Finding Needles in Haystacks:

Tools for Finding Structure in Large Datasets

Brian D. Ripley

Workshop 24–26 June 2000;   MSRI, Berkeley
‘Mathematics and Computational Biology of Genome Analysis’

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

Outline

- data visualization
  - projection methods
  - multi-dimensional scaling
  - self-organizing maps
  - clustering

- magnetic resonance imaging (MRI) of human brains
  - partially supervised clustering
  - ageing and Alzheimer's Disease

- functional MRI
  - making t-statistic maps
  - robustness and calibration

Visualization

Challenge is to explore data in more than two or perhaps three dimensions.

via projections

Principal components is the most obvious technique: kD projection of data with largest variance matrix (in several senses). Usually ‘shear’ the view to give uncorrelated axes.

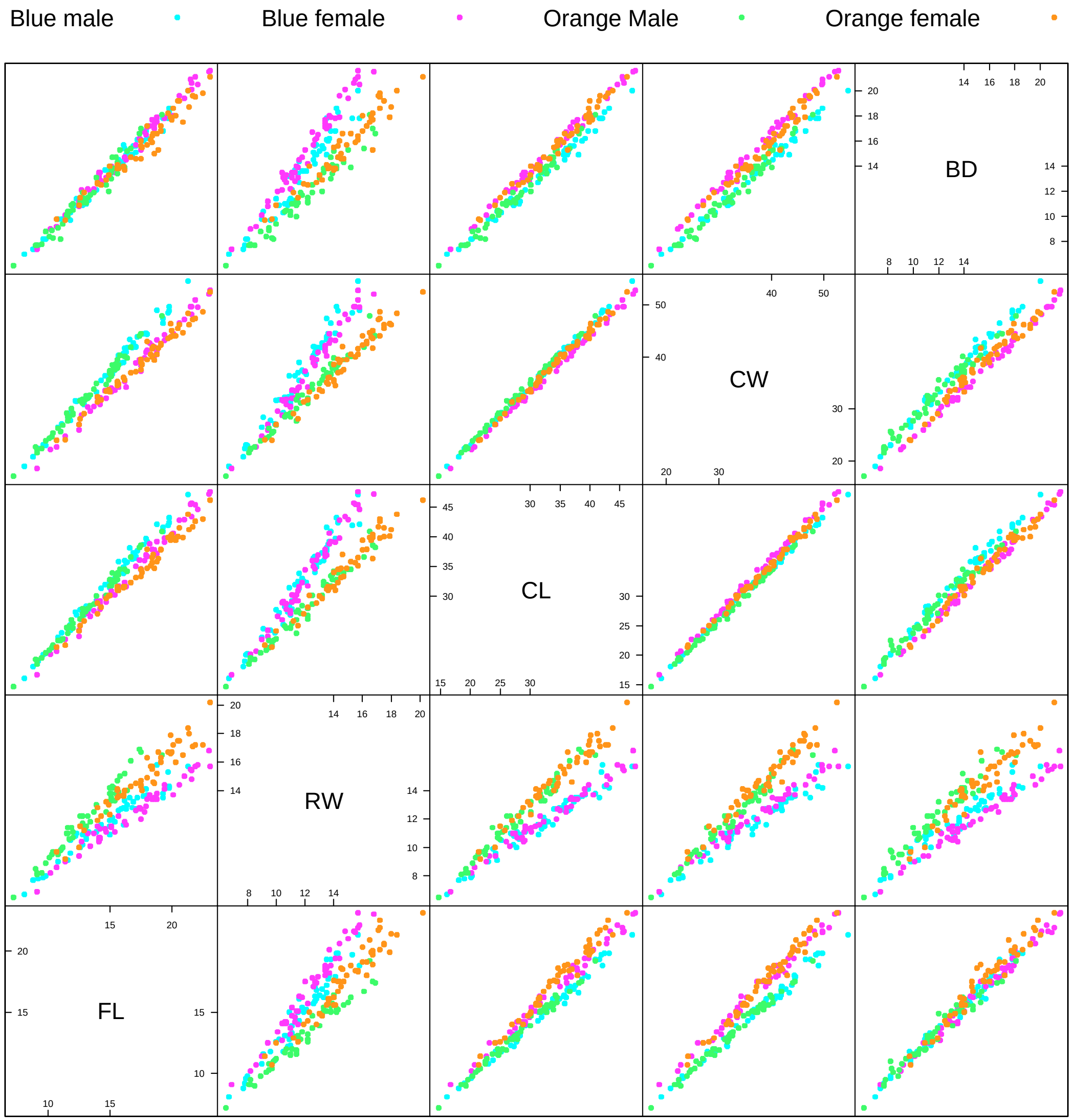
Lots of other projections looking for ‘interesting’ views, for example groupings, outliers, clumping. Known as (exploratory) *projection pursuit*.

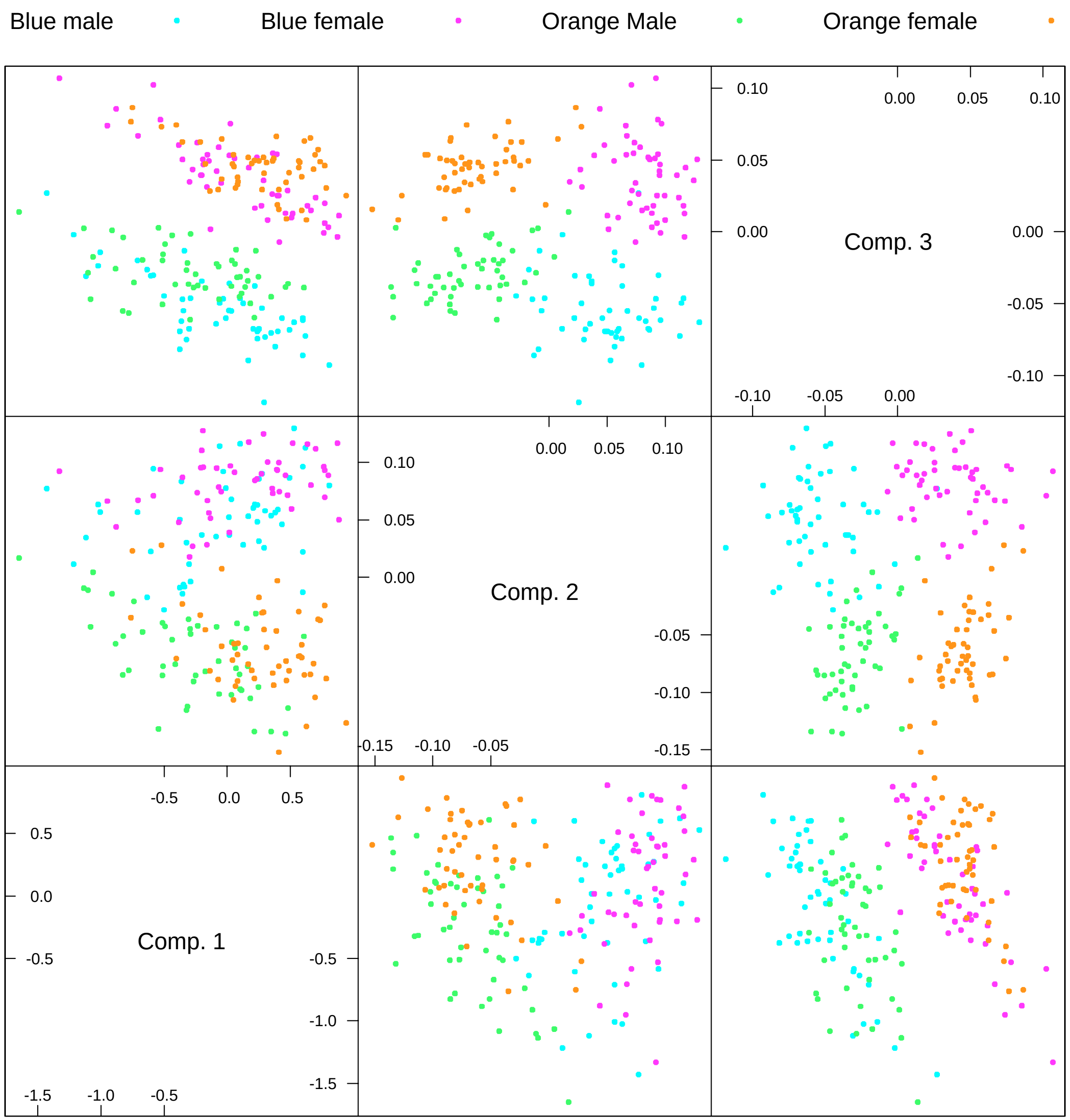
Implementation via numerical brute-force: freely available in XGobi.
‘Random’ searching (so-called *grand tours*) are not viable even in 5D.

Leptograpsus variegatus Crabs

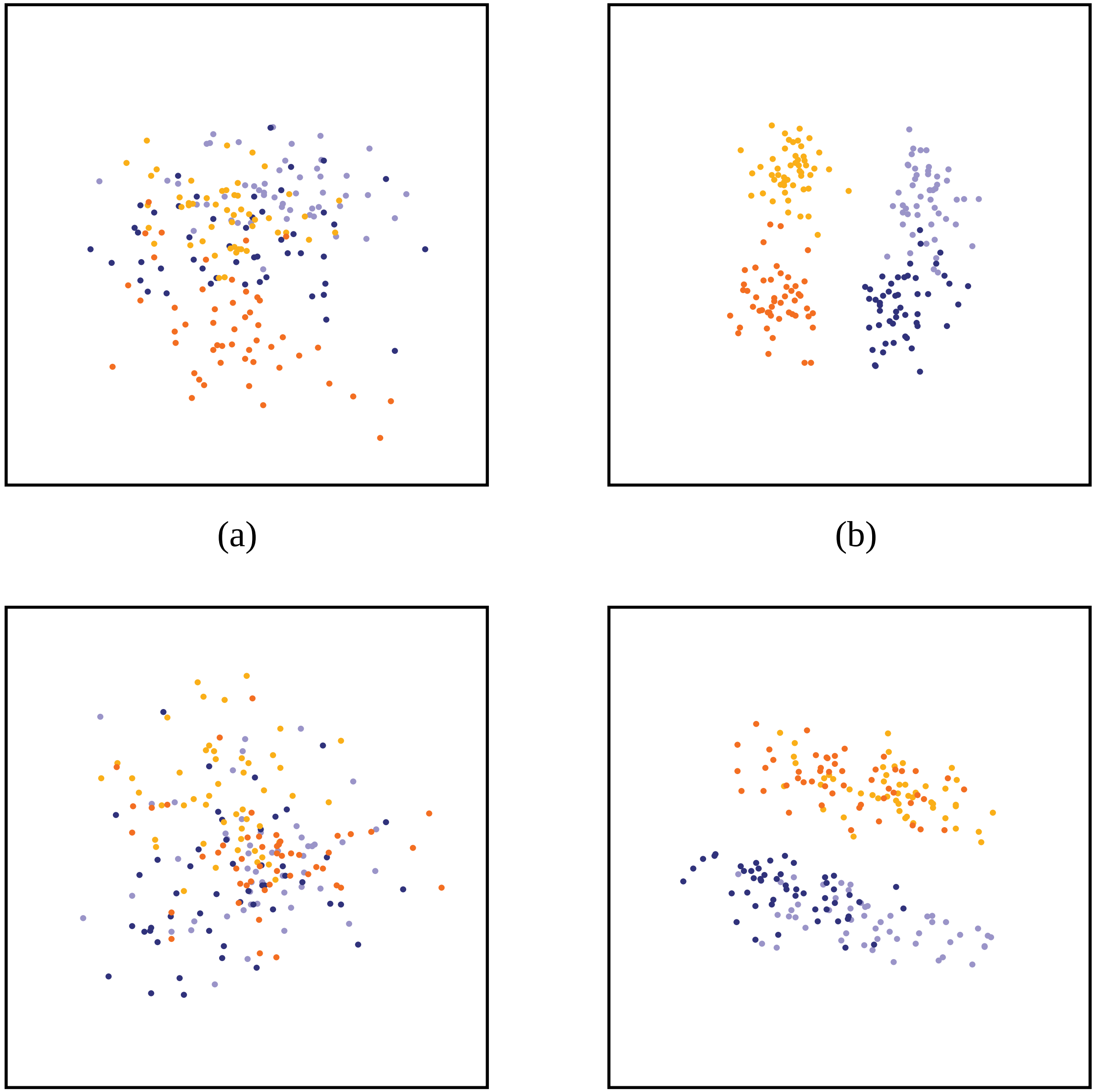
200 crabs from Western Australia. Two colour forms, blue and orange; collected 50 of each form of each sex. Are the colour forms species?

Measurements of carapace (shell) length `CL` and width `CW`, the size of the frontal lobe `FL`, rear width `RW` and body depth `BD`


Blue male ●
Blue female ●
Orange Male ●
Orange female ●
BD
14 16 18 20
20
18
16
14
14
12
10
8
8 10 12 14
CW
40 50
50
40
30
20
20 30
CL
30 35 40 45
45
40
35
30
30
25
20
15
15 20 25 30
RW
14 16 18 20
20
18
16
14
14
12
10
8
8 10 12 14
FL
15 20
20
15
15
10
10 15



First three principal components on log scale.



(a)

(b)

(c)

(d)

Projections of the *Leptograpsus* crabs data found by projection pursuit. View (a) is a random projection. View (b) was found using the natural Hermite index, view (c) by the Friedman–Tukey index and view (d) by Friedman's (1987) index.

Independent Components Analysis

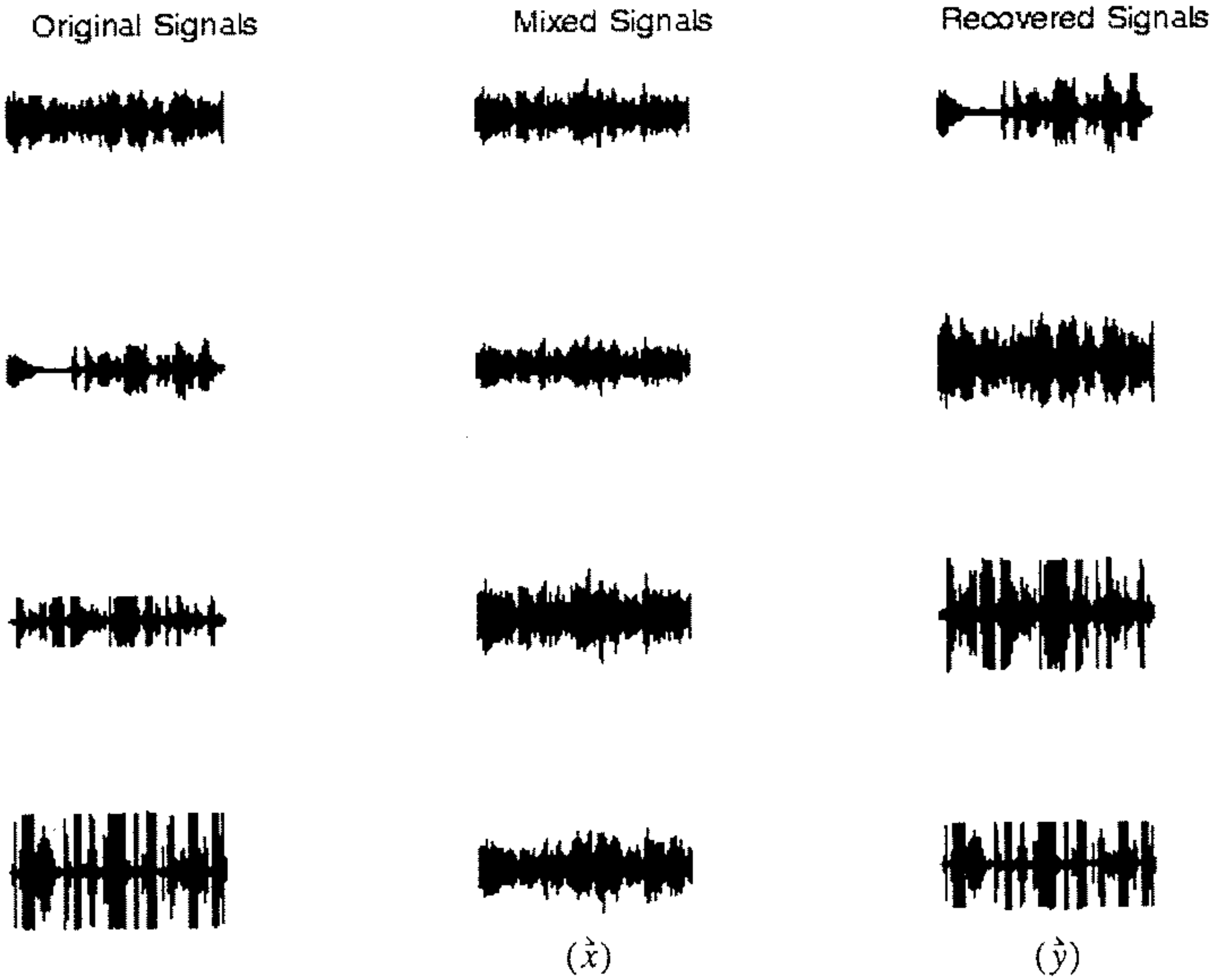
A ‘hot topic’ that has moved from field to field over the last decade. Originally(?) used for blind source signal separation in geophysics.

A projection pursuit technique in which the objective is to find k independent linear combinations. So minimize entropy difference between joint kD projection distributions and the product of their marginals.

Many local minima. No guarantee that you will find k signals not k noise sources. Choice of k may be crucial.

Many impressive results: but often every other visualization technique finds them. *‘In the land of the blind . . . .’*

A close relative of *factor analysis* and other latent variable methods.



ICA experiment, from Deco & Obradovic (1996).

Multidimensional Scaling

Aim is to represent distances between points well.

Suppose we have distances (d_{ij}) between all pairs of n points, or a *dissimilarity* matrix. Classical MDS plots the first k principal components, and minimizes

$$\sum_{i \neq j} d_{ij}^2 - \widetilde{d}_{ij}^2$$

where $(\widetilde{d}_{ij})$ are the Euclidean distances in the kD space.

More interested in getting small distances right. Sammon (1969) proposed

$$\min E(d, \widetilde{d}) = \frac{1}{\sum_{i \neq j} d_{ij}} \sum_{i \neq j} \frac{(d_{ij} - \widetilde{d}_{ij})^2}{d_{ij}}$$

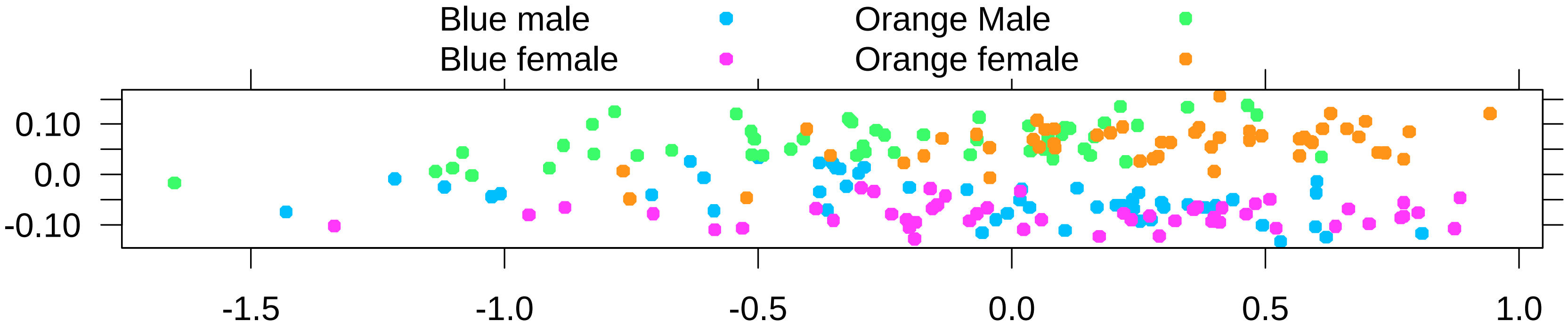
Shepard and Kruskal (1962–4) proposed only to preserve the ordering of distances, minimizing

$$STRESS^2 = \frac{\sum_{i\neq j} \left[\theta(d_{ij}) - \widetilde{d}_{ij}\right]^2}{\sum_{i\neq j} \widetilde{d}_{ij}^2}$$

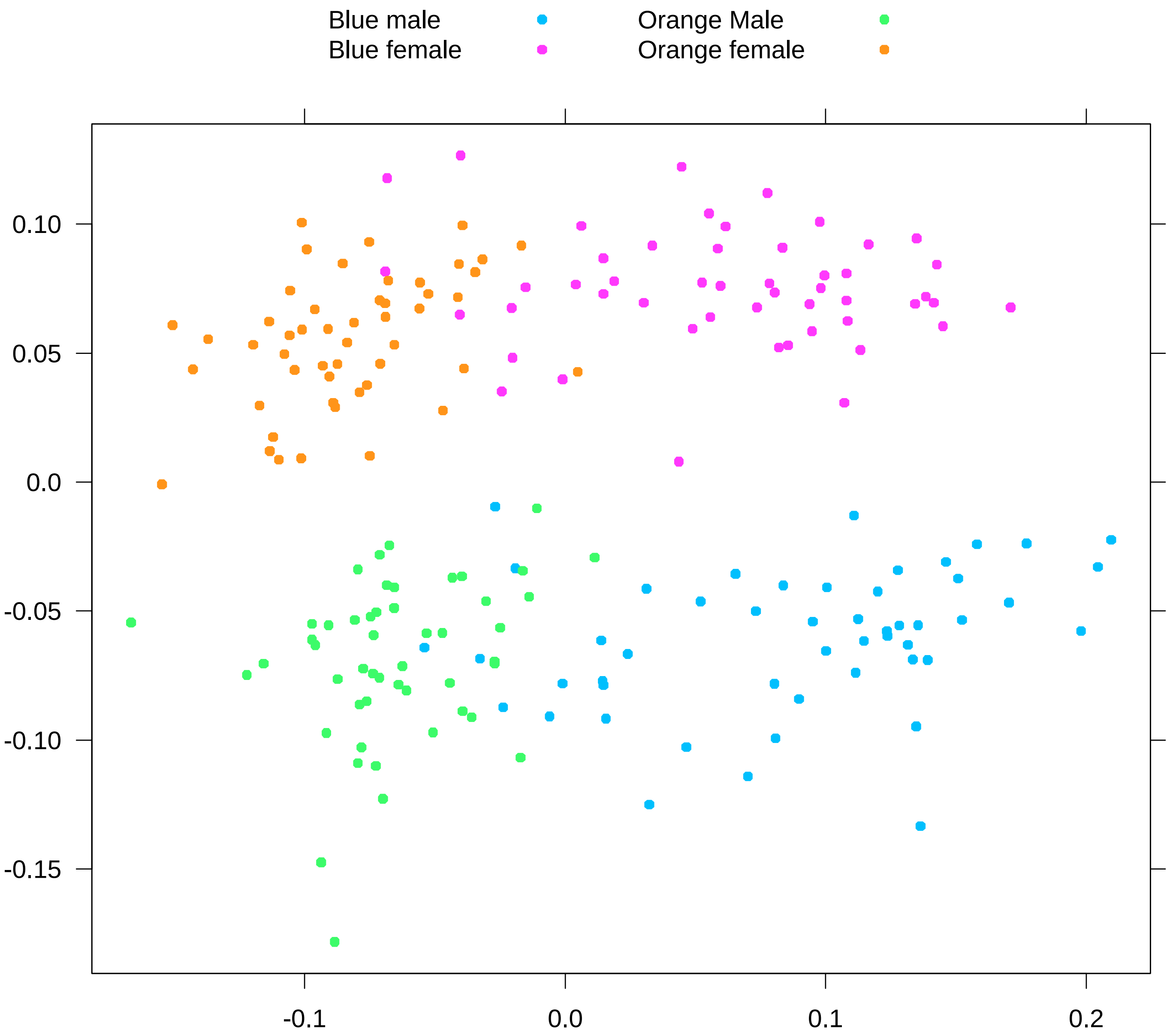
over both the configuration of points and an increasing function θ.

The optimization task is quite difficult and this can be slow.

Multidimensional scaling



An order-preserving MDS plot of the (raw) crabs data.

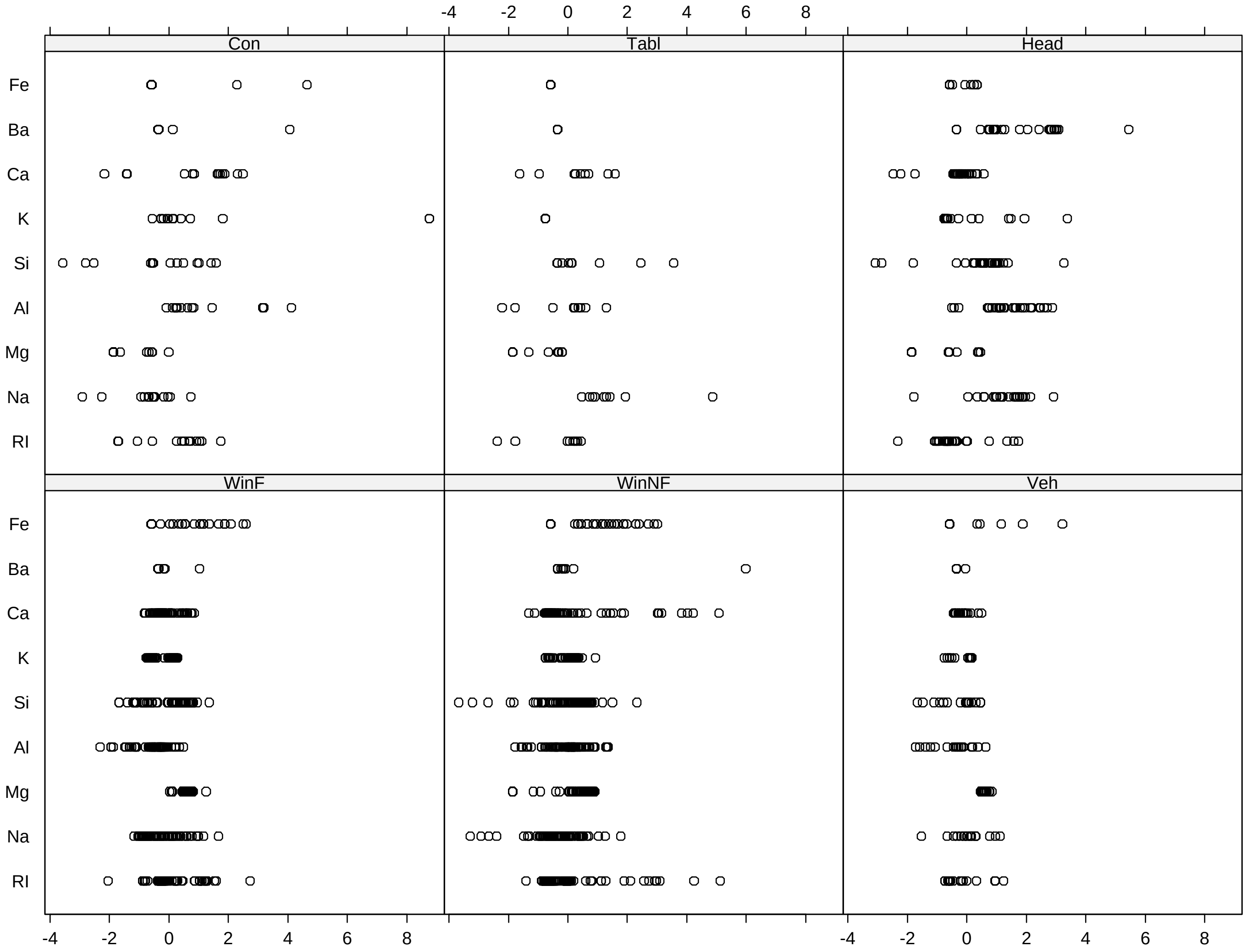


After re-scaling to (approximately) constant carapace area.

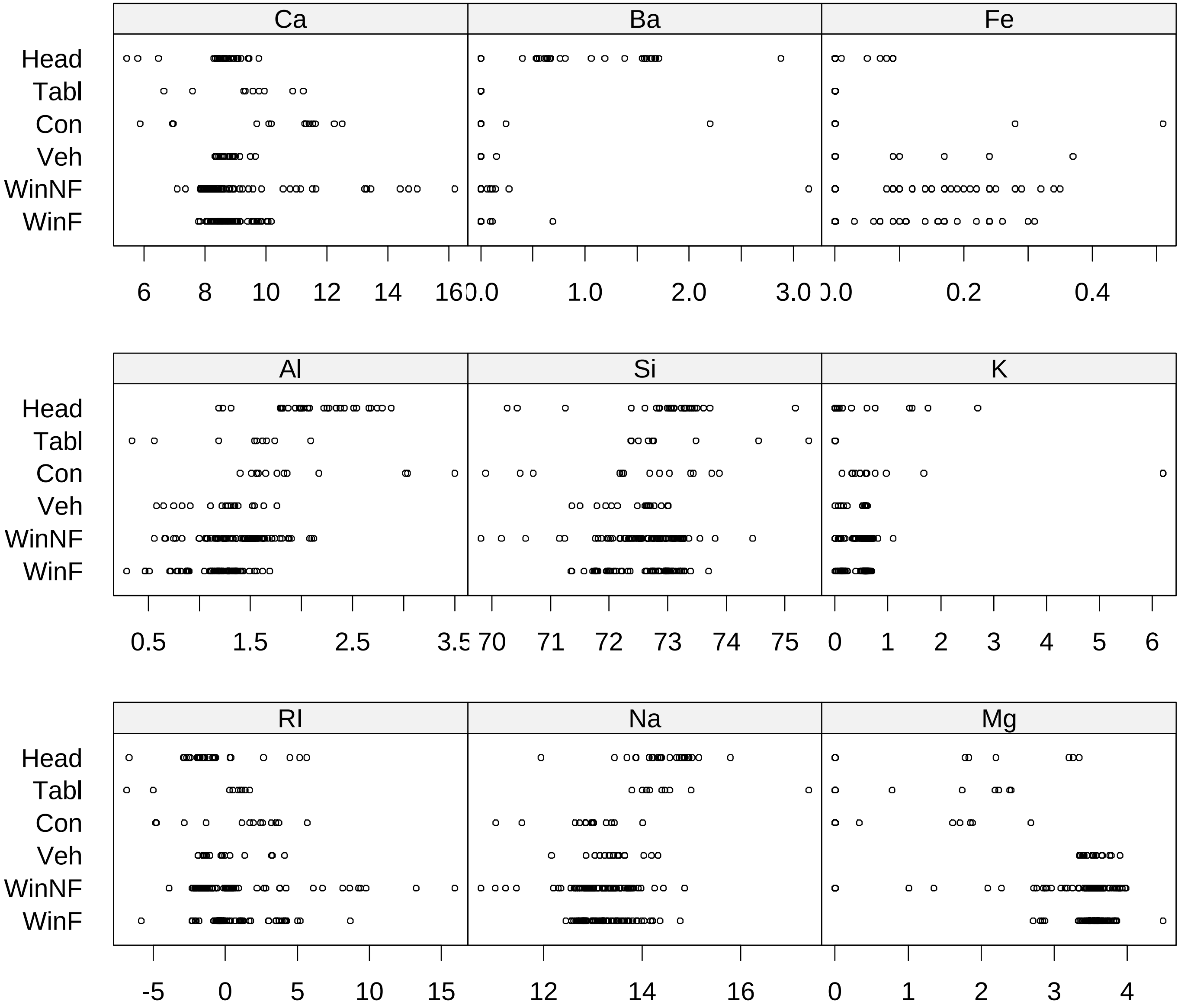
A Forensic Example

Data on 214 fragments of glass collected at scenes of crimes. Each has a measured refractive index and composition (weight percent of oxides of Na, Mg, Al, Si, K, Ca, Ba and Fe).

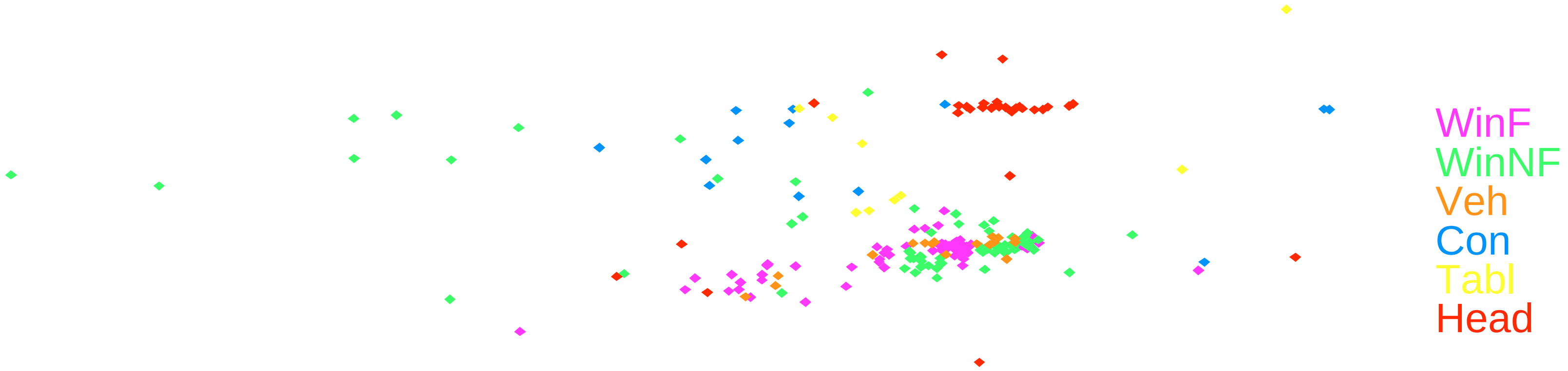
Grouped as window float glass (70), window non-float glass (76), vehicle window glass (17) and other (containers, tableware, headlamps) (22).



Strip plot by type of glass.



Strip plot by type of analyte.



Isotonic multidimensional scaling representation.

Kohonen’s Self-Organizing Maps

Kohonen describes his own motivation as:

> ‘I just wanted an algorithm that would effectively map similar patterns (pattern vectors close to each other in the input signal space) onto contiguous locations in the output space.’
>
> Kohonen (1995, p. VI)

He interpreted ‘contiguous’ via a rectangular or hexagonal 2-D lattice.

In K-*means clustering* the data are split into K groups, and each example is assigned to the cluster whose representative $\boldsymbol{m}_j$ is nearest to the example. The cluster representatives (‘centre’) are then adjusted to be the centroid of the group, and iteration gives a simple, finite, algorithm.

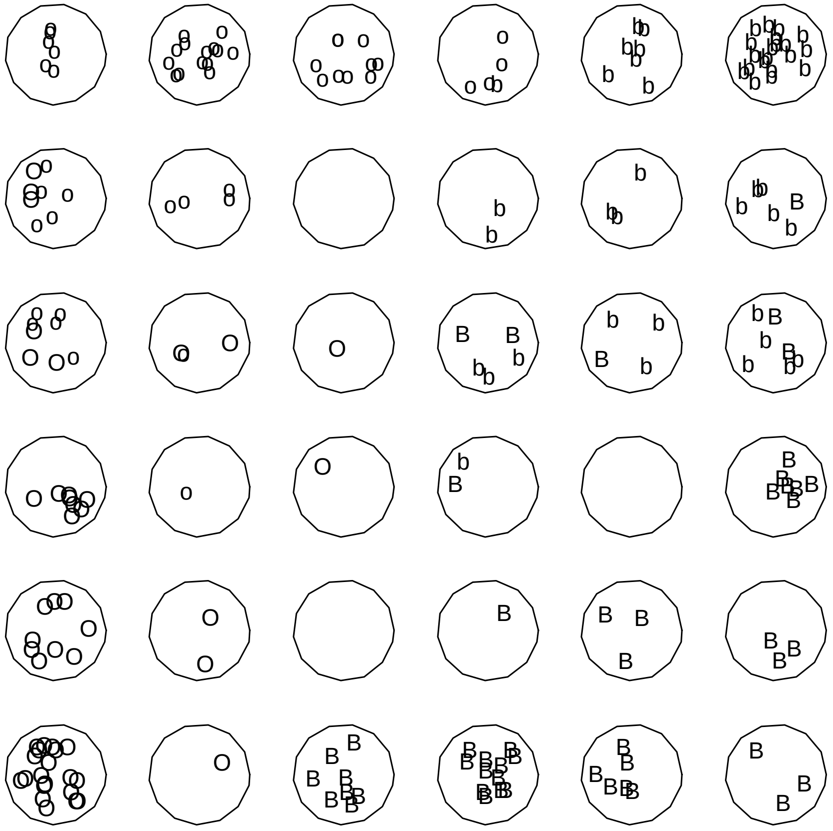
In SOM (self-organizing mapping) the representatives $(\boldsymbol{m}_j)$ are arranged on a regular grid, with representatives at nearby points on the grid are more similar that those which are widely separated.

Examples are presented in turn until convergence. The $\boldsymbol{m}_j$ are initially assigned at random. Whenever an example $\boldsymbol{x}$ is presented, the closest representative $\boldsymbol{m}_j$ is found. Then

$$\boldsymbol{m}_i \leftarrow \boldsymbol{m}_i + \alpha[\boldsymbol{x} - \boldsymbol{m}_i] \qquad \text{for all neighbours } i\ .$$

Both the constant α and the definition of 'neighbour' change with time.

A cruder form of MDS, but one that scales to 100,000+ examples.



SOM mapping of the crabs data to a 6×6 grid. The labels of those examples mapped to each cluster are distributed randomly within the circle representing the cluster.

Clustering

General idea is to divide data into groups such that the points within a group are more similar to each other than to those in other groups.

Important details:

- The number k of groups may or may not be known.
- May wish to allow ‘outliers’ assigned to no group.
- Could allow overlap in group membership (‘fuzzy clustering’).

Note that need a measure of (dis)similarity or distance between a point and a group of points.

A Clustering of Cluster Methods

- Agglomerative hierarchical methods.
  - Produces a set of clusterings, usually one for each $k = n, \dots, 2$.
  - Main differences are in calculating group–group dissimilarities from point–point dissimilarities.
  - Computationally easy.
- Optimal partitioning methods.
  - Produces a clustering for fixed K.
  - Need an initial clustering.
  - Lots of criteria to optimize, some based on (joint normal) probability models.
  - Can have distinct ‘outlier’ group(s).
- Divisive hierarchical methods.
  - Produces a set of clusterings, usually one for each $k = 2, \dots, K \ll n$.
  - Computationally nigh-impossible.
  - Most available methods are *monothetic* (split on one variable at each stage).

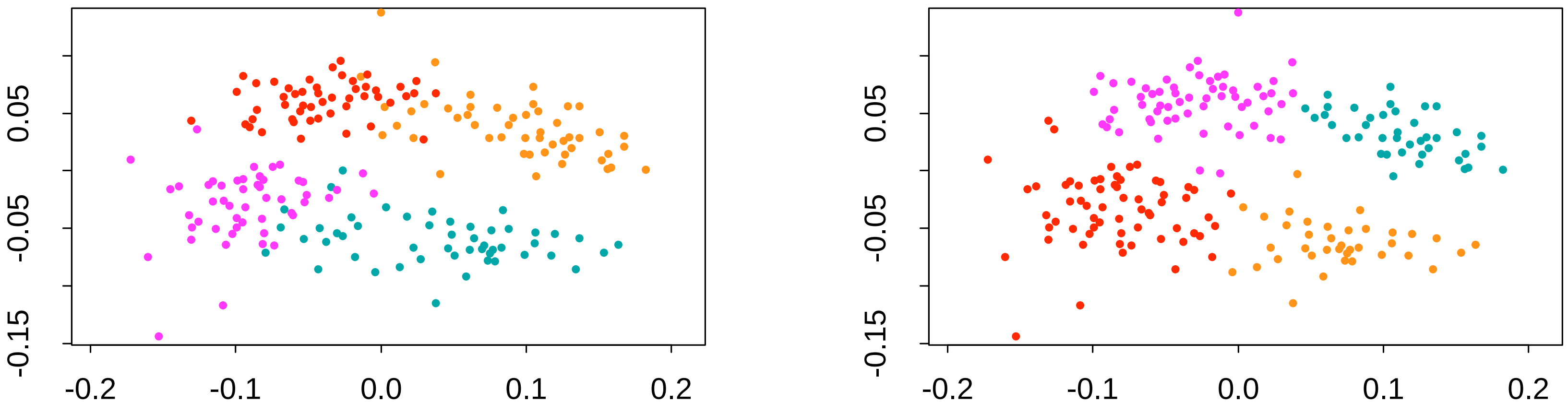
References

Comprehensive reference:
Gordon, A. D. (1999) *Classification.* Second Edition.

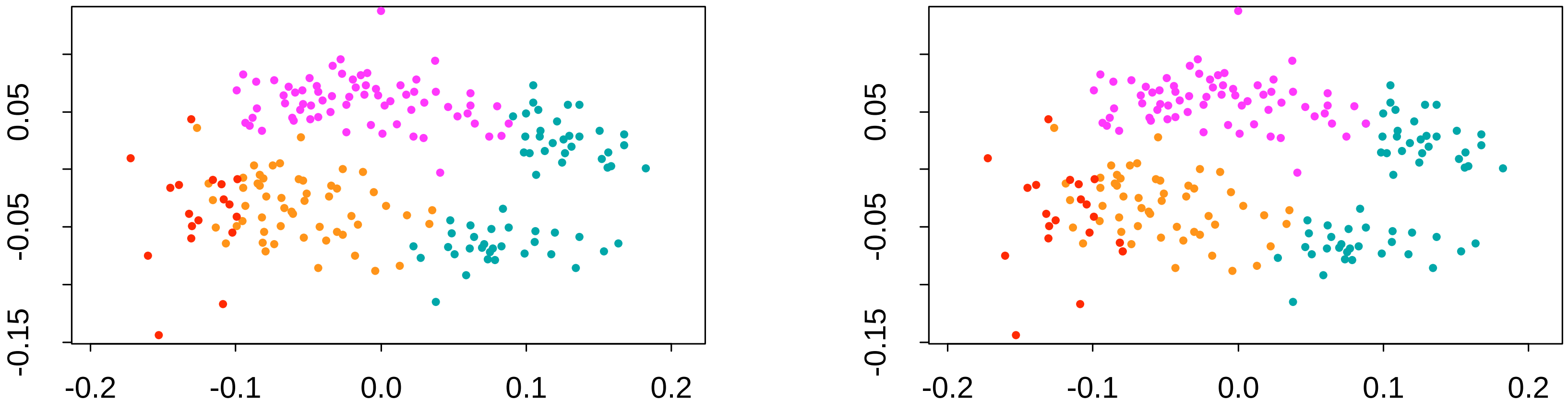
Good introduction:
Kaufman, L. and Rousseeuw, P. J. (1990) *Finding Groups in Data*.

An example

The *Leptograpsus* crabs data, with 4 groups (known in advance here). Same information as available to projection pursuit and MDS.

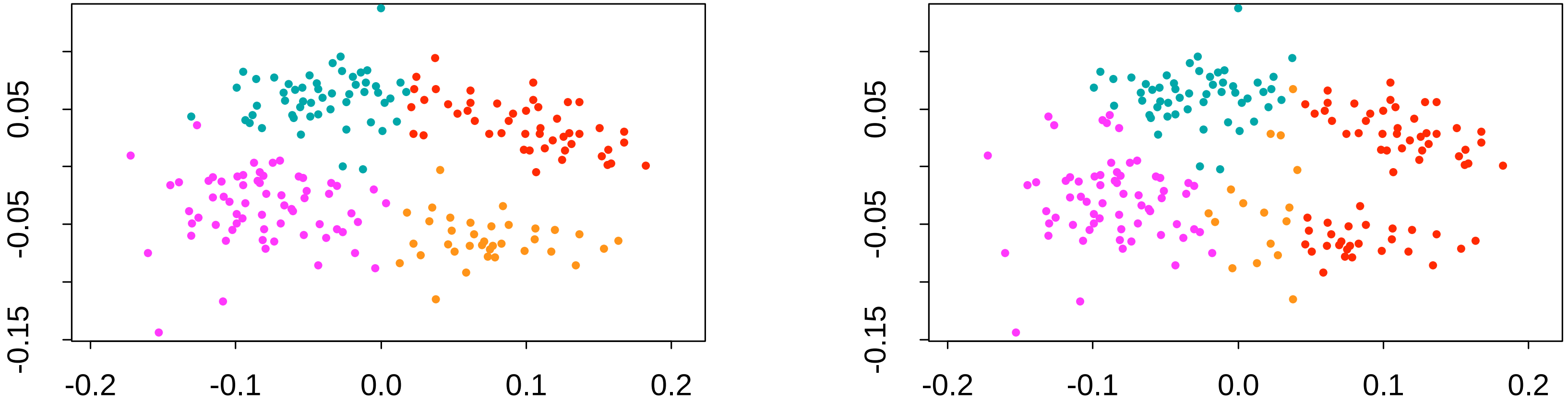


Left: True. Right: $(K = 4)$–means.



Left: complete-link hierarchical clustering.

Right: ‘maximum likelihood’ clustering with ellipsoidal clusters.



Left: Macnaughton-Smith *et al.*'s divisive method.

Right: 'hardened' classification from fuzzy clustering.

Alternative approach

Model as mixture of (multivariate) normal distributions. Hard, but some progress with MCMC approaches in the Bayesian paradigm.

Case Study:

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)

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 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.

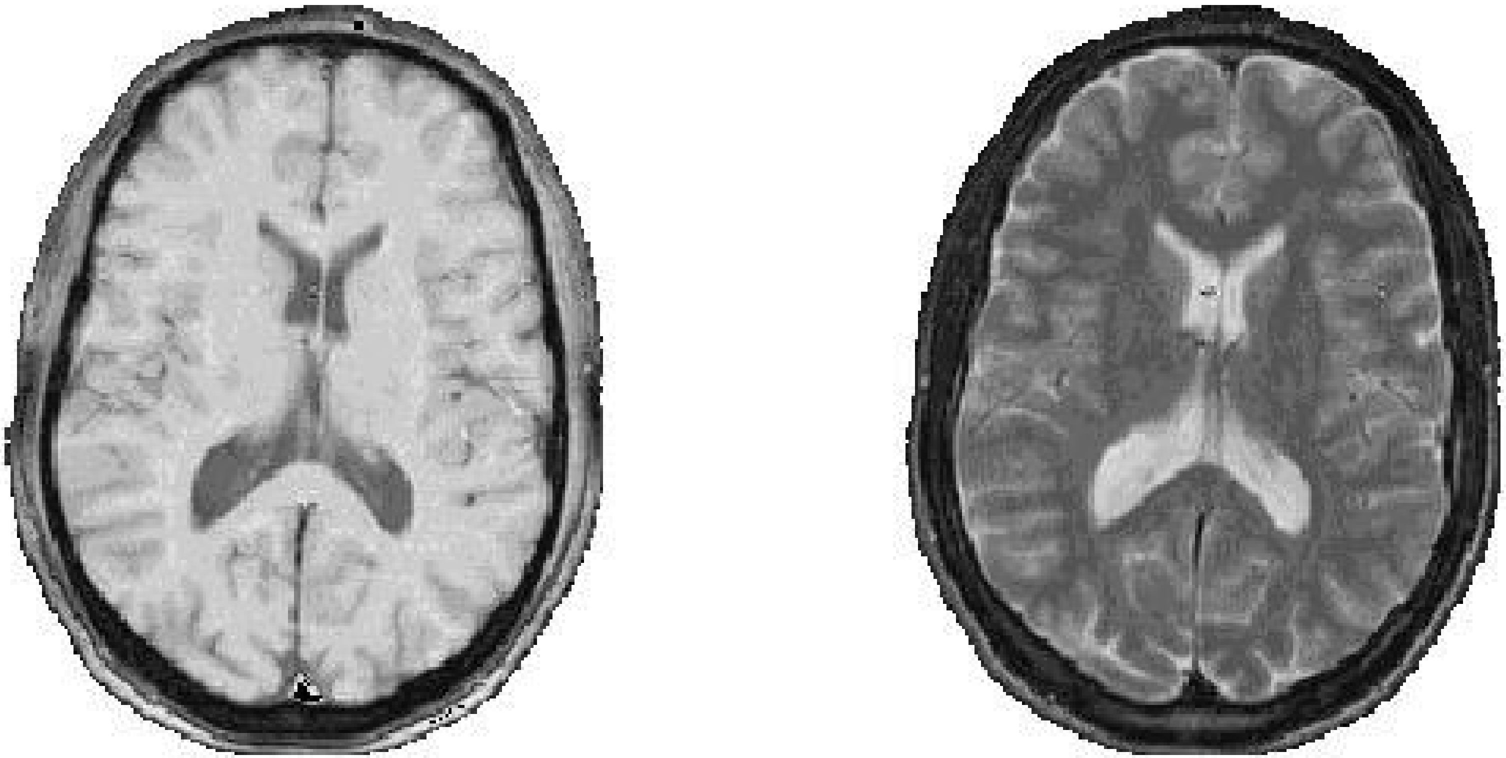
Pilot Study

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’.

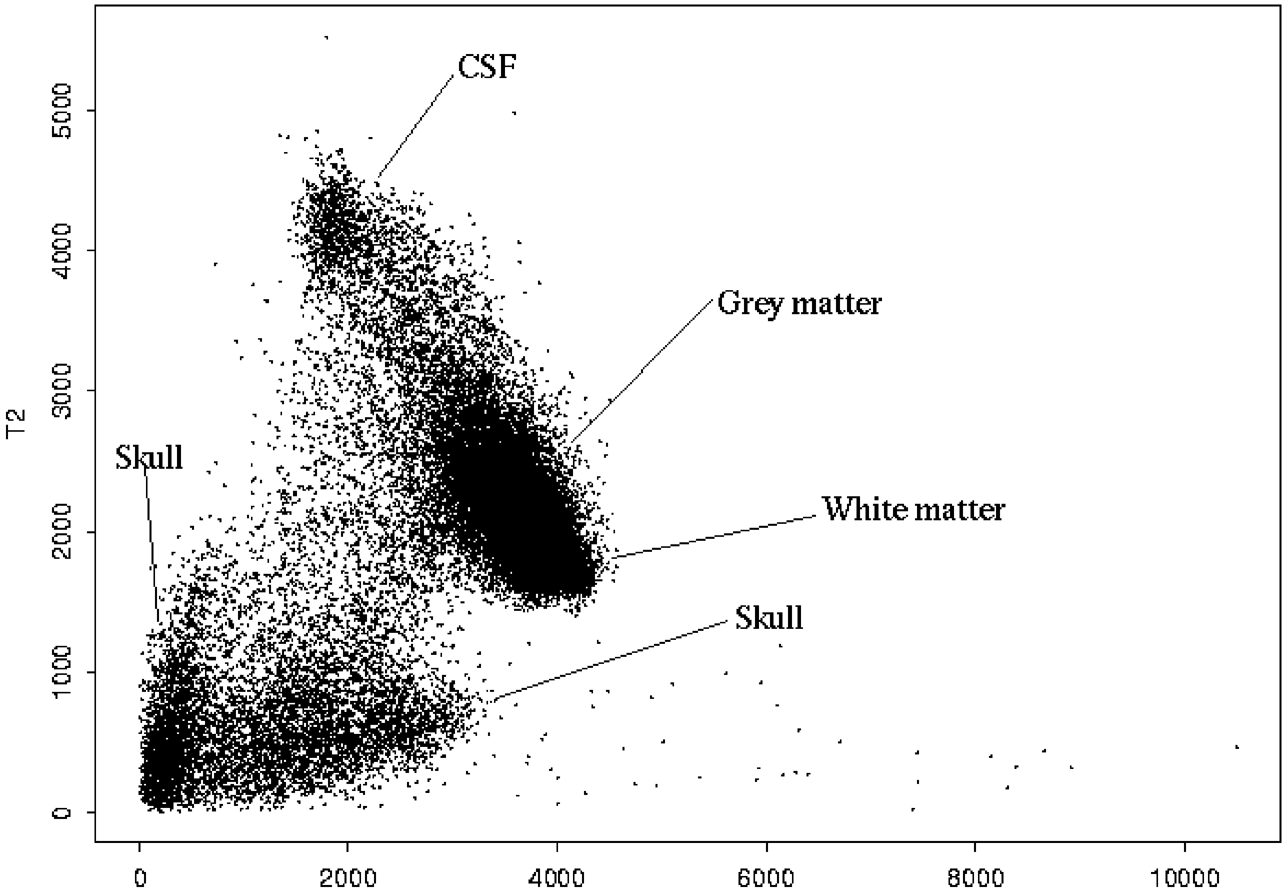
Initial aim is reliable automatic segmentation.

Some Data



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

This slice is of 172×208 pixels. 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.

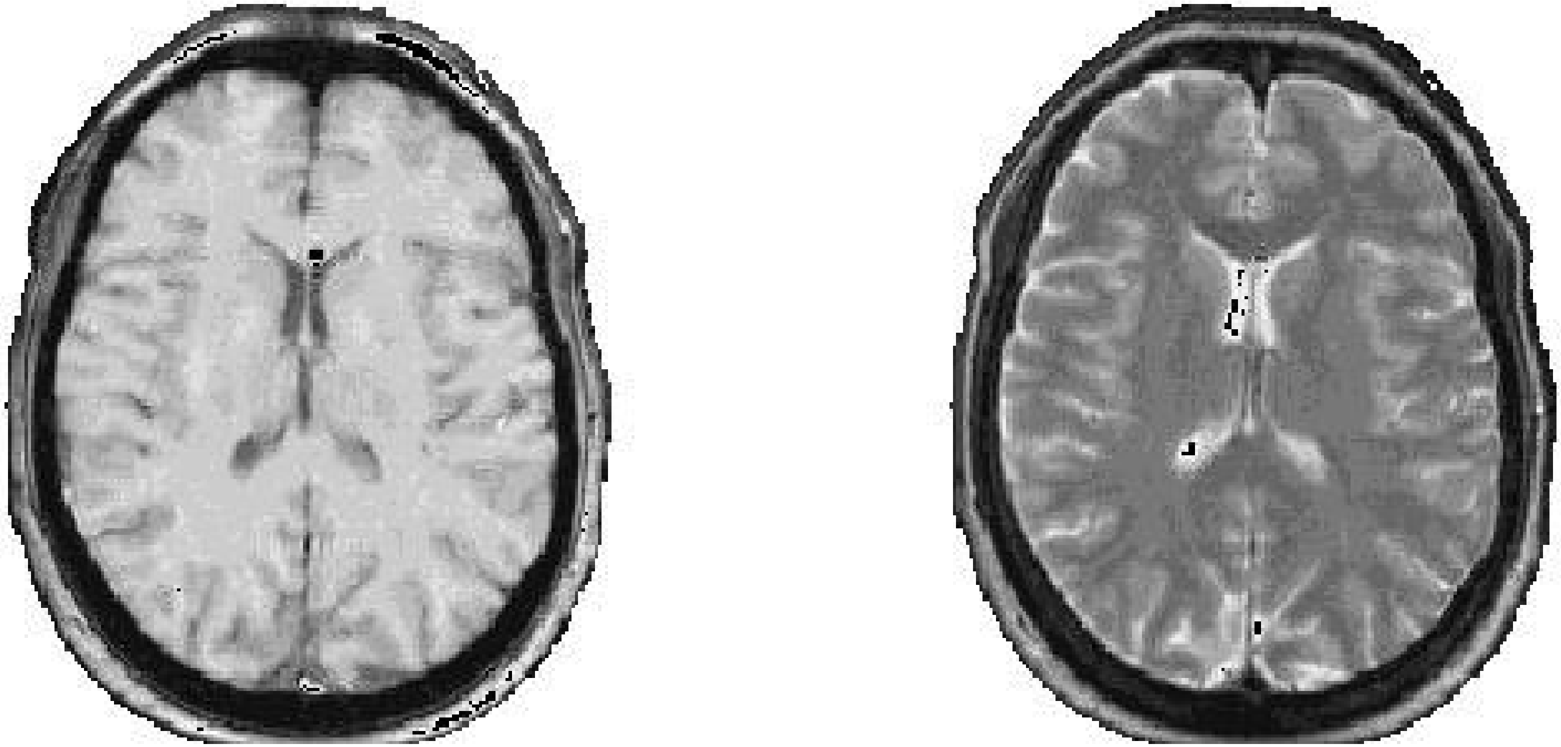
Application/Results

6 component model

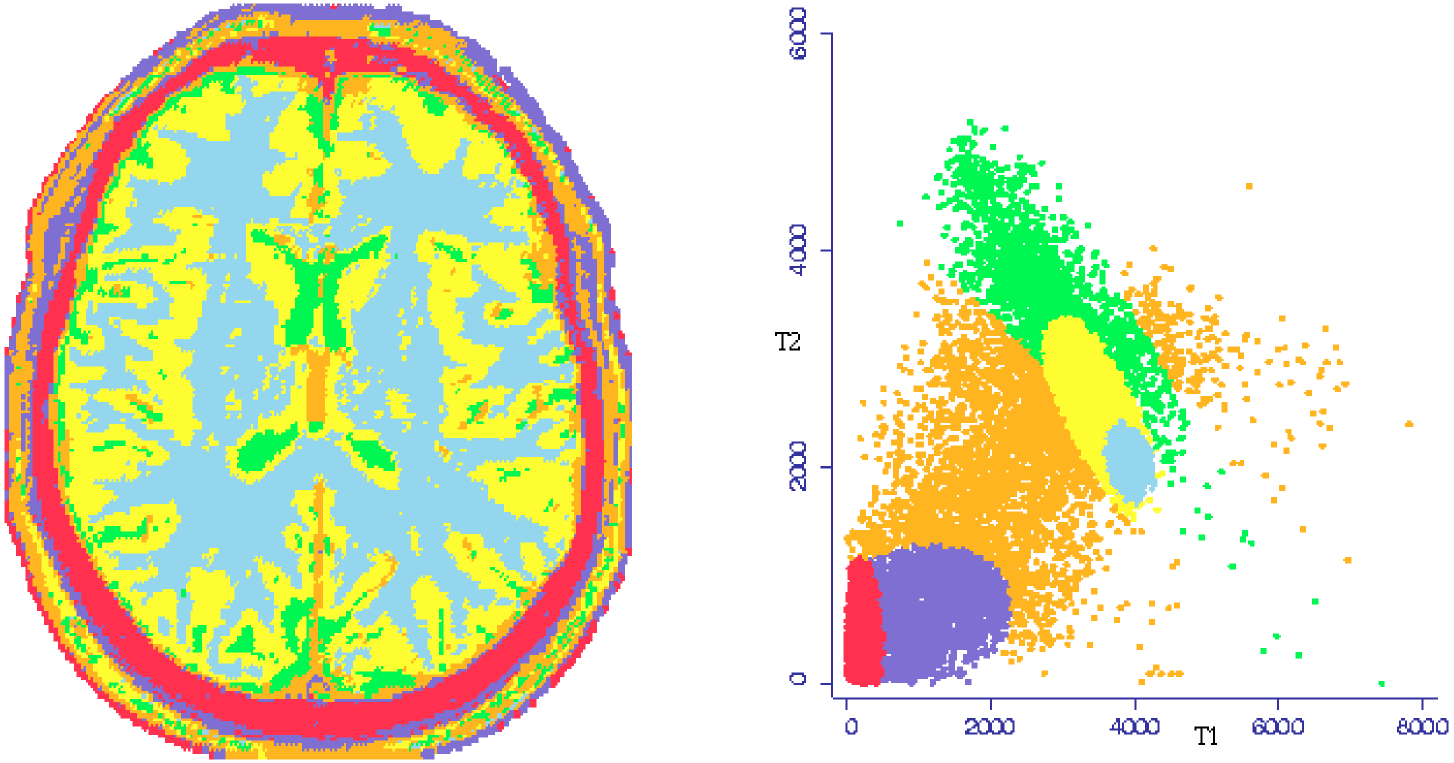
- CSF
- White matter
- Grey matter
- Skull type 1
- Skull type 2
- Outlier component (fixed mean and large variance)

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)

Case Study:

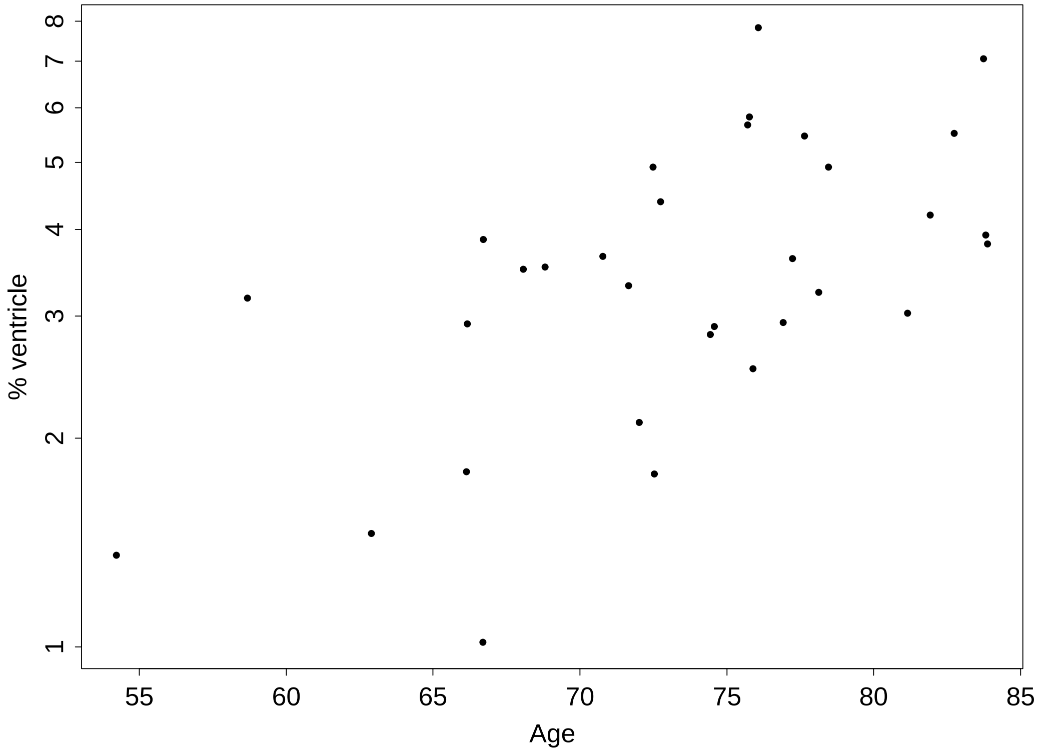
Structural MRI of Ageing and Dementia

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

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.

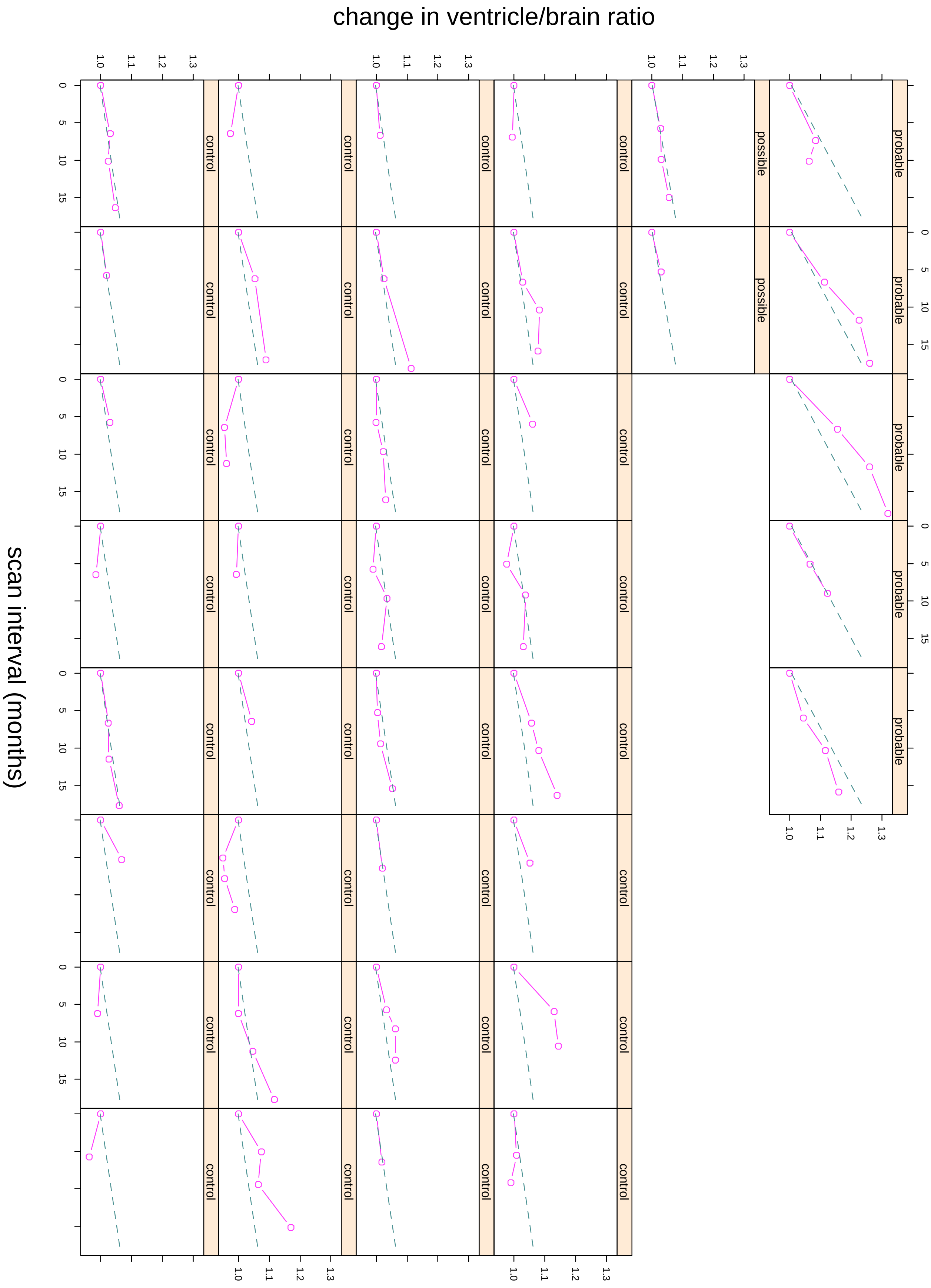
How large should n be?

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.

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

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)?

Case Study:

Magnetic Resonance Imaging of Brain Function

Joint work with Jonathan Marchini.

Data, background and advice provided by Stephen Smith (Oxford Centre for Functional Magnetic Resonance Imaging of the Brain).

‘Functional’ Imaging

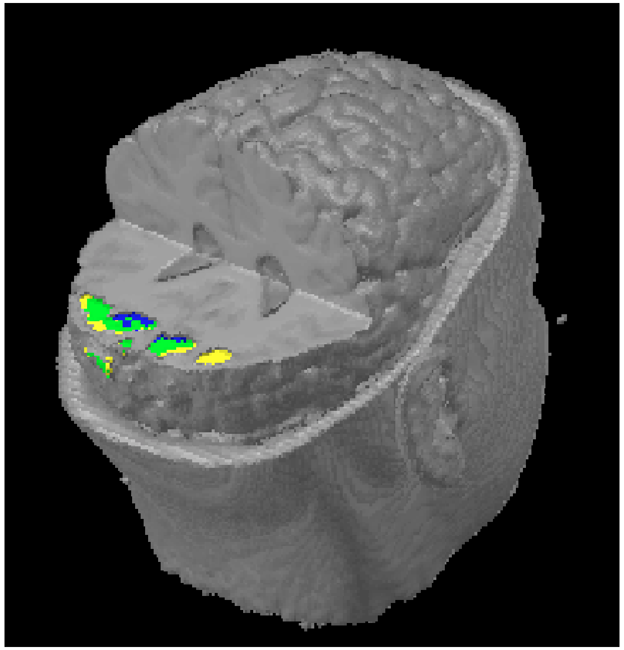
Functional PET and MRI are used for studies of brain function: give a subject a task and see which area(s) of the brain ‘light up’.

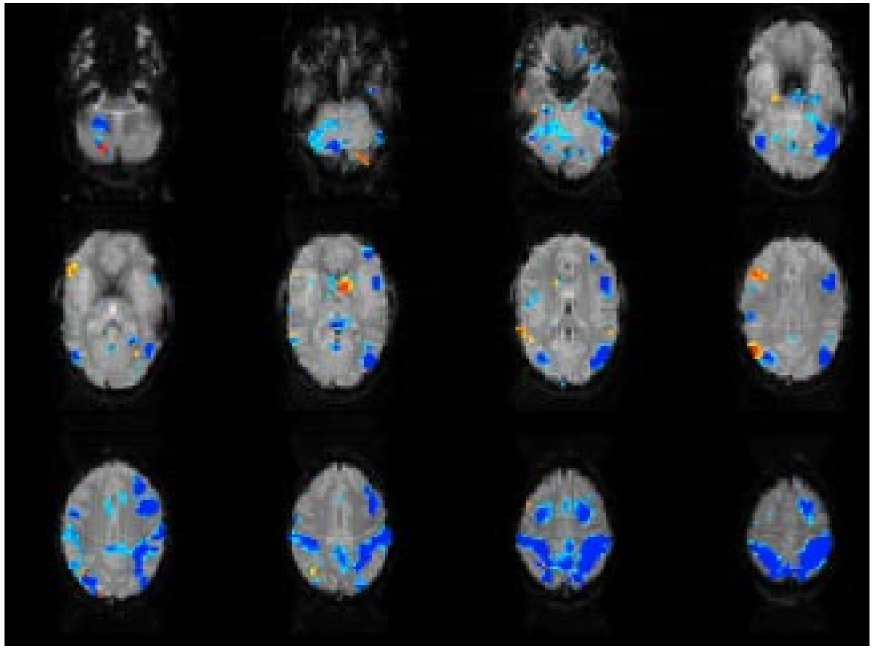
Functional studies were done with PET in the late 1980s and early 1990s, now fMRI is becoming possible (needs powerful magnets—that in Oxford is 3 Tesla). Down to $1 \times 1 \times 3$ mm voxels.

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.

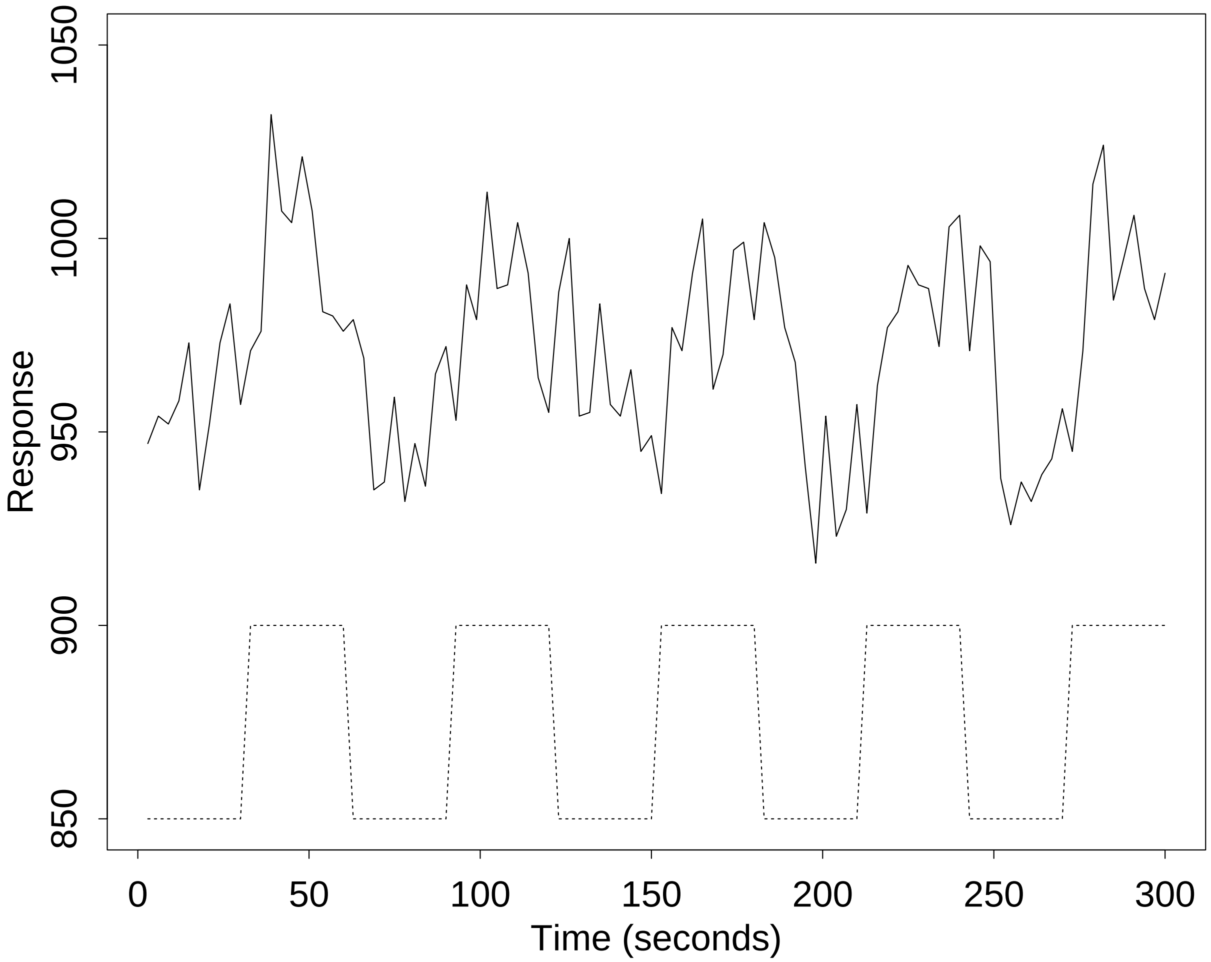
fMRI has a higher spatial resolution, and temporal resolution of around one second. So most commonly stimuli are applied for a period of about 30 secs, images taken around every 3 secs, with several repeats of the stimulus being available for one subject.

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





A Look at Some Data



A real response (solid line) in an area of activation from the visual experiment. The periodic boxcar shape of the visual stimulus is shown below.

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.

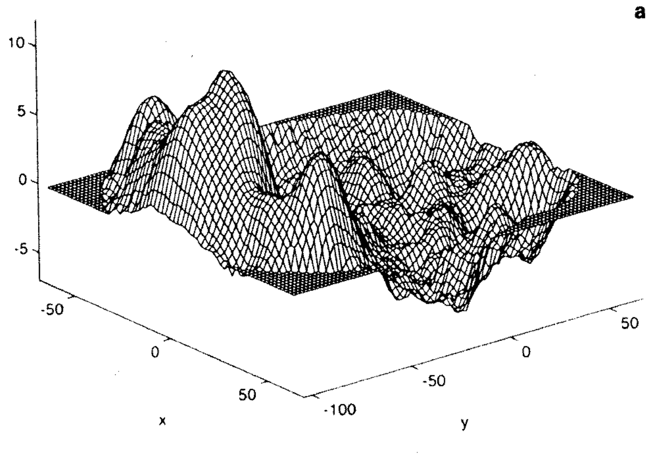
The t-statistic is in PET of a comparison between states over a number of subjects, voxel by voxel. Thus the numerator is an average over subjects of the difference in response in the two states, and the denominator is an estimate of the standard error of the numerator.

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

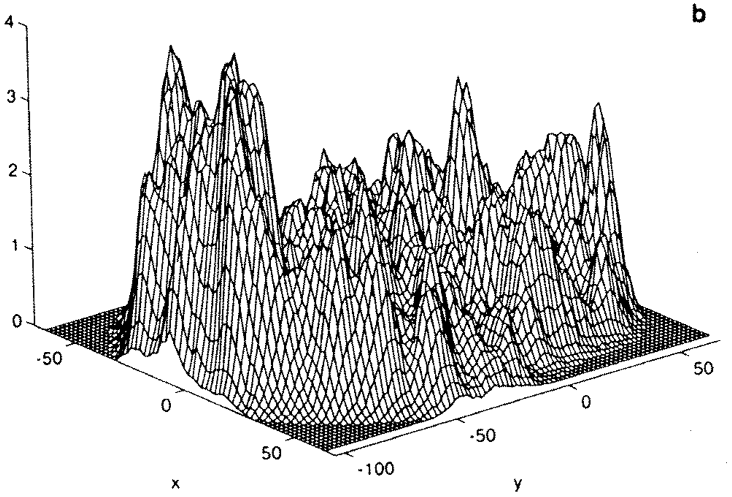
The details also vary widely between releases of and users of the programs.

Example PET Statistics Images

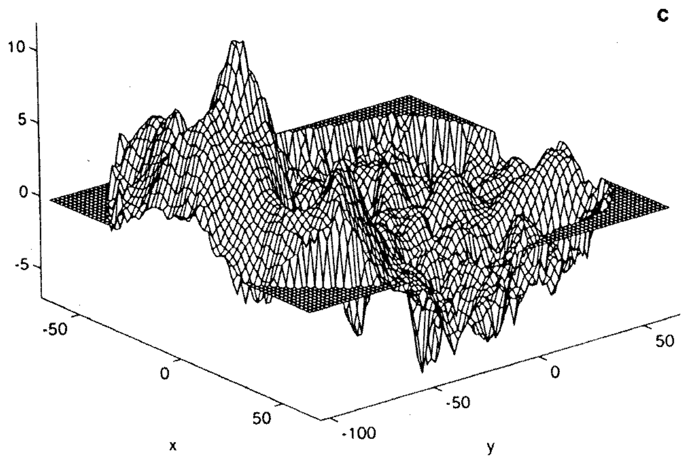
From Holmes *et al* (1996).



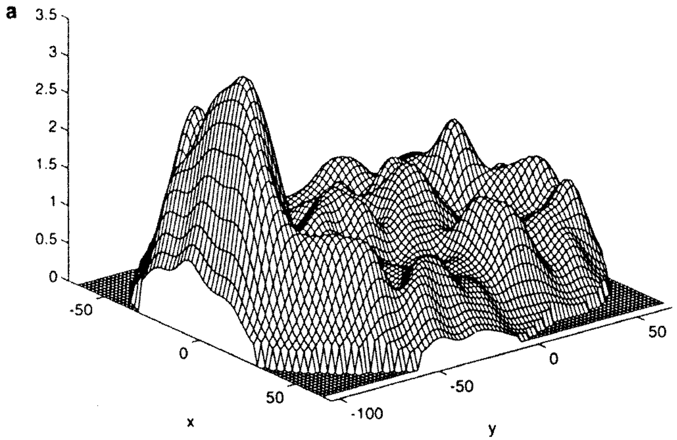
Mean difference image.



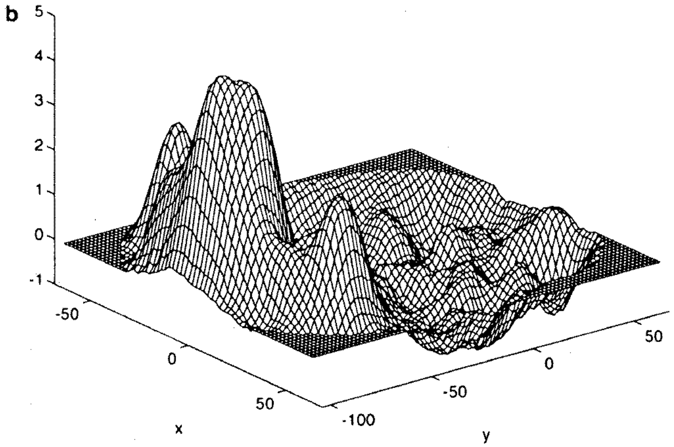
Voxel-wise variance image.



Voxel-wise t–statistic image.



Smoothed variance image.



Resulting t–statistic image.

The GLM approach

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

The signal is usually filtered by a matrix S, so the model becomes

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

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$(\hat{\beta})$? (For a t-test of some component being non-zero.)

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:

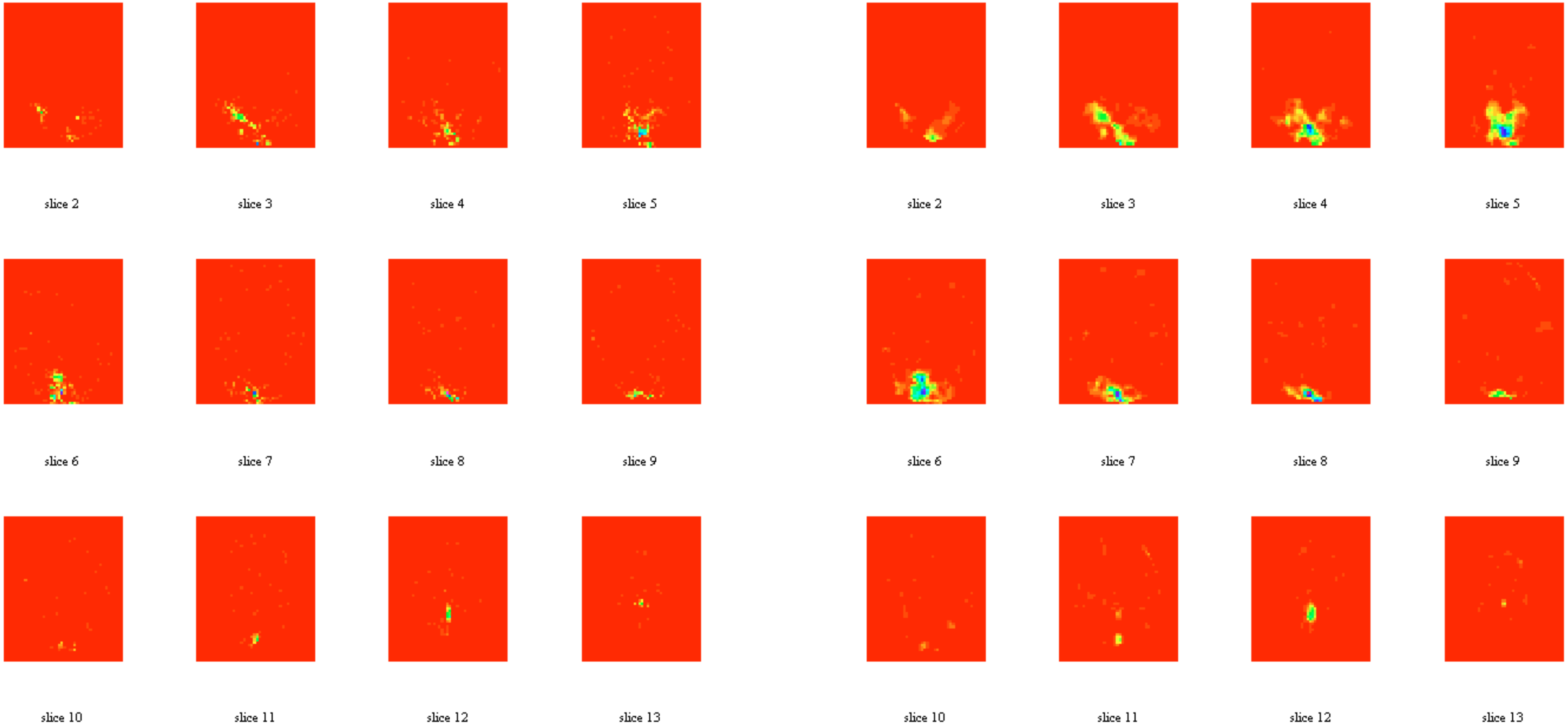
1. (High) level crossings of Gaussian stochastic processes
   (Worsley *et al*): *Euler characteristics*.
2. Randomization-based analysis (Holmes *et al*) across replications.
3. Variability within the time series at a voxel.

fMRI Example

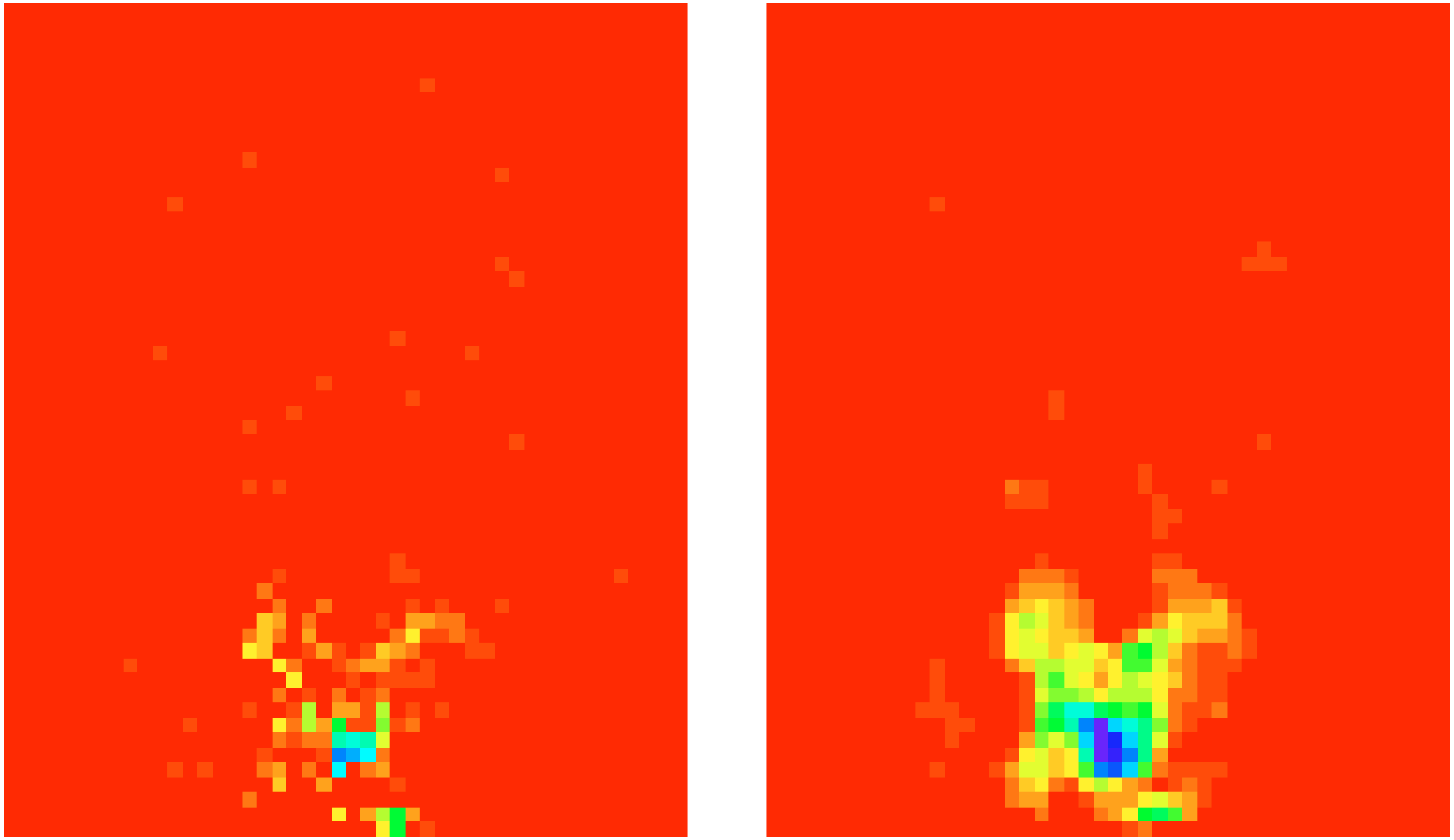
Data on $64 \times 64 \times 14$ grid of voxels. (Illustrations omit top and bottom slices and areas outside the brain, all of which show considerable activity, probably due to registration effects.)

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.

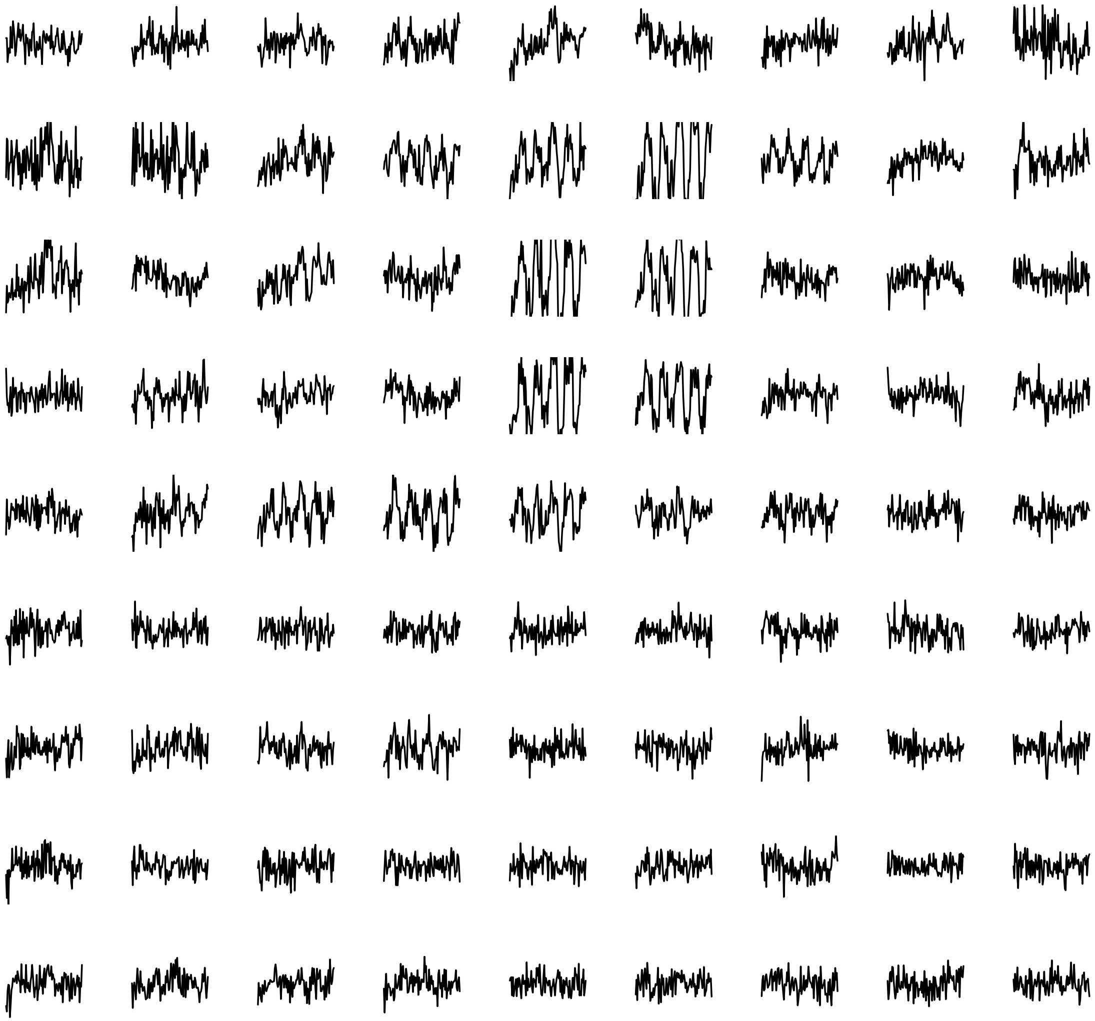


SPM99 t–statistic images, with spatial smoothing on the right



Slice 5, with spatial smoothing on the right

A Closer Look at Some Data

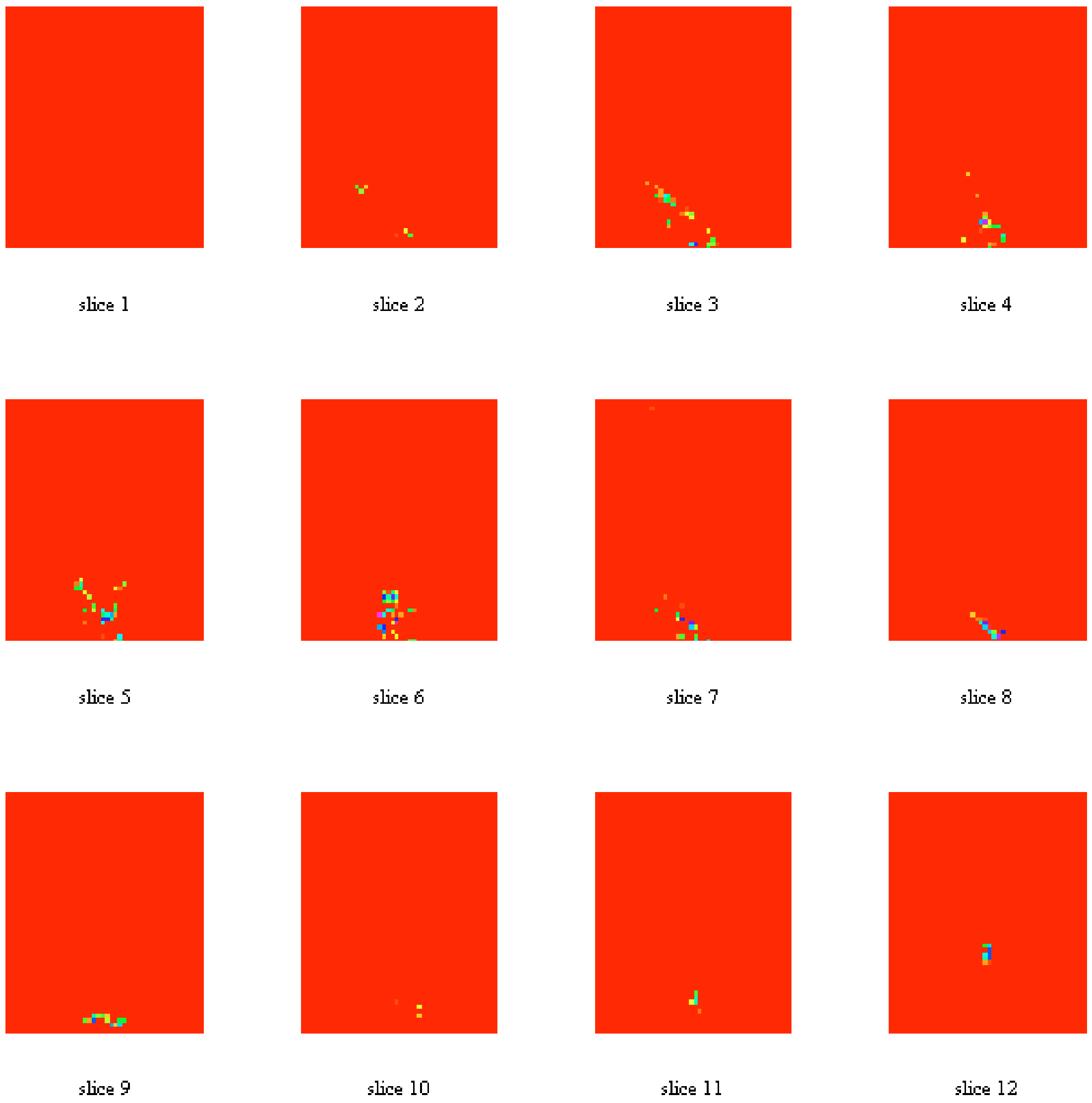


A 9×9 grid in an area of slice 5 containing activation.

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 / R on ca 400MHz PC).



slice 1

slice 2

slice 3

slice 4

slice 5

slice 6

slice 7

slice 8

slice 9

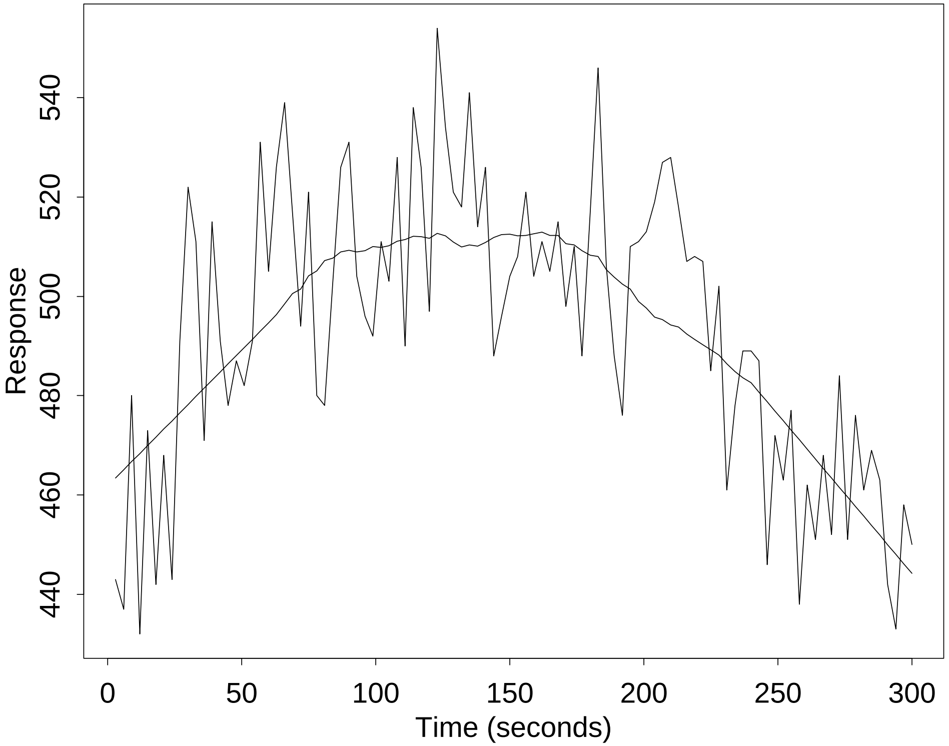
slice 10

slice 11

slice 12

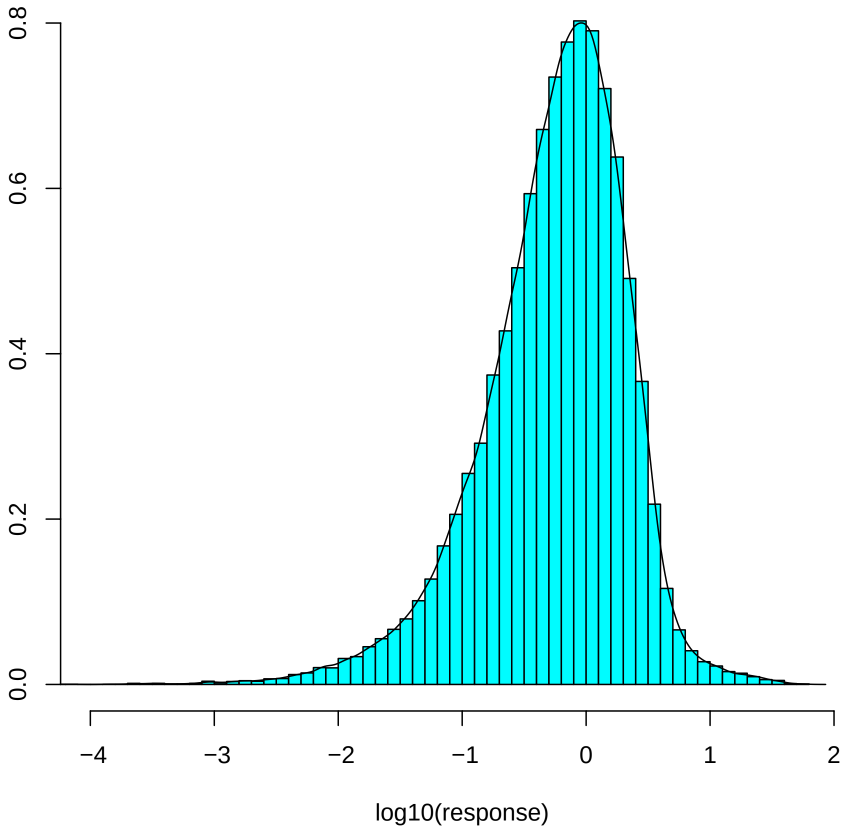
Log abs filtered response, with small values coloured as background (red). Threshold for display is $p < 10^{-5}$ (and there are ca 20,000 voxels inside the brain here).

Trend-removal



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

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

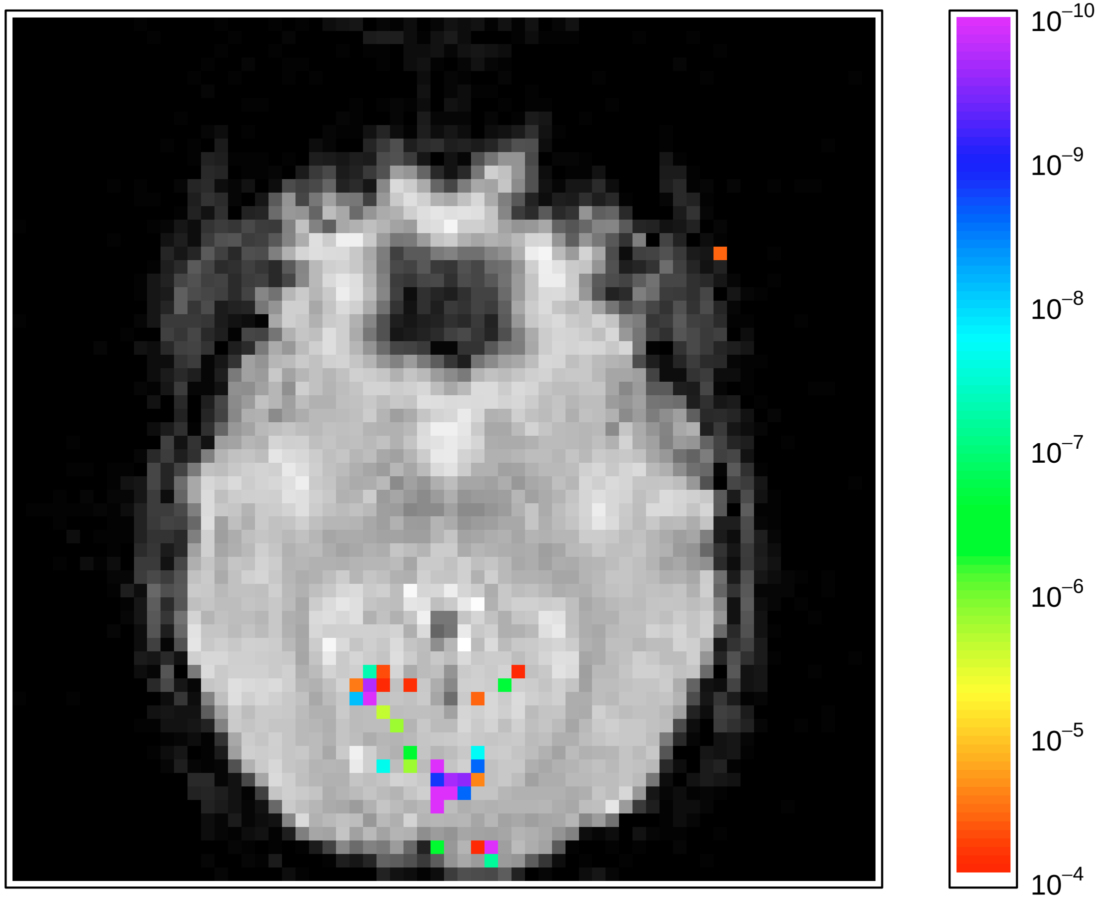


Histogram of log filtered response, for an image with activation.

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

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.

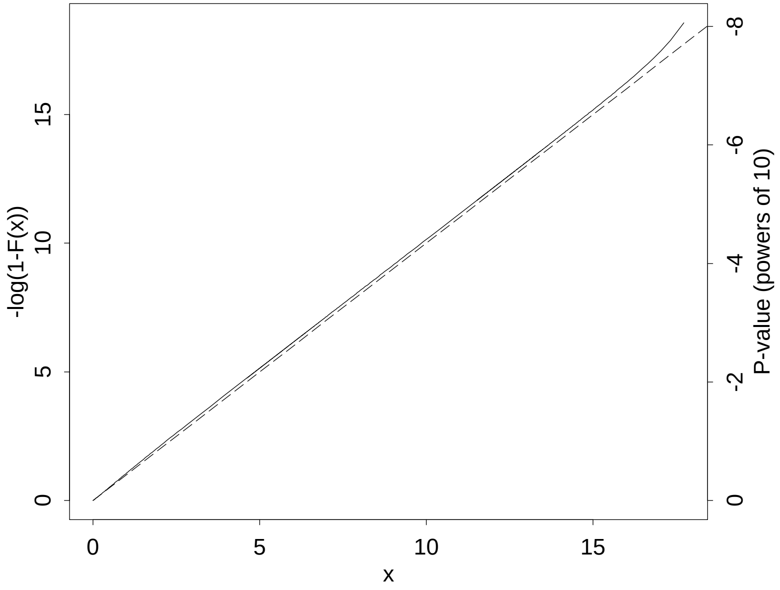


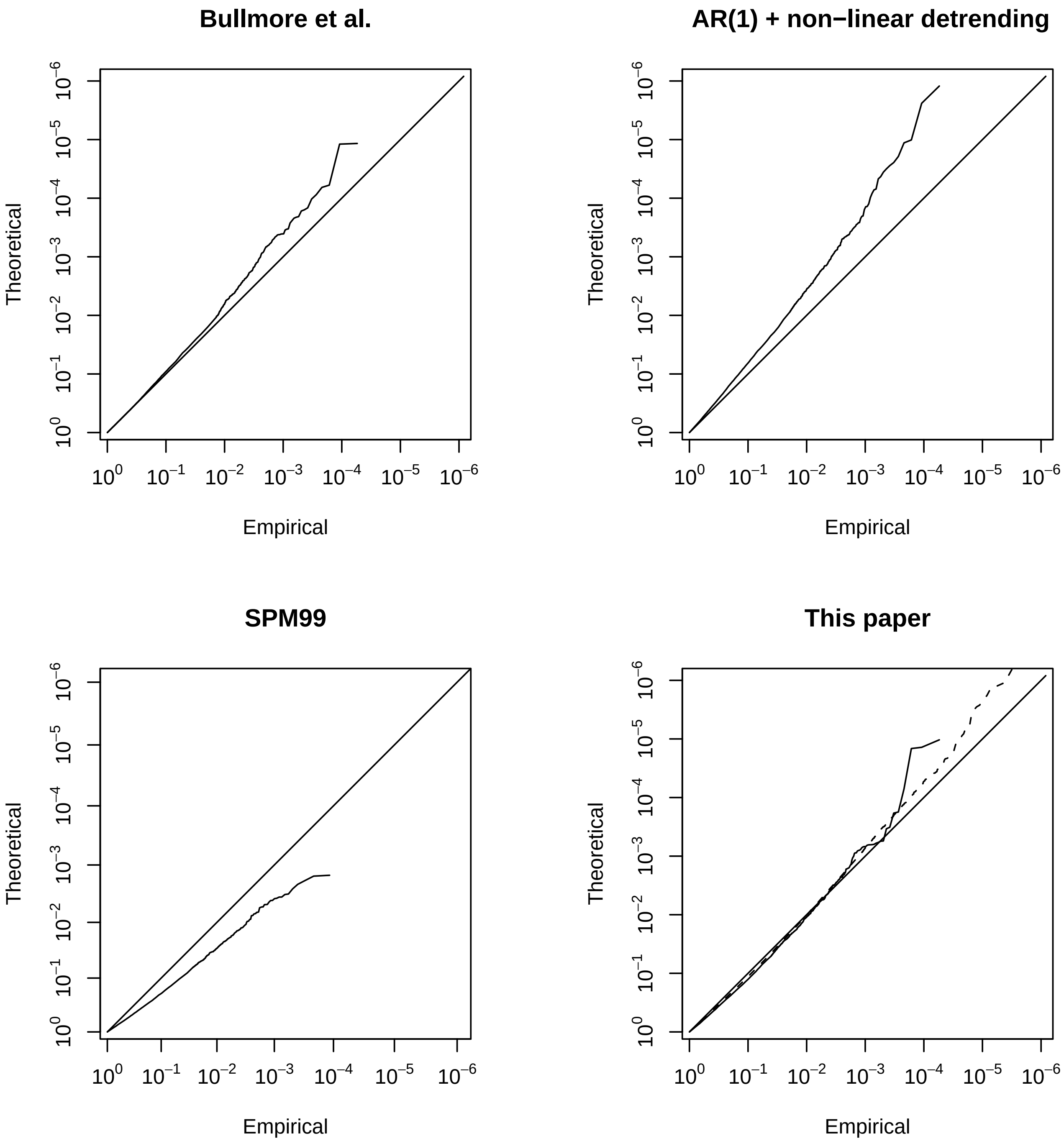
Calibration

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}