhomogeneity in the magnetic field.

(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.

Modelling the data

Each data point (representing a pixel) consists of one T1 and one T2 value

Observations come from a mixture of sources so we use a finite normal mixture model

$$f(y;\Psi) = \sum_{i=1}^{g} \pi_i \phi(y;\mu_i,\Sigma_i)$$

where the mixing proportions, π_i, are non-negative and sum to one and where $\phi(y;\mu_i,\Sigma_i)$ denotes the multivariate normal p.d.f with mean vector μ and covariance matrix Σ.

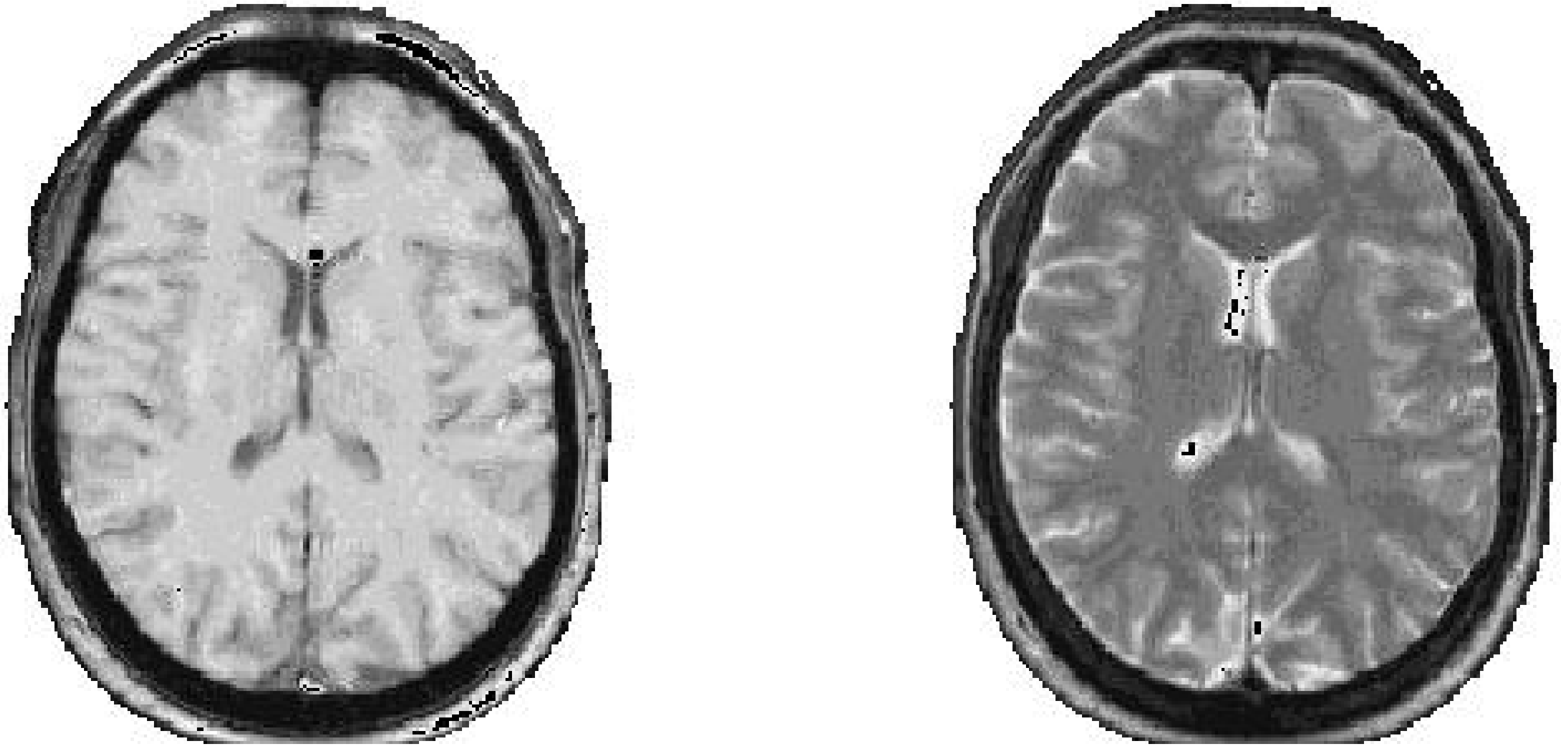
Application/Results

6 component model

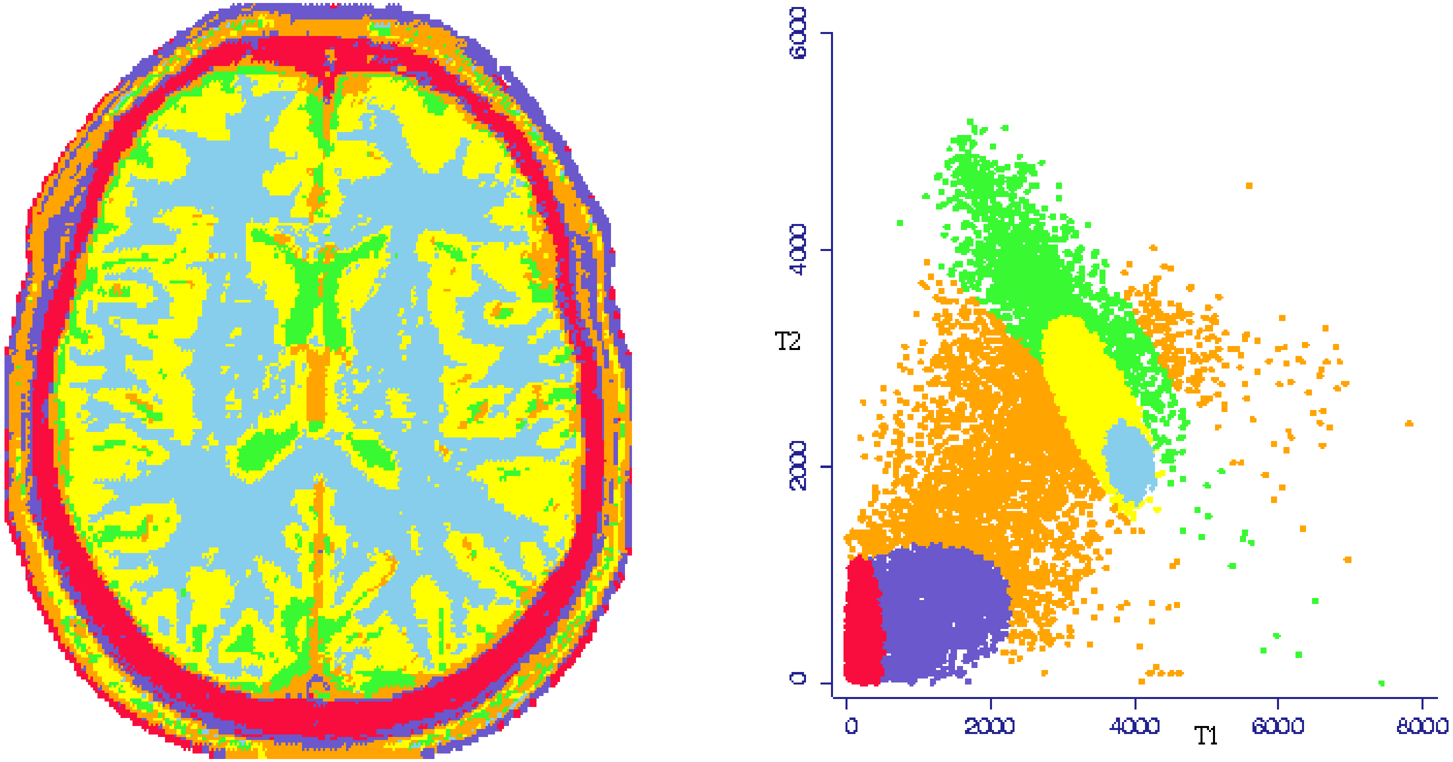
- CSF
- White matter
- Grey matter
- Skull type 1
- Skull type 2
- Outlier component (fixed mean and large variance)

Initial estimates chosen manually from one image and used in the classification of other images.

A Second Dataset



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



Classification image (left) and associated T1/T2 plot (right)

SPM

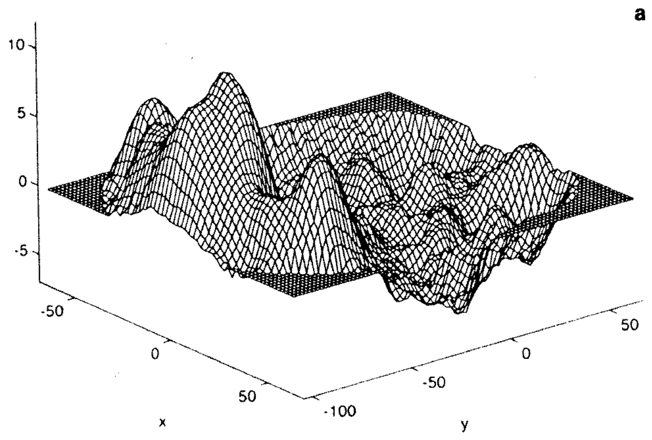
‘Statistical Parametric Mapping’ is a widely used program and methodology of Friston and co-workers, originating with PET. The idea is to map ‘t-statistic’ images, and to set a threshold for statistical significance.

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.

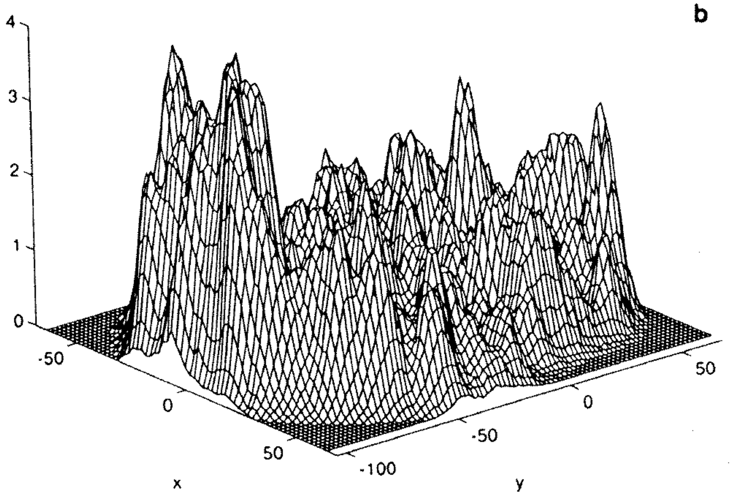
The details differ widely between studies, in particular if a pixel-by-pixel or global estimate of variance is used.

Example PET Statistics Images

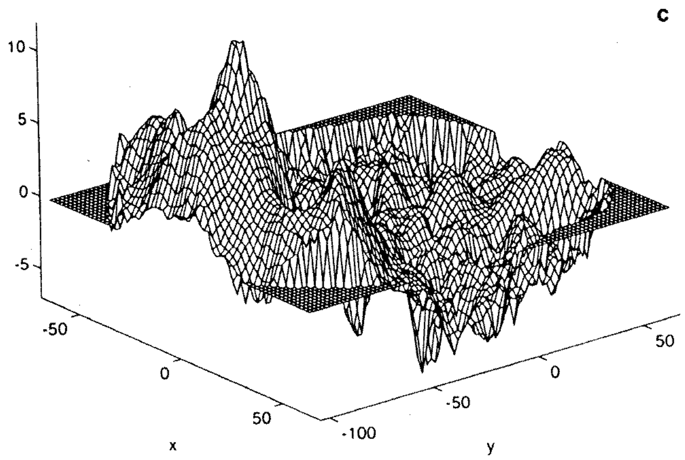
From Holmes *et al* (1996).



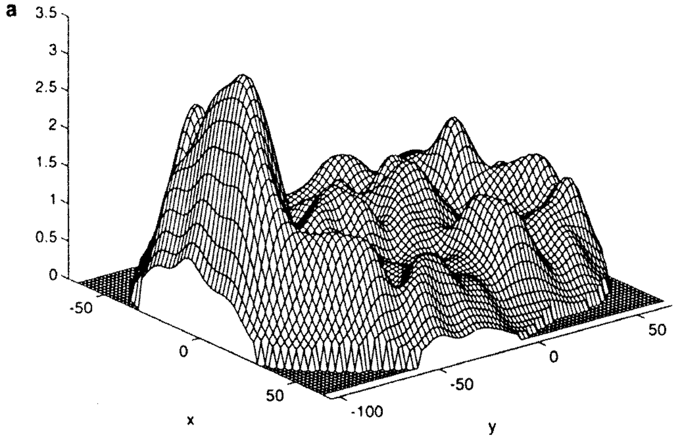
Mean difference image.



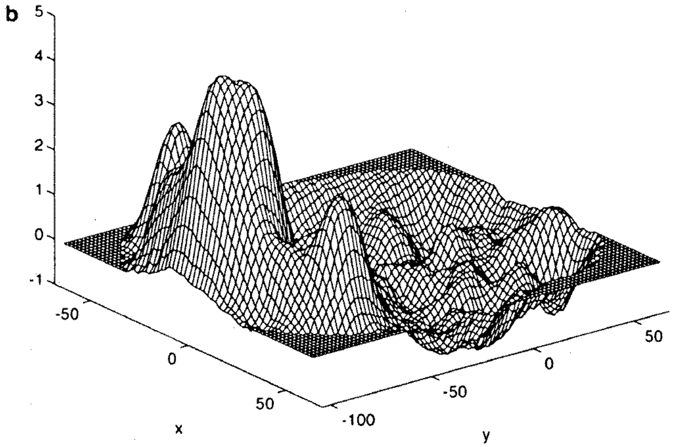
Voxel-wise variance image.



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) may over-compensate if the voxel values are far from independent.

Three main approaches:

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

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.

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 × 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.

Holmes *et al.* actually used a restricted randomization, keep the balance of their 12 pairs into 6 A–B and 6 B–A pairs.

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.

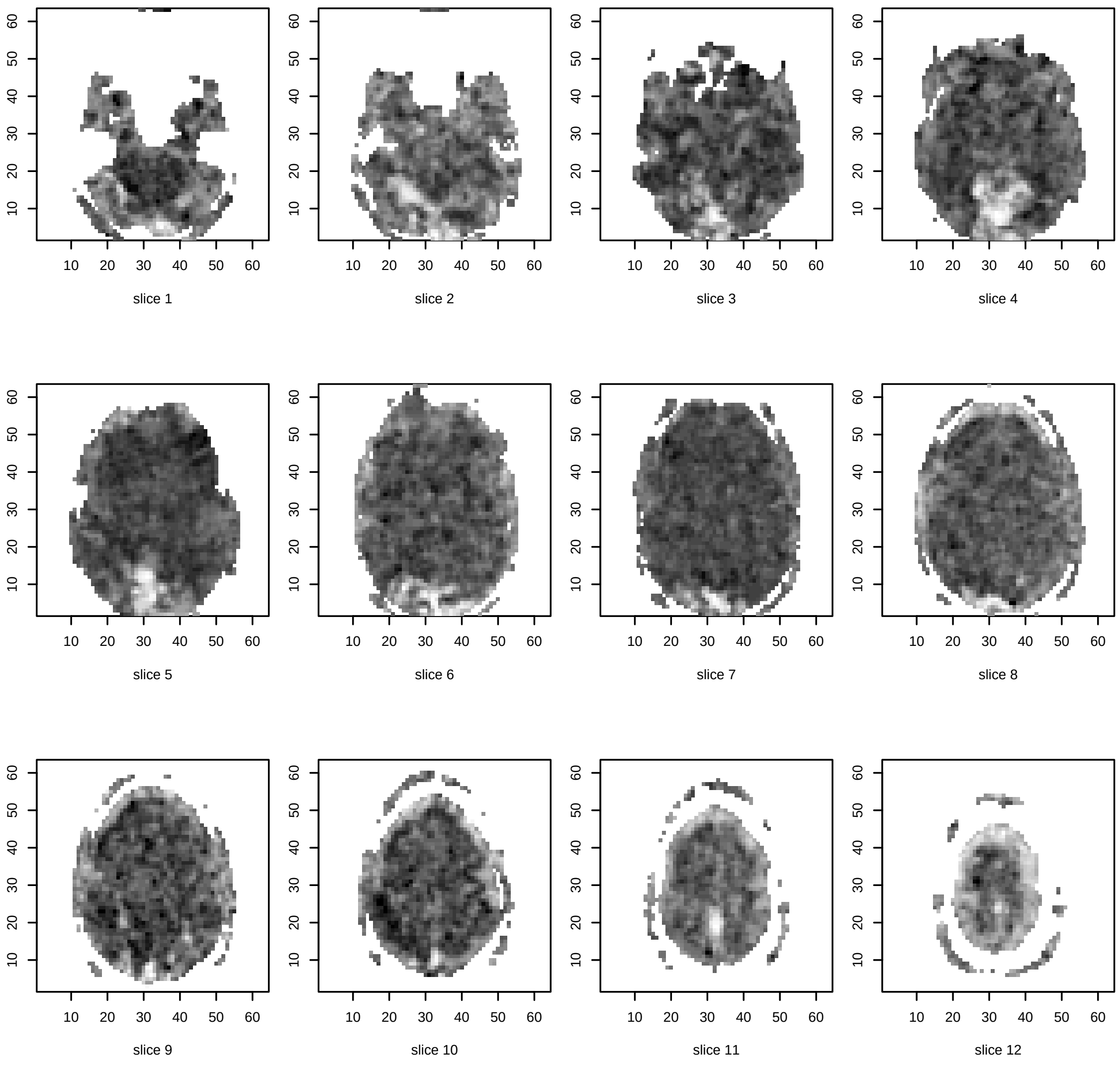
More generally, we want the mean of the A and B phases, and there will be a delayed response (approximately known) giving a cross-over effect. Instead, use a matched filter ($\sin$ wave?) to extract effect, and estimated autocorrelations (like Hannan estimation) or spectral theory to estimate variability. 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 areas outside the brain, all of which show considerable activity, probably due to registration effects.)

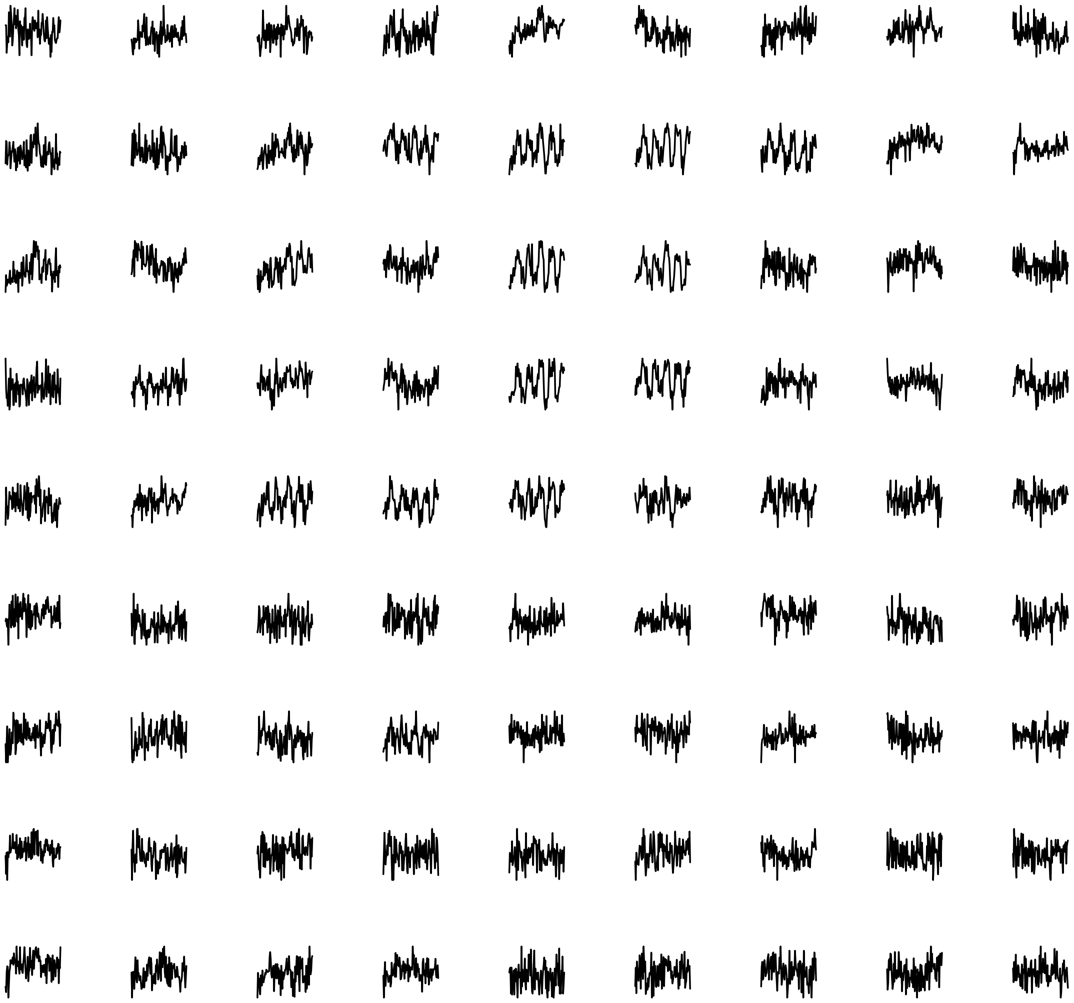
A series of 100 images at 3 sec intervals: a visual stimulus (a striped pattern) was applied after 30 secs for 30 secs, and the A–B pattern repeated 5 times. In addition, an auditory stimulus was applied with 39 sec ‘bursts’.

Conventionally the images are filtered in both space and time, both high-pass time filtering to remove trends and low-pass spatial 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 – for visual stimulus

A Closer Look at some Data

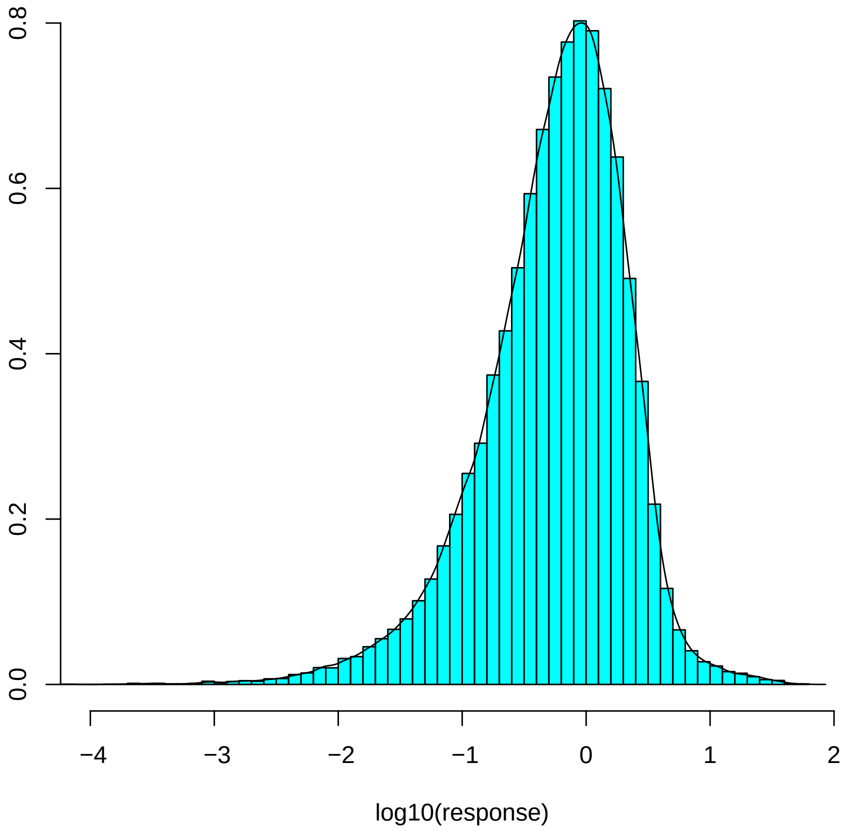


A 10×10 grid in an area of slice 4 containing activation.

Alternative Analyses

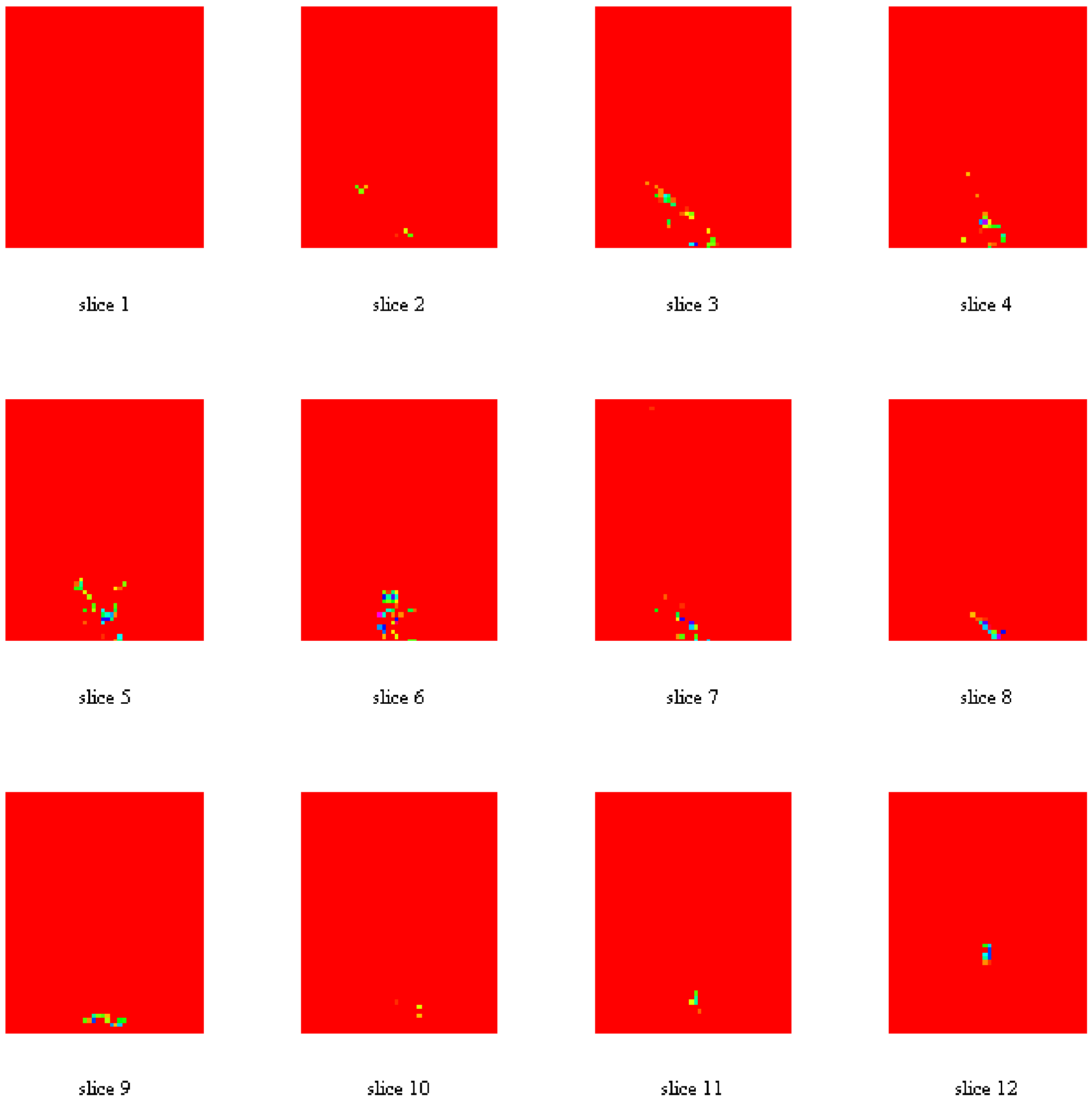
- Work with raw data.
- Non-parametric robust de-trending, Winsorizing if required.
- Work in spectral domain.
- Match a filter to the expected pattern of response (square wave input, modified by the haemodynamic response).
- Non-parametric smooth estimation of the noise spectrum at a voxel, locally smoothed across voxels.
- Response normalized by the noise variance should be Gumbel (with known parameters) on log scale.

This produced much more extreme deviations from the background variation, and much more compact areas of response. 30–100 minutes for a brain (in S / R on ca 400MHz PC).



Histogram of log filtered response, for an image with activation.

We can validate the distribution theory by looking at frequencies without stimulus, and ‘null’ images.

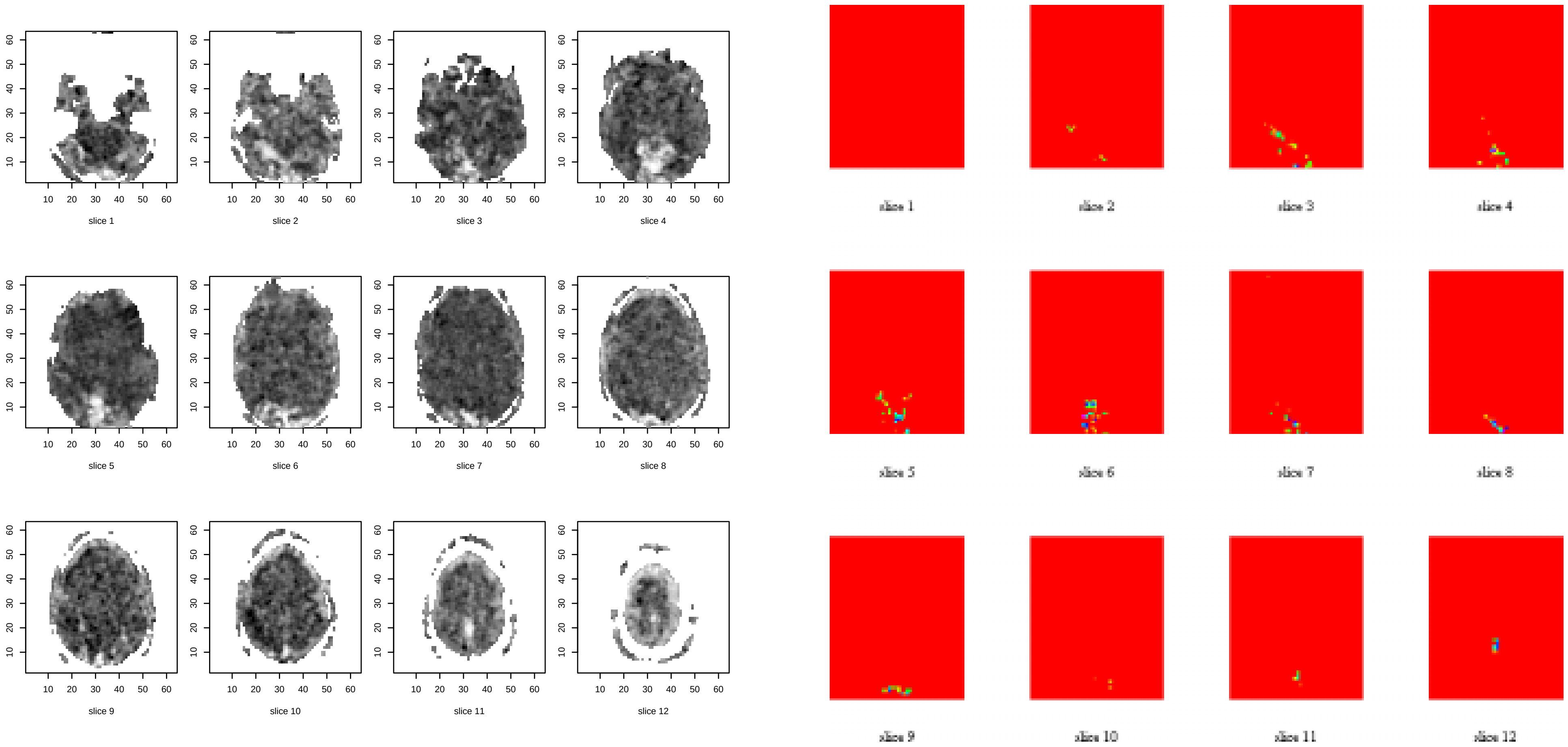


slice 1 slice 2 slice 3 slice 4

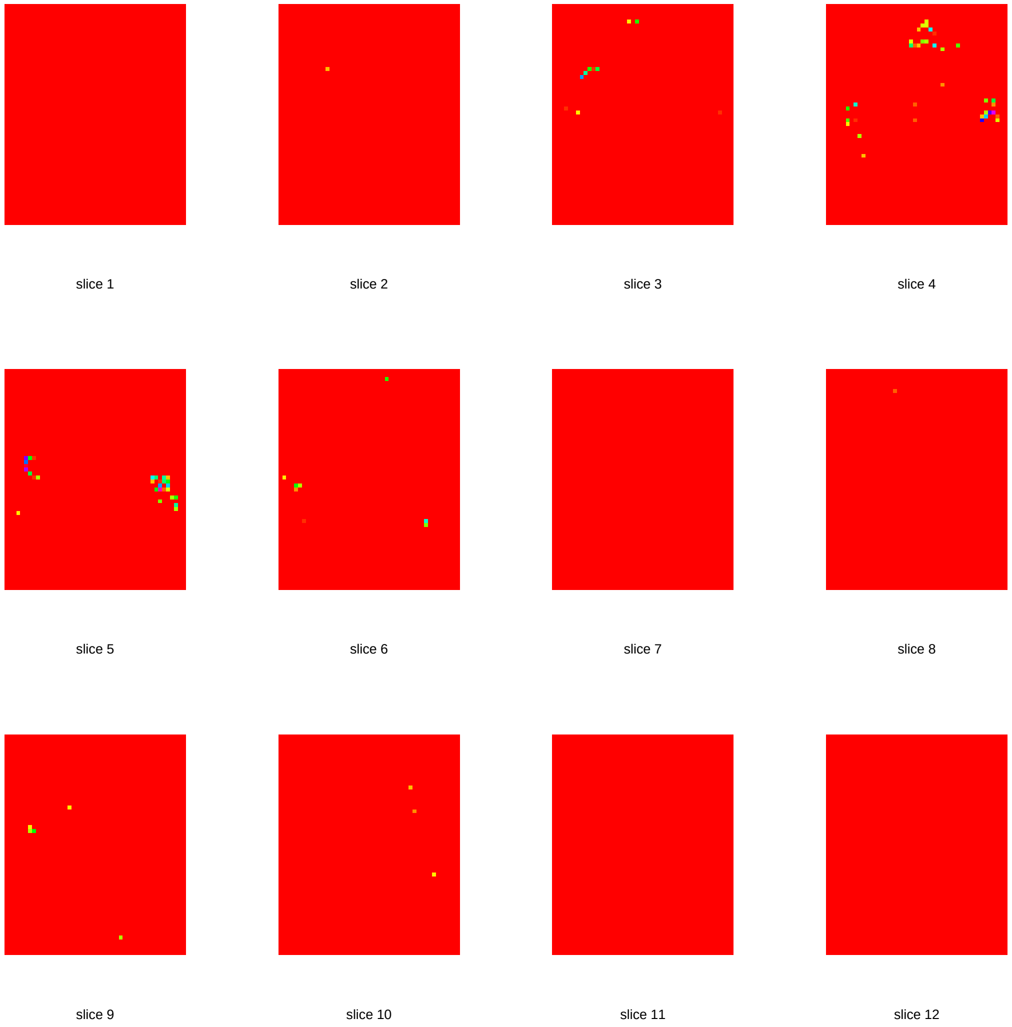
slice 5 slice 6 slice 7 slice 8

slice 9 slice 10 slice 11 slice 12

Log abs filtered response, with small values coloured as background (red). Threshold for display is $p < 10^{-5}$ (and there are ca 20,000 voxels inside the brain here).



Comparison of *t*-statistics and our analysis.



slice 1

slice 2

slice 3

slice 4

slice 5

slice 6

slice 7

slice 8

slice 9

slice 10

slice 11

slice 12

The auditory sub-experiment.

A Related Approach

Lange & Zeger (1997) *Applied Statistics* has a related approach, in which they model the haemodynamic response by gamma probability density functions, and fit a regression estimating the autocorrelation structure of the time series (in Fourier domain).

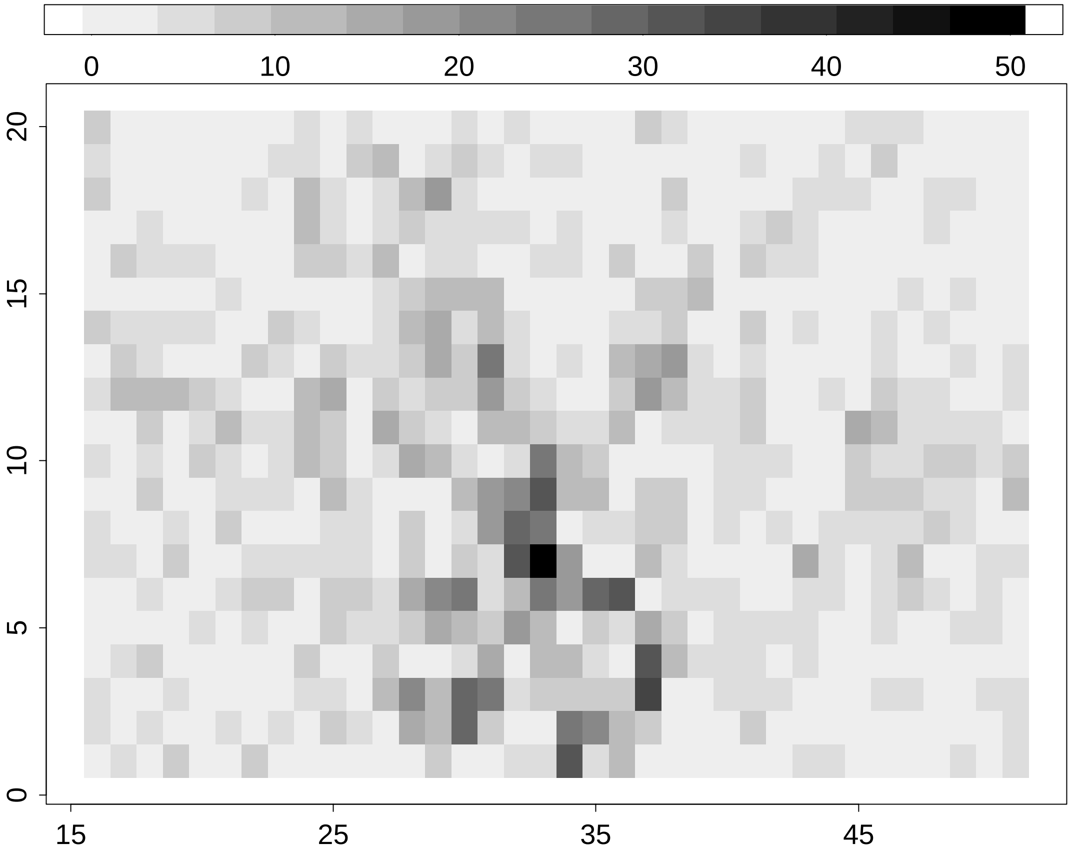
Thus their model is that the response is a modified square wave of the form

$$\beta\Lambda(t;\theta) = \beta \int_0^t \lambda(t-u;\theta)\,\mathrm{d}u, \qquad 0 \le t \le 30$$

continued in the obvious way. The parameters β and θ are then fitted by generalized least squares at each voxel, with a general stationary covariance function for the errors (estimated locally in space).

It is hard to estimate the haemodynamic response where there is little neurological activity: for local estimation (as in Lange–Zeger) this might be at a handful of voxels.

In our implementation of the Lange–Zeger procedure we find (2 hours)



Estimates of β from a region in the bottom centre of slice 4.

Conclusions

- Look at your data (even if it is on this scale: 2 million points per experiment).
- Data ‘cleaning’ is vital for routine use of such procedures.
- Fit your theory to your analysis, not *vice versa*.
- It is amazing what can be done in high-level languages on cheap computers.