Mining Information from Brain Images

Brian D. Ripley

Professor of Applied Statistics
University of Oxford

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

Outline

Part 1: Characterizing Alzheimer's Disease

Joint work with Kevin Bradley, Radiologist at OPTIMA (Oxford Project to Investigate Memory and Ageing).

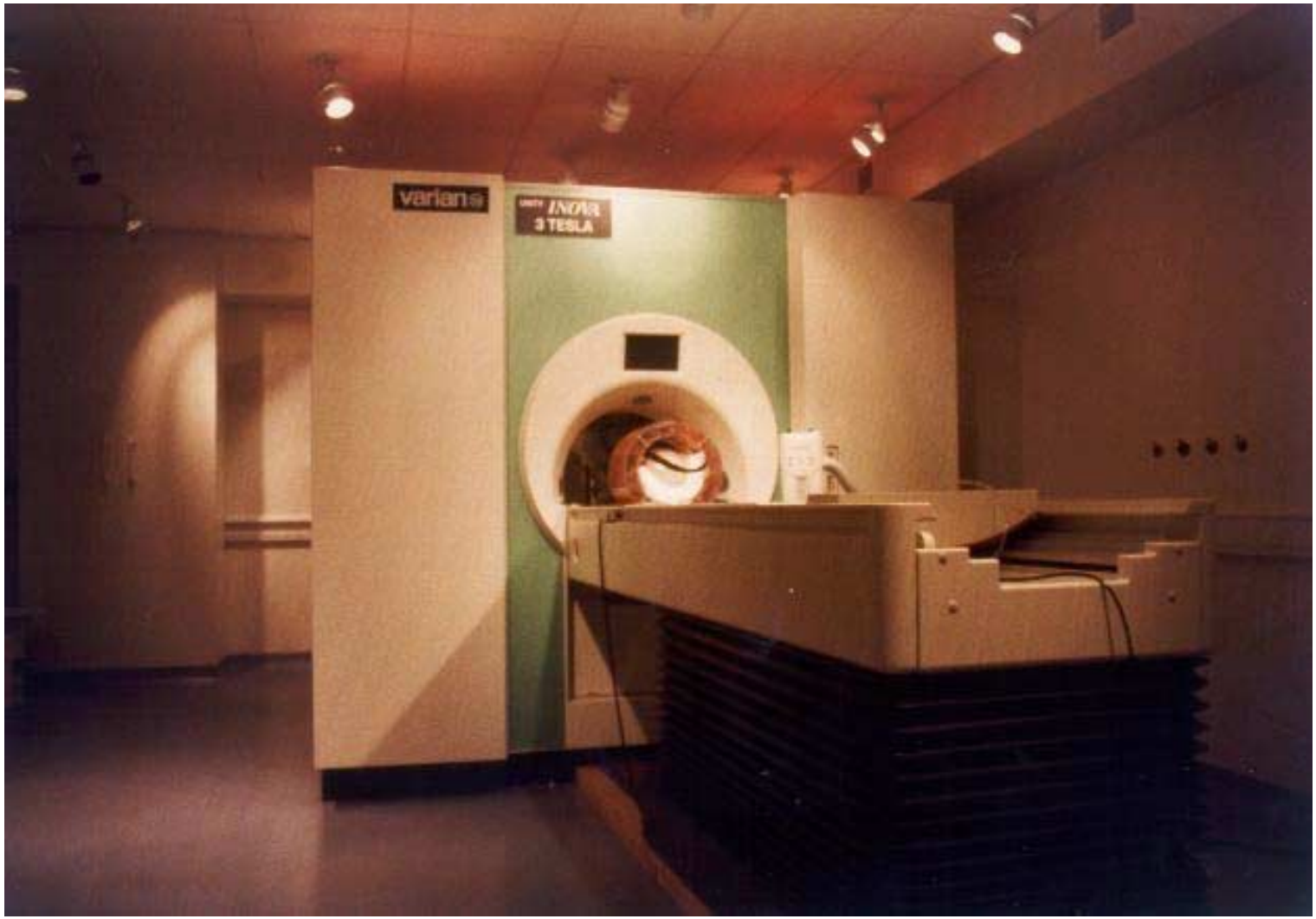
Part 2: Magnetic Resonance Imaging of Brain Structure

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, Oxford)

Part 3: Statistical Analysis of Functional MRI Data

Joint work with Jonathan Marchini.
Data, background and advice provided by Stephen Smith (Oxford Centre for Functional Magnetic Resonance Imaging of the Brain) and Nick Lange (McLean Hospital, Harvard).

Magnetic Resonance Imaging



Magnetic resonance imaging is a non-invasive way of examining a living brain as it functions. We will only consider human brains, but work is also done on rat brains, which are smaller and can be given more interference.

In MRI can trade temporal, spatial and spectral resolution. Data is acquired in Fourier domain over a period, down to around 3 seconds. The subject is immersed in a strong magnetic field (needs powerful magnets—that in Oxford is 3 Tesla) and responses to input signals allow information to be collected simultaneously spatially. The sorts of information are

- spin decay rates of hydrogen atoms, longitudinally and transversally (T1 and T2).
- BOLD, measurements of presence of oxygenated blood.
- perfusion of blood.
- presence of specific chemicals.

Collect one or more over ca 45 minutes in the scanner.

Data rates

A typical experiment will collect information on up to $256 \times 256 \times 30$ voxels. This could be collected at up to 200 time points, and on several quantities.

Typically one experiment yields 1–100 million observations. There are be a few hundred sessions (a few on up to 100 subjects) in the course of a medical/psychological/drug study.

Data collection is expensive, especially on research machines.

Computation

Everything you see (except the movies) was computed in S on a modest machine, typically a 300–500 Mhz chip with 128–512Mb of RAM.

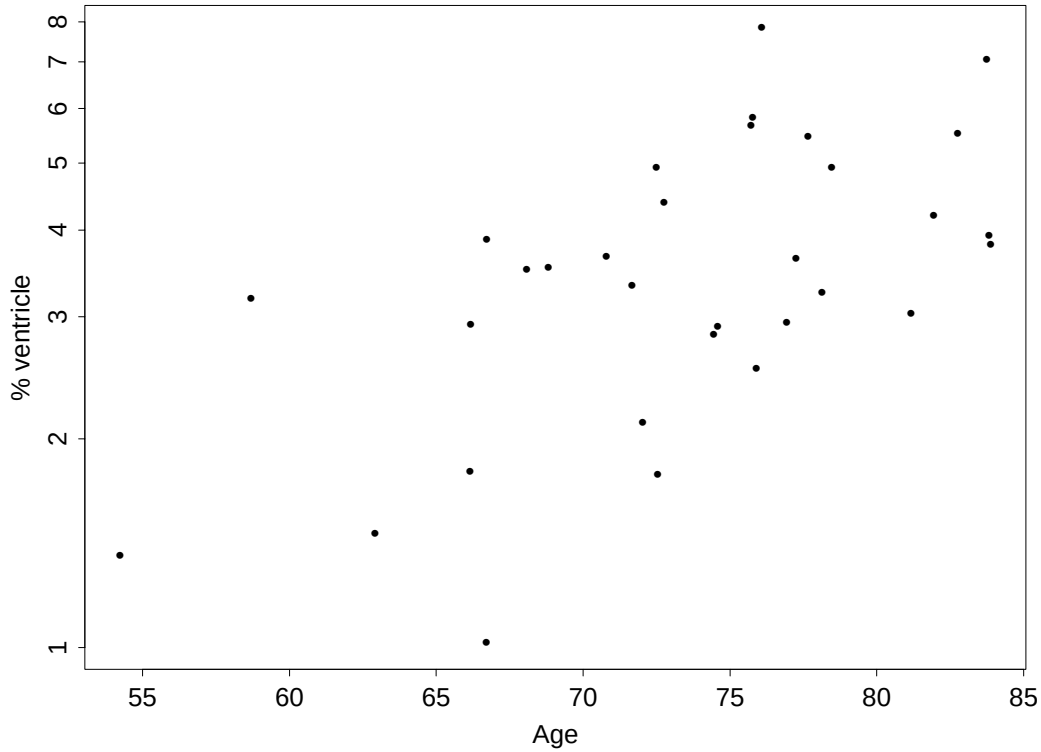
Part 1:

Characterizing Alzheimer's Disease *via* serial structural MRI

Structural MRI of Ageing and Dementia

Everyone's brain shrinks with age (0.4% per year), and not uniformly.

Disease processes, for example Alzheimer's Disease (AD) change both the overall rate and the differences in rates in different parts of the brain.



Use serial structural MRI, probably of two measurements n months apart.

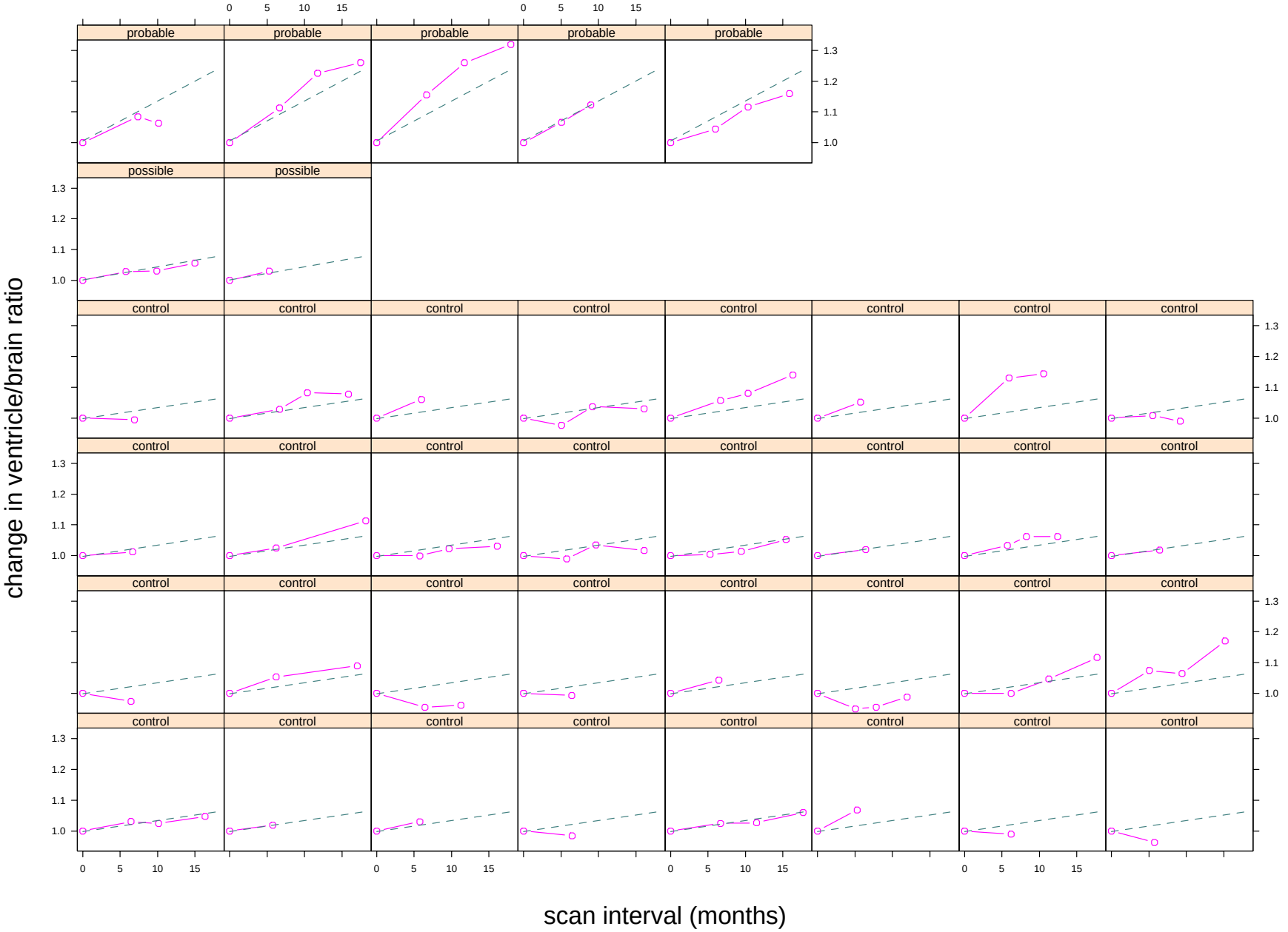
How large should n be?

How many patients are needed? (Parallel study by Fox *et al*, 2000, *Archives of Neurology*.)

Study with 39 subjects, most imaged 3 or 4 times over up to 15 months.

Three groups, ‘normal’ (32), ‘possible’ (2) and ‘probable (5).

Given the ages, expect a substantial fraction of ‘normals’ to have pre-clinical AD.


probable
possible
control
change in ventricle/brain ratio
scan interval (months)
0
5
10
15
1.0
1.1
1.2
1.3

Statistical Analysis

Major source of variation is between subjects. Not many ‘abnormals’, and usually the diseased group is more variable than the normals.

Choose to use linear mixed-effects models (NLME of Pinheiro & Bates). Relative size of the random effects answers the questions.

How not to do it

Fox *et al* has 18 normals, 18 AD, 9 of each sex in each group. They used the elementary sample-size formulae for detecting differences between two arms of the trial.

Hypothesis was that a drug would give a 20% reduction in the excess overall brain shrinkage in AD patients. Concluded that 168 subjects were needed in each arm of the trial.

That’s the two-sided formula! What is the variability in the treatment group (pilot size 0)?

Part 2:

Magnetic Resonance Imaging of Brain Structure

Neurological Change

Interest 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 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.

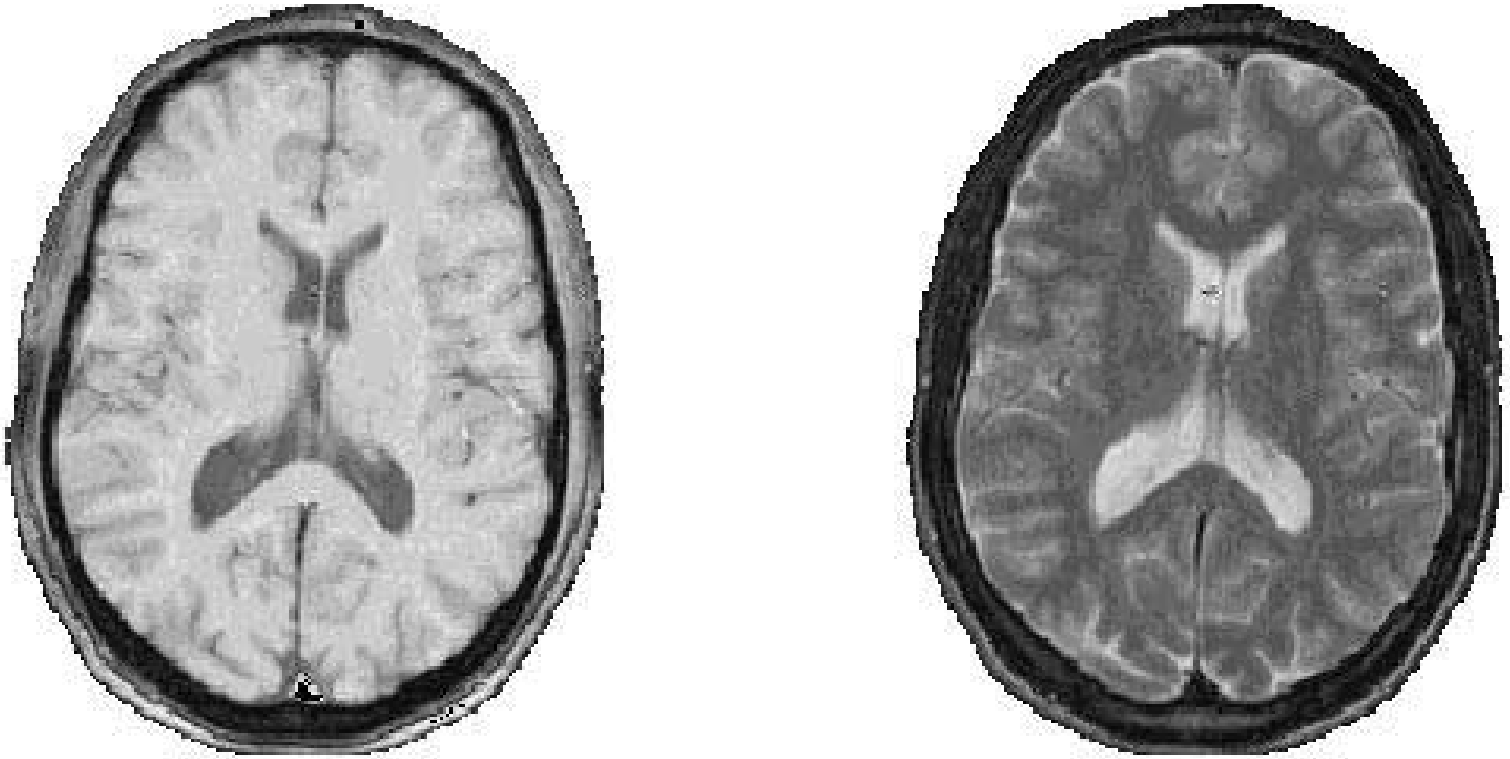
Pilot Study

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’.

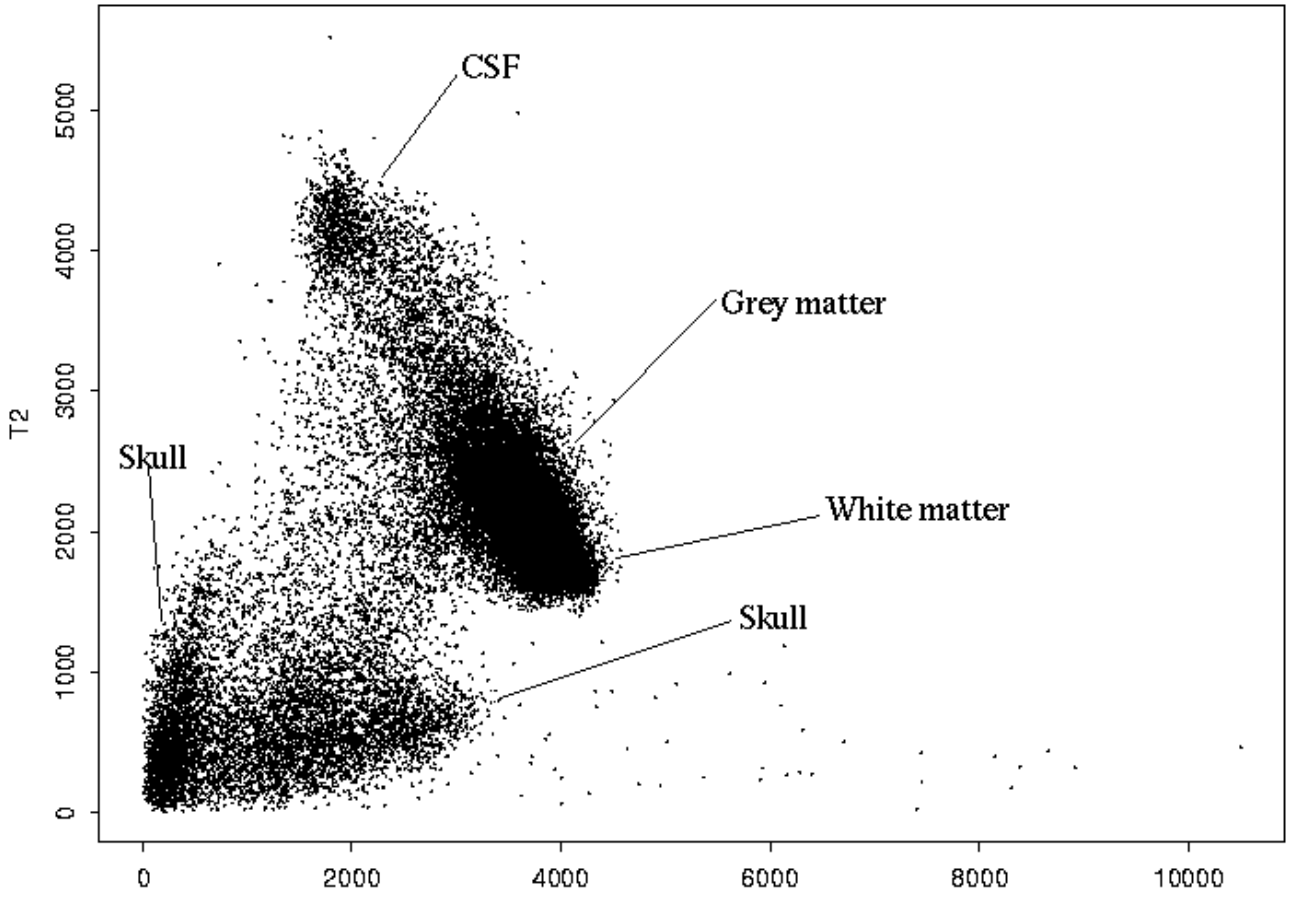
Initial aim is reliable automatic segmentation. To be applied to a set of patients recovering from severe head injuries.

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.

Don't believe what you are told: almost everything we were told about image imperfections from the physics was clearly contradicted by the data.

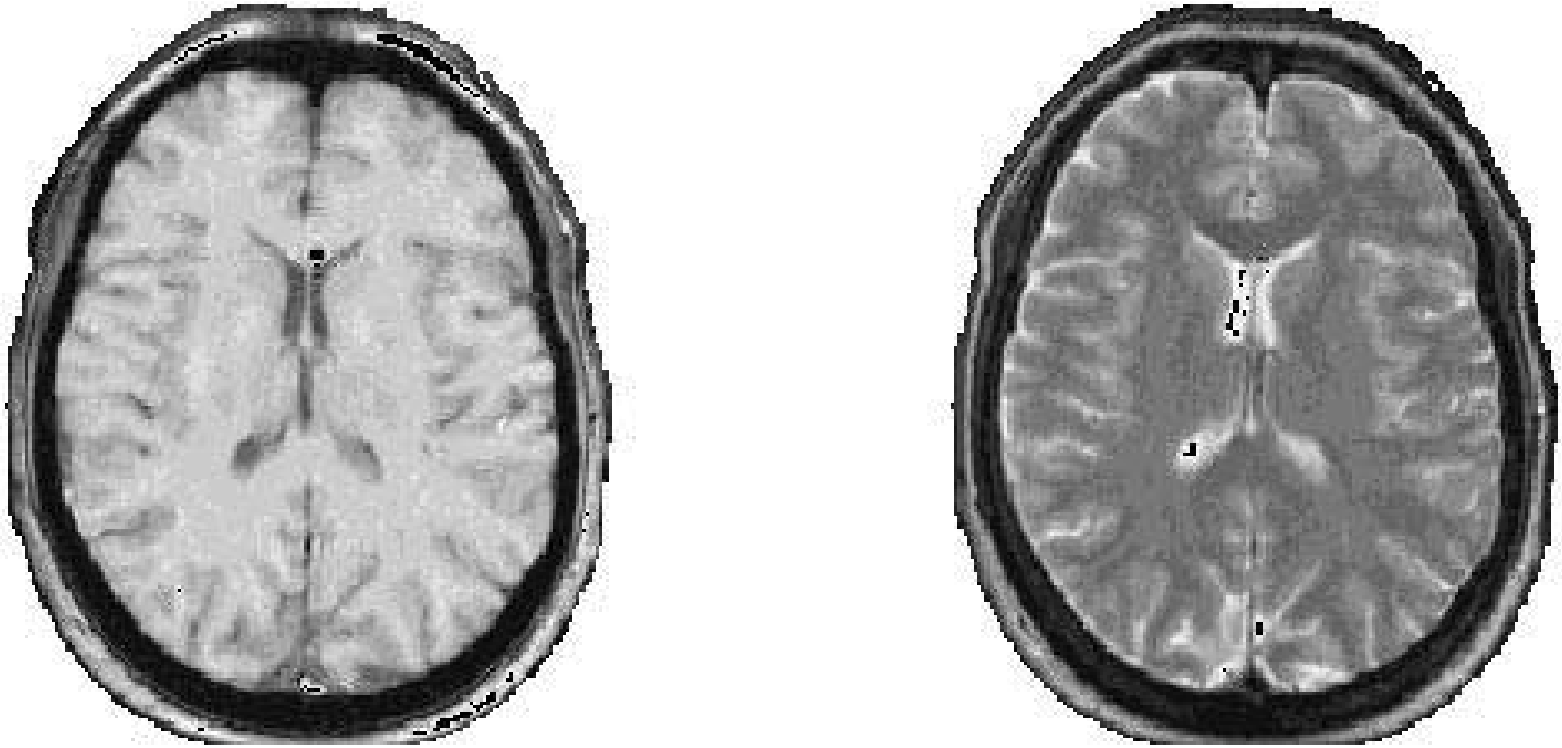
Application/Results

6 component model

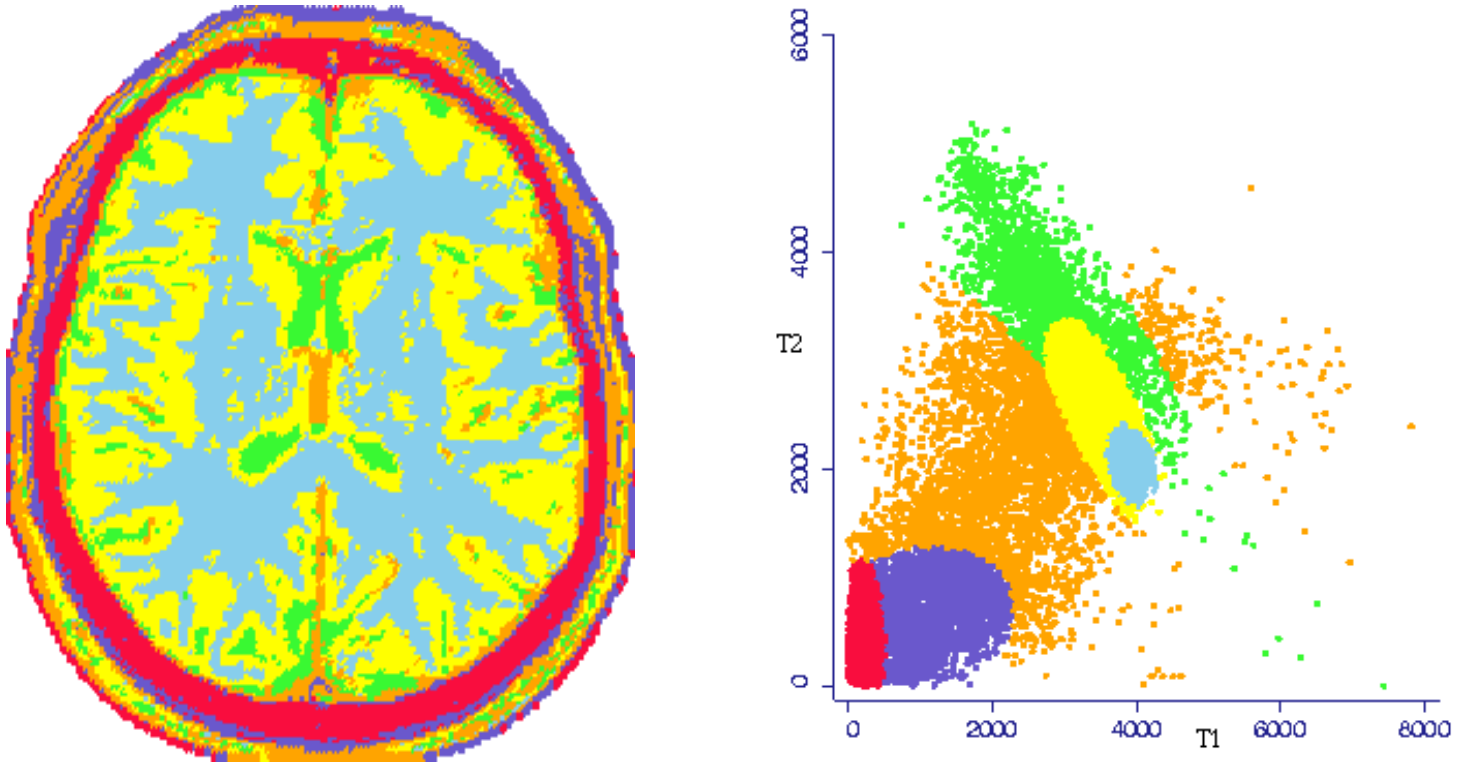
- CSF
- White matter
- Grey matter
- Skull type 1
- Skull type 2
- Outlier component (fixed mean and large variance)

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), training the 6-component mixture model from its fit on the reference subject.

Part 2:

Statistics of Functional MRI Data

‘Functional’ Imaging

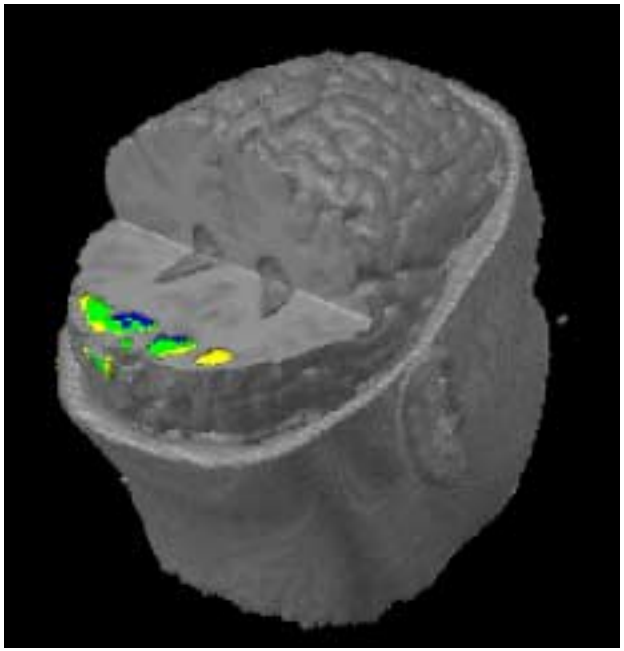
Functional PET (positron emission spectroscopy: needs a cyclotron) and MRI are used for studies of brain function: give a subject a task and see which area(s) of the brain ‘light up’.

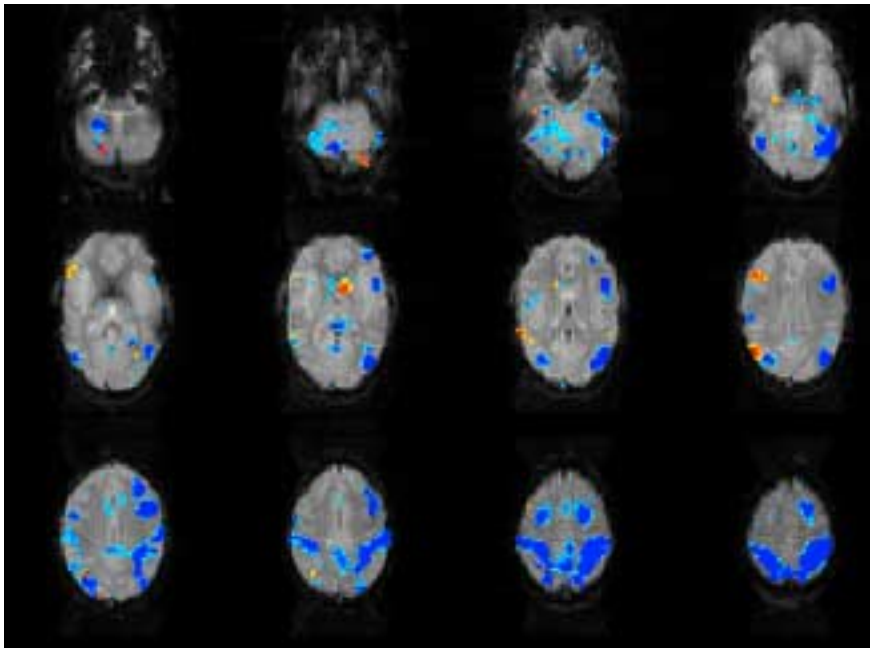
Functional studies were done with PET in the late 1980s and early 1990s, now fMRI is becoming possible. Down to $1 \times 1 \times 3$ mm voxels.

PET has lower resolution, and relies on injecting agents into the blood stream. 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 and temporal resolution. So most commonly stimuli are applied for a period of 10–30 secs, images taken around every 3 secs, with several repeats of the stimulus being available for one subject.

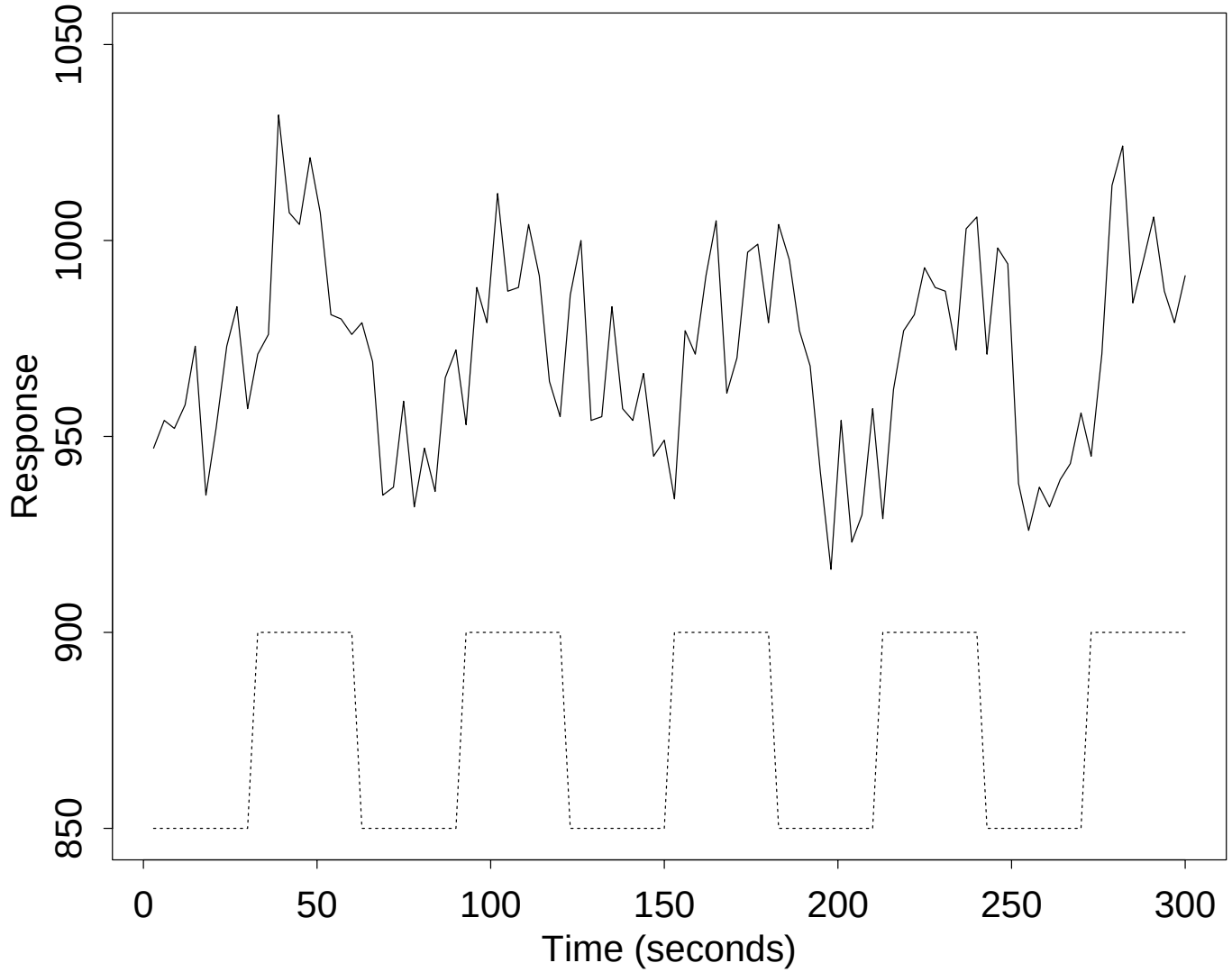
The commonly addressed statistical issue is ‘has the brain state changed’, and if so where?





Left: A pain experiment. Blue before drug administration, green after, yellow both.

Right: A verbal/spatial reasoning test, averaged over 4 subjects. 12 slices, read row-wise from bottom of head to top. Blue=spatial, red=verbal.



A real response (solid line) from a 100-scan (TR=3sec) dataset in an area of activation from the visual experiment. The periodic boxcar shape of the visual stimulus is shown below.

The GLM approach

A SAS-ism: it means linear models. May take the autocorrelation of the noise (in time) into account.

- May or may not remove overall mean.
- May or may not remove trends.
- Design matrix may contain
  - An overall mean
  - A linear or other trend
  - The signal convolved with a linearly-parametrized HRF: often a couple of gammas or a gamma density and its derivative.
- The signal is usually filtered by a matrix S, so the model becomes

$$SY = SX\beta + S\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim N(0, \sigma^2 V(\theta))$$

Variations on the theme in a series of papers by Friston *et al*:

> Analysis of functional MRI time-series
>
> Analysis of functional MRI time-series revisited
>
> Analysis of functional MRI time-series revisited—again

All fit β by least-squares, but differ in the estimate used of σ^2. The third paper suggests a scaled version of the residual MSq.

Not exactly unknown statistical theory (to statisticians)!

Two main issues:

1. What is the best estimate $\hat{\beta}$ of β?
2. What is a good (enough) estimate of its null-hypothesis variability, var$(\hat{\beta})$?

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.

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.

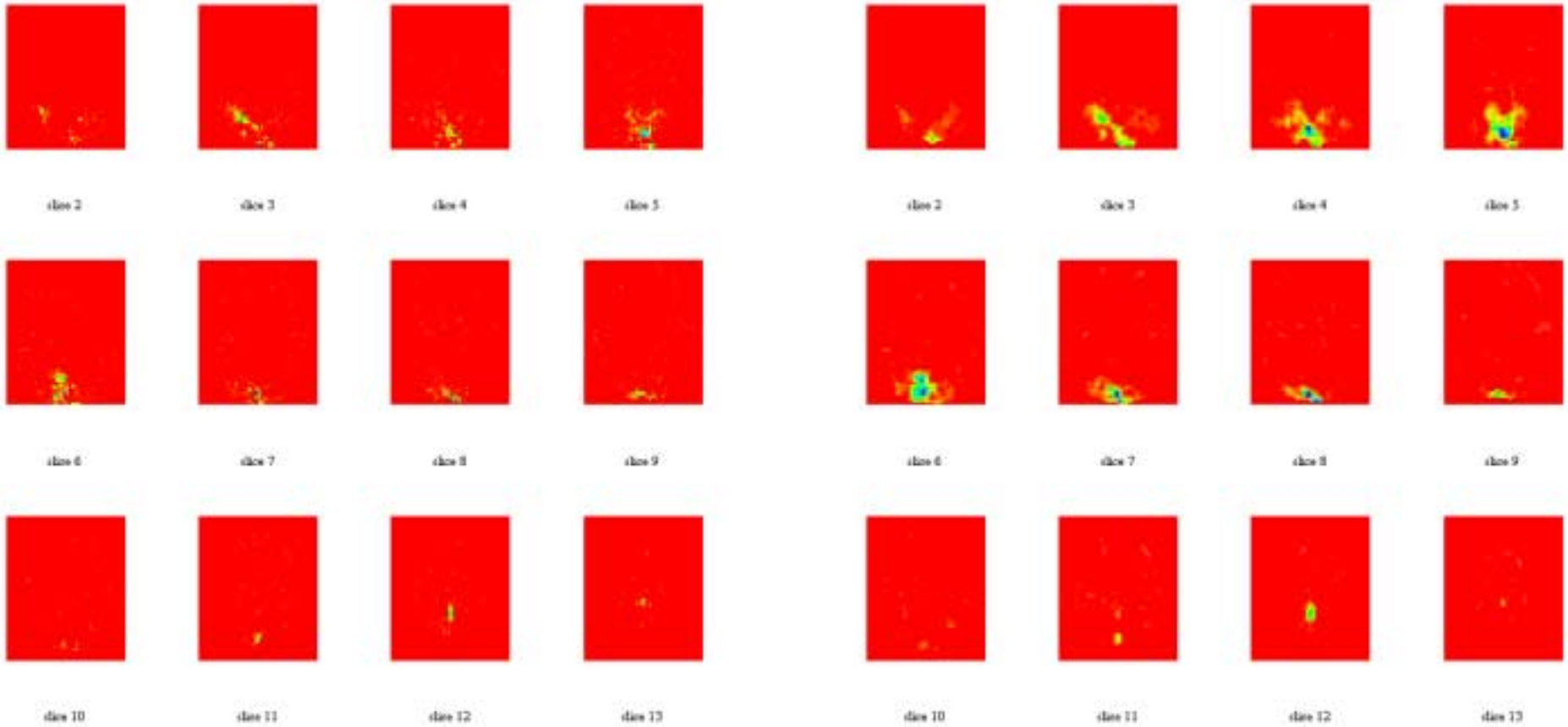
More generally, we want the mean of the A and B phases, and there will be a delayed response (approximately known) giving a cross-over effect. Instead, use a matched filter ($\sin$ wave?) to extract effect, and estimated autocorrelations (like Hannan estimation) or spectral theory to estimate variability.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

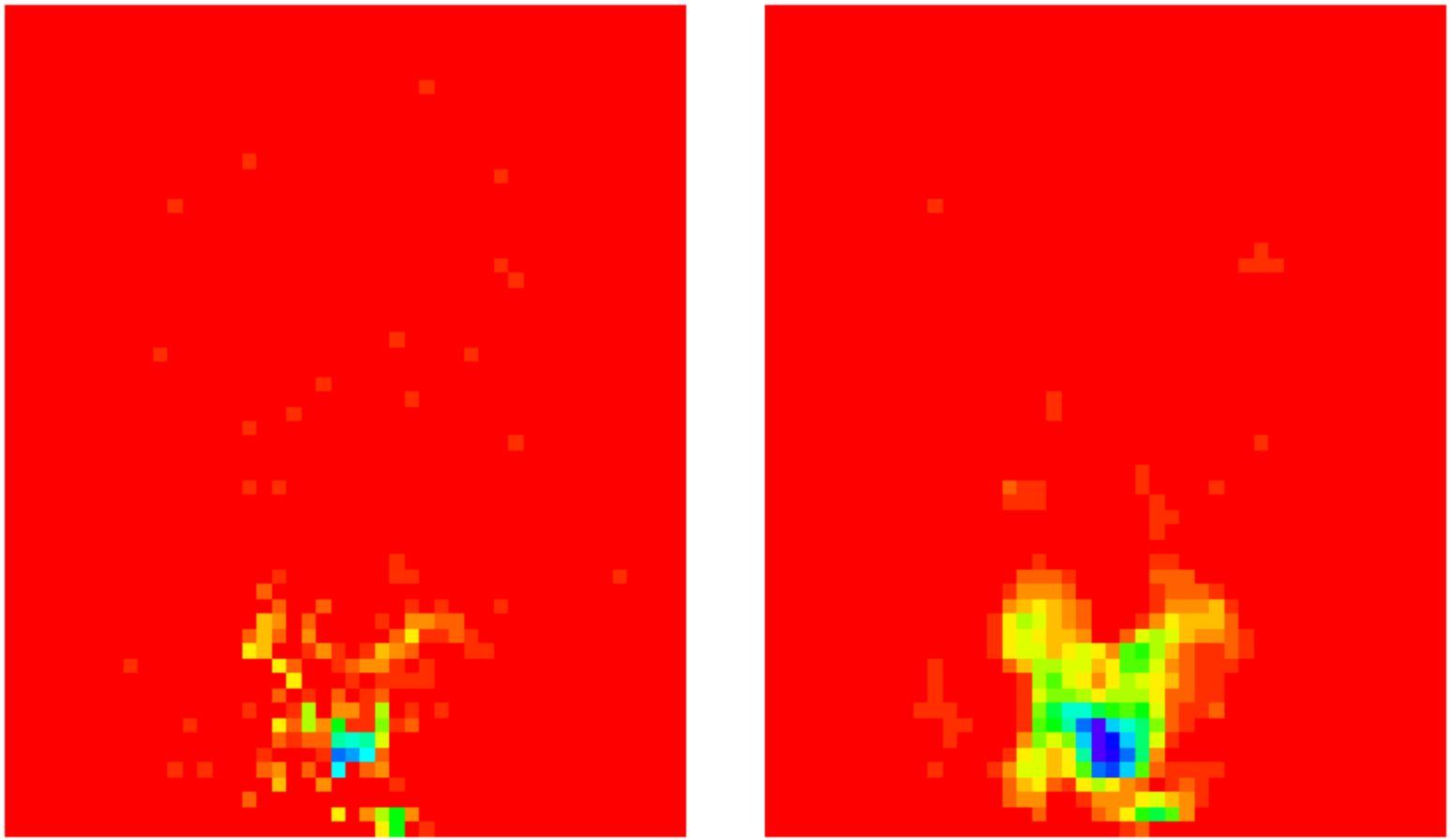
Data on $64 \times 64 \times 14$ grid of voxels. (Illustrations omit top and bottom slices and areas outside the brain, all of which show considerable activity, probably due to registration effects.)

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.

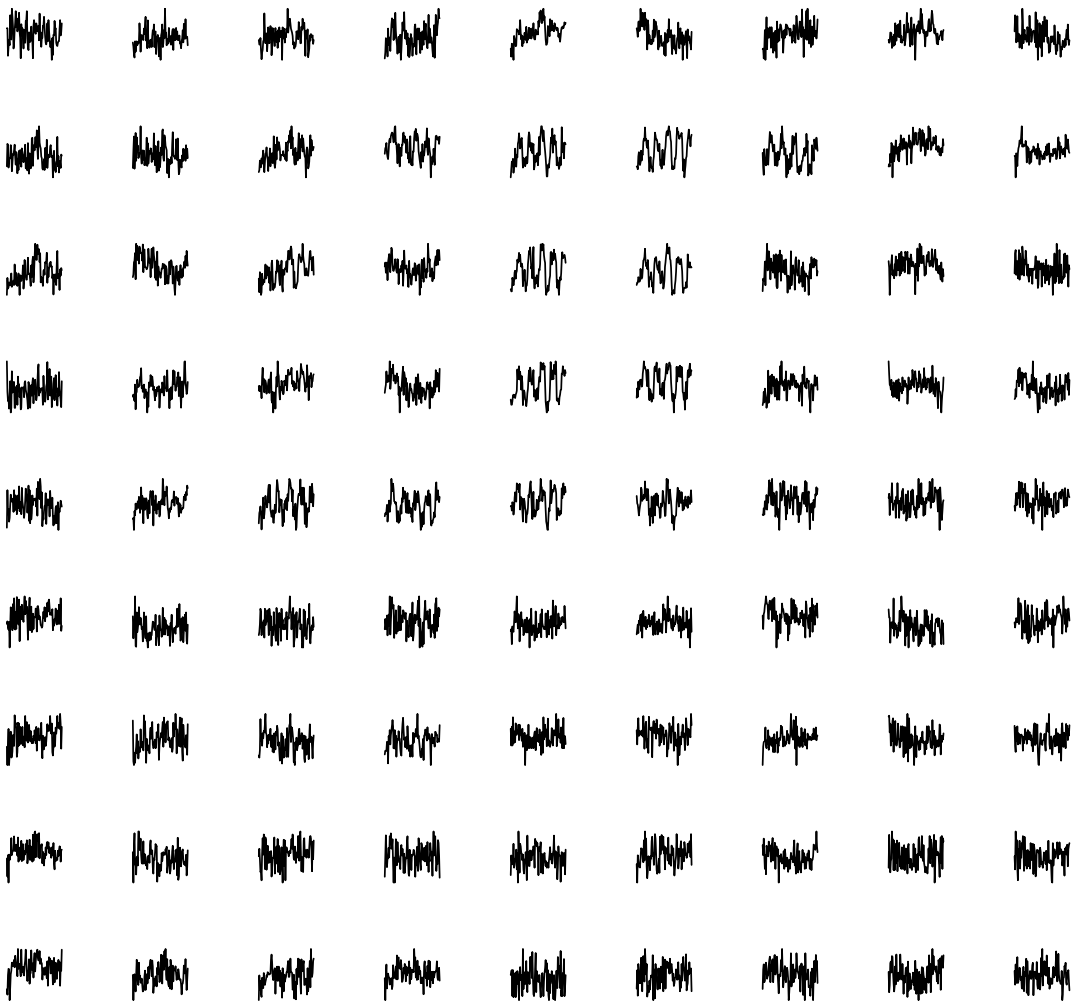


SPM99 t–statistic images, with spatial smoothing on the right



Slice 5, with spatial smoothing on the right

A Closer Look at some Data

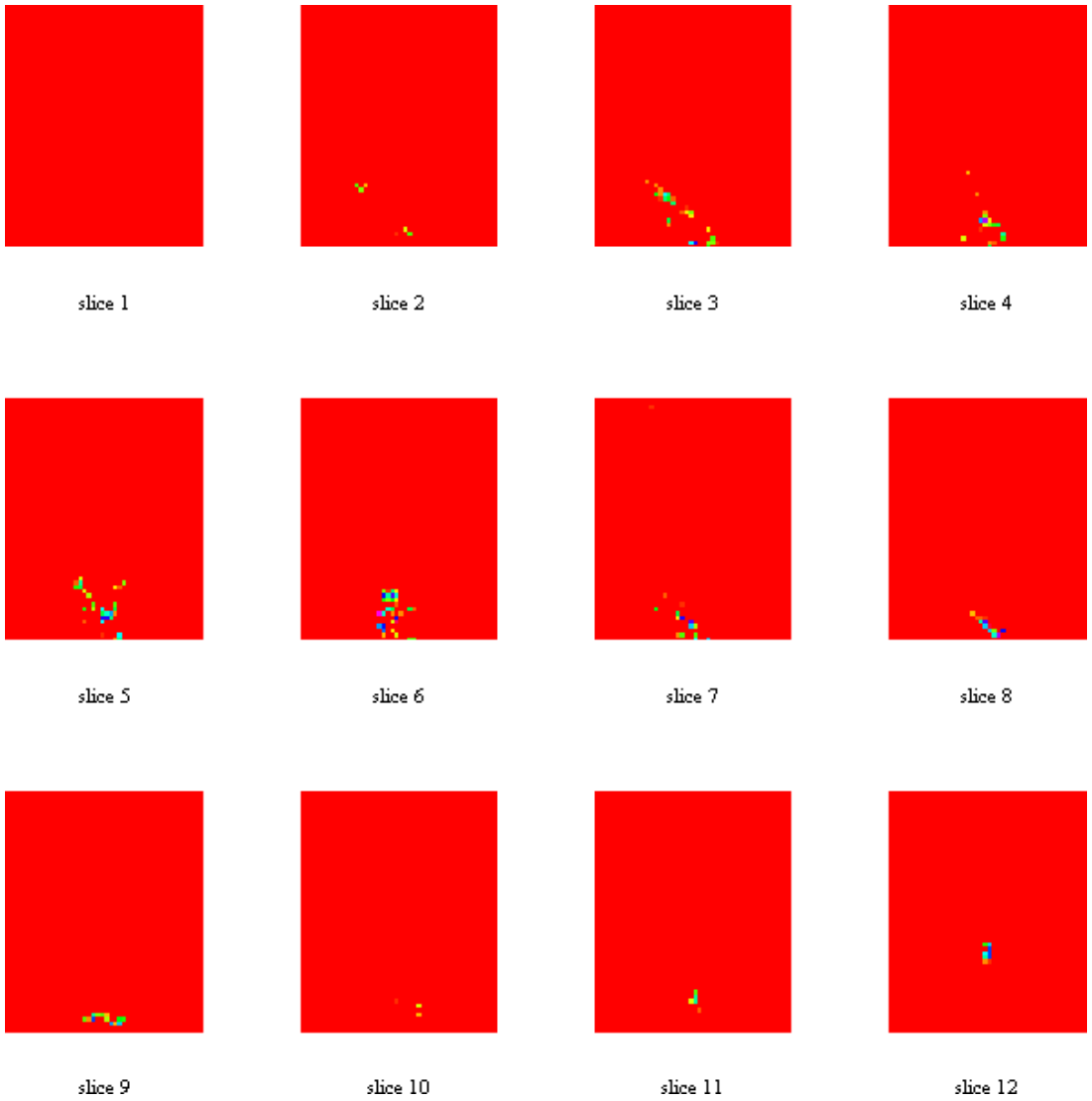


A 10×10 grid in an area of slice 5 containing activation.

Alternative Analyses

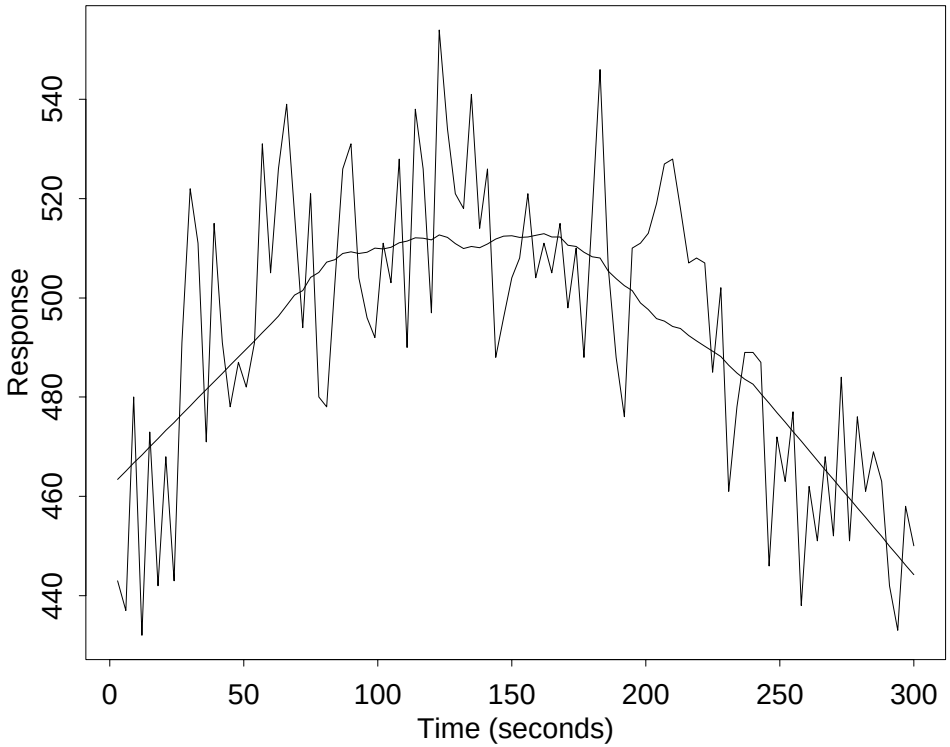
- Work with raw data.
- Non-parametric robust de-trending, Winsorizing if required.
- Work in spectral domain.
- Match a filter to the expected pattern of response (square wave input, modified by the haemodynamic response).
- Non-parametric smooth estimation of the noise spectrum at a voxel, locally 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 on ca 400MHz PC).



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).

Trend-removal

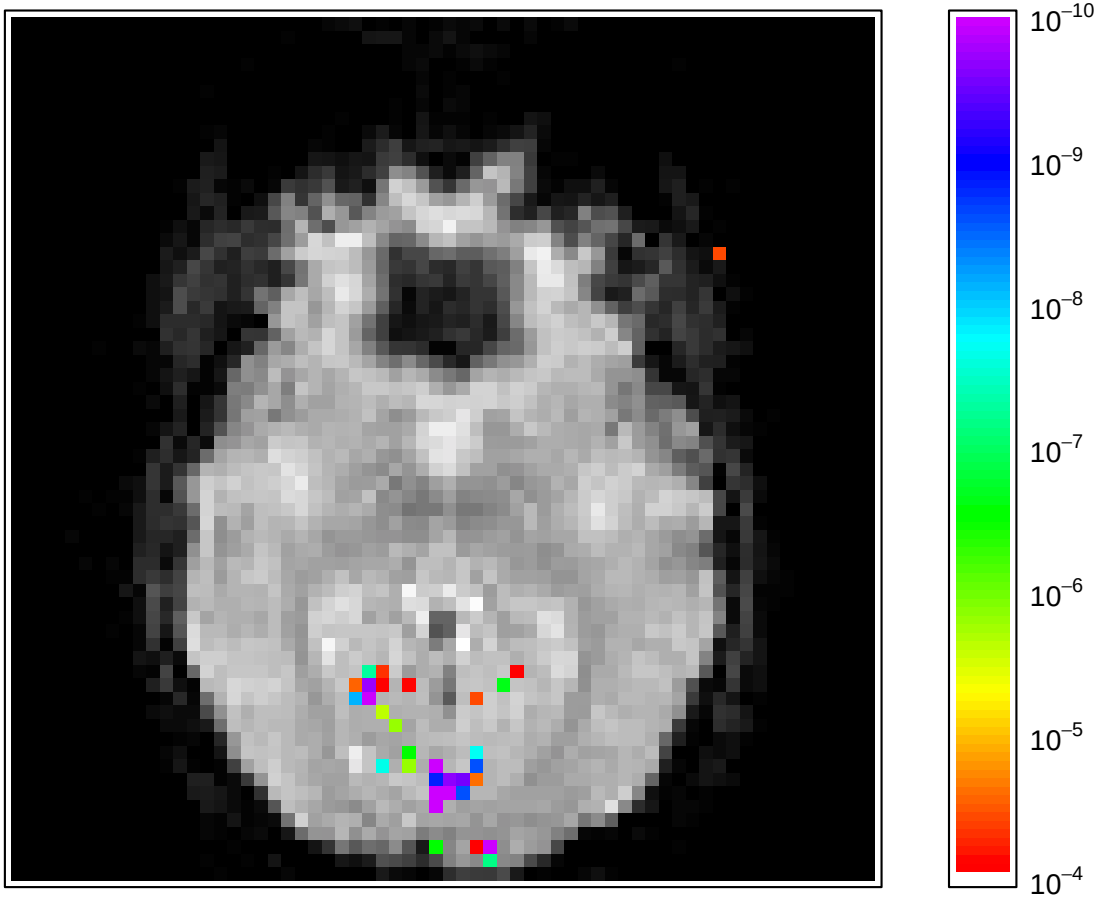


A voxel time series from the dataset showing an obvious non-linear trend.

We used a running-lines smoother rejecting outliers (and Winsorizing the results).

Plotting p values

p-value image of slice 5 thresholded to show p-values below 10^{-4} and overlaid onto an image of the slice. Colours indicate differential responses within each cluster. An area of activation is shown in the visual cortex, as well as a single ‘false-positive’, that occurs outside of the brain.

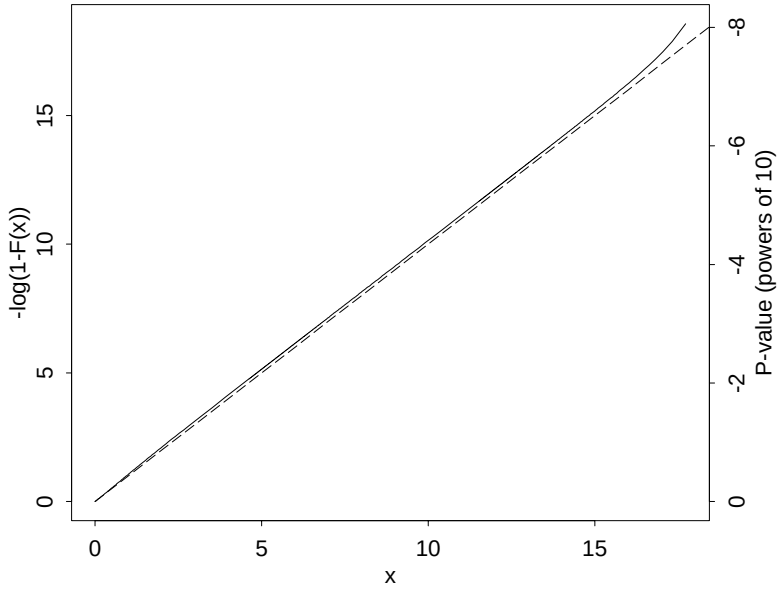


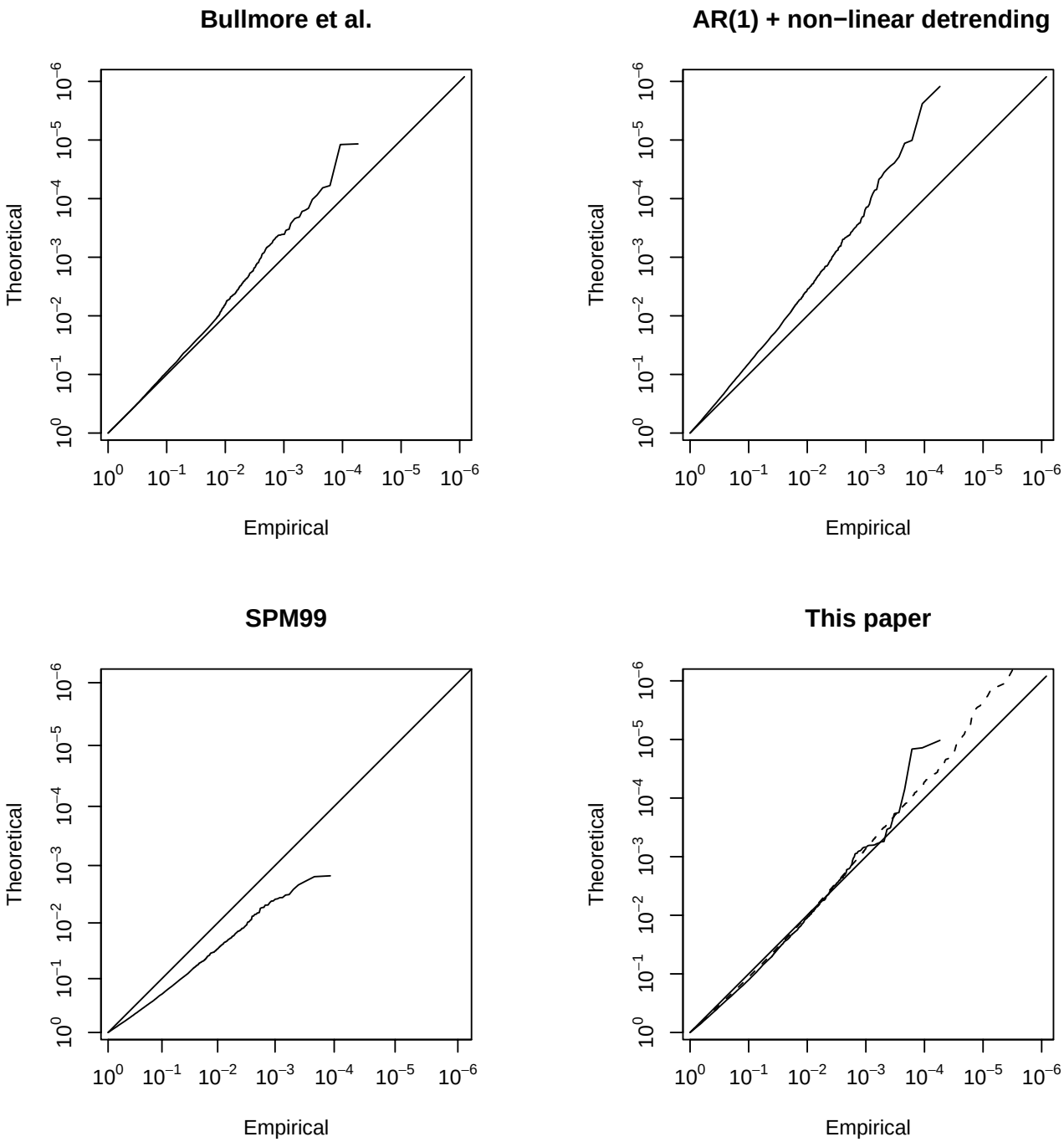
Calibration

Before we worry about multiple comparisons, are the t-statistics (nearly) t-distributed?

Few people have bothered to check, and those who did (Bullmore, Brammer *et al*, 1996) found they were not.

We can use null experiments as some sort of check.
In our analysis we can use other frequencies to self-calibrate, but *we* don't need to:




Bullmore et al.
AR(1) + non−linear detrending
SPM99
This paper
Theoretical
Empirical
10^{0} 10^{-1} 10^{-2} 10^{-3} 10^{-4} 10^{-5} 10^{-6}

Conclusions

- Look at your data (even if it is on this scale: millions of points per experiment).
- Data ‘cleaning’ is vital for routine use of such procedures.
- Fit your theory to your analysis, not *vice versa*.
- You need to be sure that the process is reliable, as no one can check on this scale.
- There is no point in building sophisticated methods on invalid foundations.
- It is amazing what can be done in high-level languages on cheap computers.