Maths for Brain Imaging: Lecture 10

W.D. Penny
Wellcome Department of Imaging Neuroscience,
University College, London WC1N 3BG.

December 13, 2006

1 Contents

- Central Limit Theorem
- Independent Component Analysis
- Application: Removing EEG artefacts
- Discriminant analysis
- Application: Estimating perceptual state from fMRI

2 Central Limit Theorem

Given n samples from *any* probability distribution, the distribution of the sample mean becomes Gaussian as $n \to \infty$. For proof see [6]. A caveat is that the sample variance must be finite.

More formally, if $y_1, y_2, ..., y_n$ are Independent and Identically Distributed (IID) random variables with $E[y_i] = \mu$ and $Var[y_i] = \sigma^2 < \infty$ and

$$\begin{aligned} \bar{y}_n &= \frac{1}{n}\sum_{i=1}^{n} y_i \\ u_n &= \sqrt{n}\left(\frac{\bar{y}_n - \mu}{\sigma}\right) \end{aligned} \tag{1}$$

then $p(u_n)$ converges to a standard Gaussian density as $n \to \infty$. This is the Central Limit Theorem (CLT).

The CLT can be extended to Independent and Non-Identically Disributed (IND) random variables, as long as $E[y_i]$ and $Var[y_i]$ are finite.

This implies that if you have a Gaussian observation

then its a 'mixture' (average or weighted average) of non-Gaussian signals. ICA attempts to find the underlying signals by looking for projections of the observations that are most non-Gaussian. This is implemented either (i) informally, by maximising an index of non-Gaussianity such as kurtosis, $E[(y_i - \mu)^4]$ (a Gaussian has zero kurtosis) or (ii) formally by specifying a probability model where the sources are non-Gaussian.

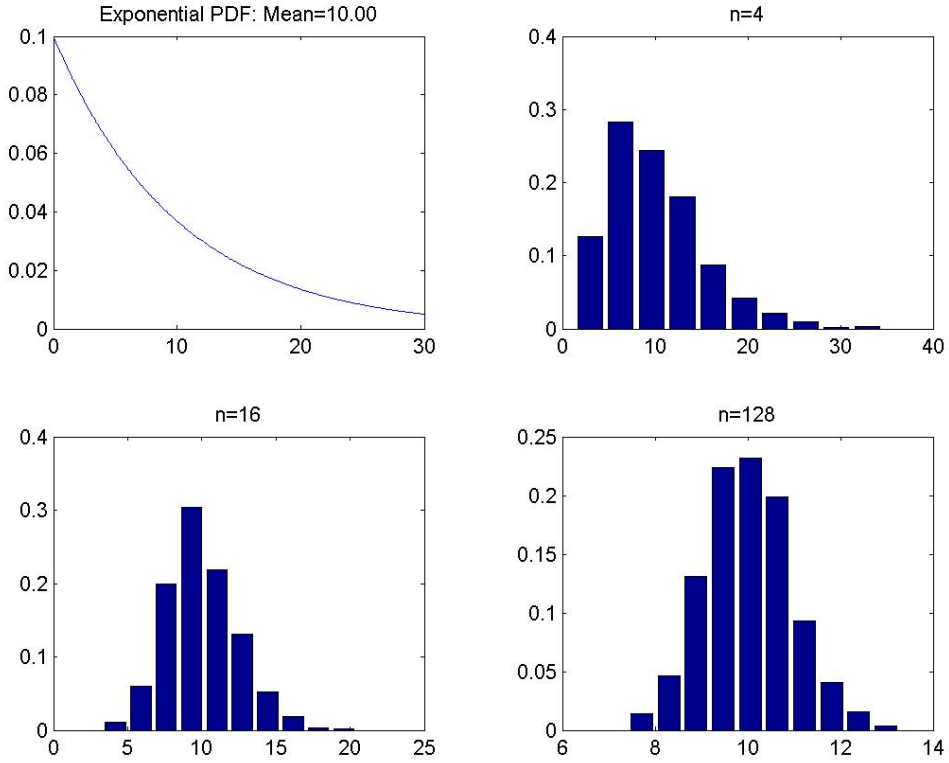


Figure 1: Take n samples from an exponential PDF, compute the sample mean. Do this multiple times to get an empirical estimate of the distribution of the sample mean. As n increases, the distribution becomes Gaussian.

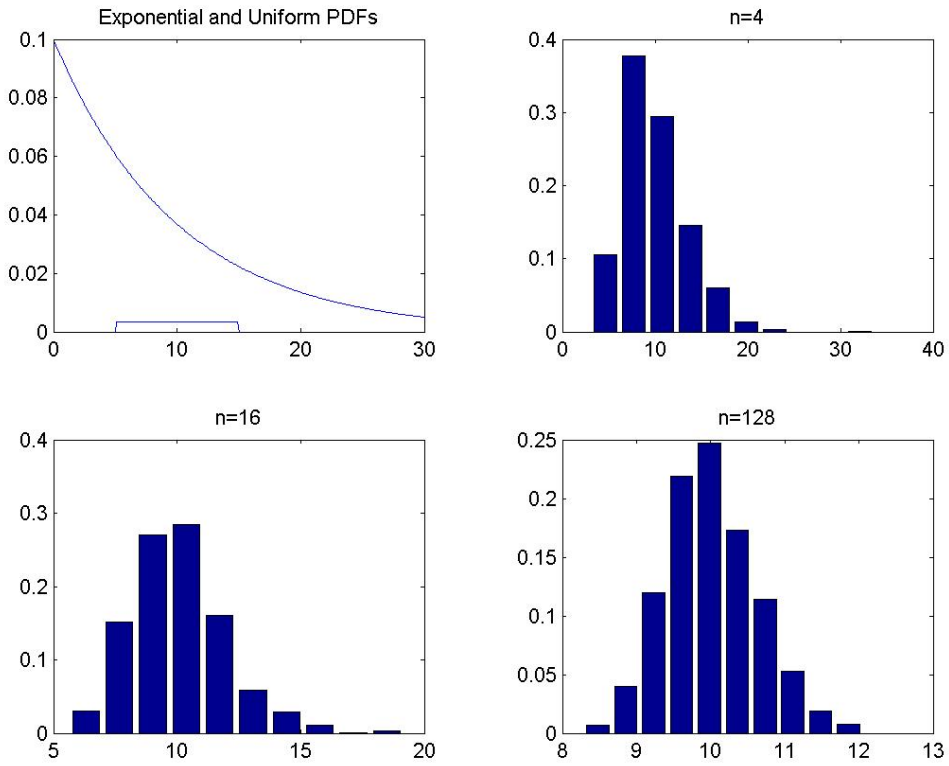


Figure 2: Take $n/2$ samples from an exponential PDF and $n/2$ from a uniform PDF, compute the sample mean. Do this multiple times to get an empirical estimate of the distribution of the sample mean. As n increases, the distribution becomes Gaussian.

3 Independent Component Analysis

In Independent Component Analysis (ICA) an M-dimensional vector observation y is modelled as

$$y = X\beta \tag{2}$$

where X is an unknown mixing matrix and β an unknown M-dimensional source vector. The matrix X is therefore $M \times M$. If we know $p(\beta)$, then using the method of transforming probability densities we can write the likelihood of an observation as

$$p(y) = \frac{p(\beta)}{|\det X|} \tag{3}$$

The determinant measures the volume of a matrix. So Eq 3. takes into account volumetric changes in the transformation, so that probability mass is preserved as we transform β into y.

ICA assumes the sources to be Independent (this is

the I in ICA)

$$p(\beta) = \prod_{i=1}^{M} p_s(\beta_i) \tag{4}$$

We can therefore write the likelihood as

$$p(y) = \frac{\Pi_{i=1}^{M} p_s(\beta_i)}{|\det X|} \tag{5}$$

The log-likelihood is then given by

$$\log p(y) = -\log|\det X| + \sum_{i=1}^{M} \log p_s(\beta_i) \tag{6}$$

We can write the unknown sources as

$$\begin{aligned} \beta &= X^{-1}y \\ &= Ay \end{aligned} \tag{7}$$

where $A = X^{-1}$ is the inverse mixing matrix. We can also write $\beta_i = \sum_{j=1}^{M} A_{ij}y_j$ and express the log-likelihood as

$$\log p(y) = \log|\det A| + \sum_{i=1}^{M} \log p_s(\sum_{j=1}^{M} A_{ij}y_j) \tag{8}$$

The log-likelihood is now a function of the data and the inverse mixing matrix.

If we have $n = 1..N$ independent samples of data, Y, the likelihood is

$$\log p(Y) = N \log |\det A| + \sum_{n=1}^{N} \sum_{i=1}^{M} \log p_s(\sum_{j=1}^{M} A_{ij} y_{nj}) \quad (9)$$

We can find A by giving this function to any optimisation algorithm. As elements of A become co-linear $|\det A| \rightarrow 0$, and the likelihood reduces. Maximising the likelihood therefore encourages sources to be different (via the first term) and encourages them to be similar to p_s ie. non-Gaussian (via the second term).

3.1 Source densities

Different ICA models result from different assumptions about the source densities $p_s(\beta_i)$. One possibility is the generalised exponential family. Another is the 'inverse

cosh' density

$$\begin{aligned} p_s(\beta_i) &= \frac{1}{cosh(\beta_i)} \\ &= \frac{1}{\exp \beta_i + \exp -\beta_i} \end{aligned} \tag{10}$$

This latter choice gives rise to the original 'Infomax' algorithm [1].

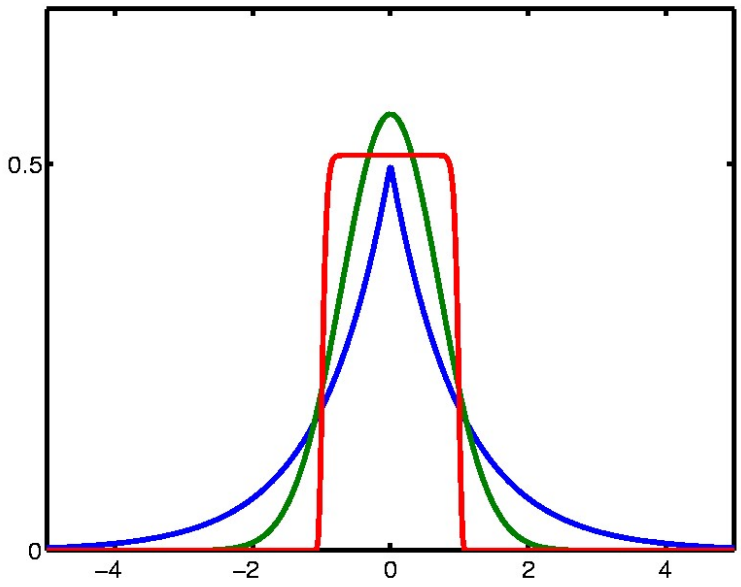


Figure 3: Generalised exponential densities $p_s(\beta_i) \propto \exp\left(-|\frac{\beta_i}{\sigma}|^R\right)$with $R = 1$ (Blue, 'Laplacian density'), $R = 2$ (Green, 'Gaussian density'), $R > 20$ (Red, 'Uniform density'). The parameter σ defines the width of the density.

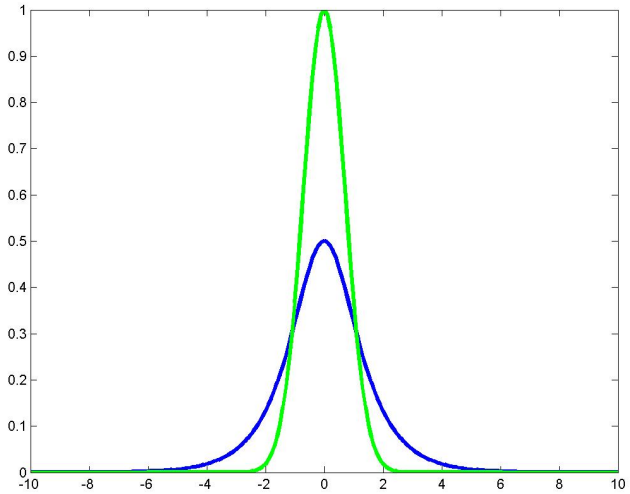
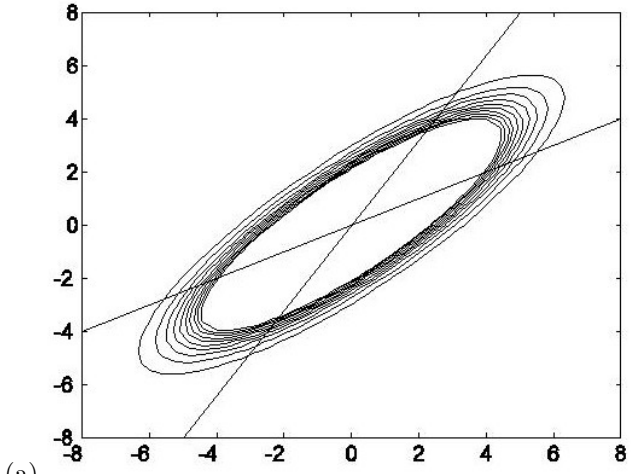
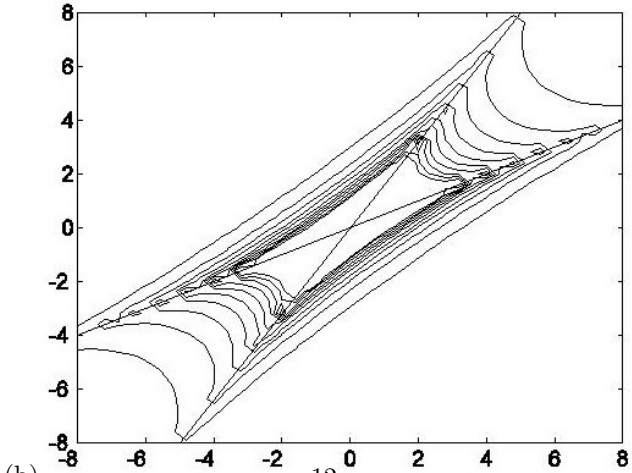


Figure 4: Inverse Cosh, $\frac{1}{\exp \beta_i + \exp -\beta_i}$ (Blue) and Gaussian, $\exp -\beta_i^2$ (Green)



(a)



(b)

Figure 5: Probability contours, $p(y)$, from 2D-ICA models with (a) Gaussian sources and (b) Heavy-tailed sources. The mixing matrices X are the same.

3.2 Removing EEG artefacts

Jung et al. [5] use ICA to remove artefacts from EEG data recorded from 20 scalp electrodes placed according to the 10/20 system and 2 EOG electrodes, all references to the left mastoid. The sampling rate was 256Hz. An ICA decomposition was implemented by applying an extended Infomax algorithm to 10-second EEG epochs to produce sources with time series that are maximally independent.

This artefact removal method compares favourably to PCA and filtering approaches, and approaches for eye-movement correction based on dipole models and regression [5]. It has been incorporated in the EEGLAB available from `http://sccn.ucsd.edu/eeglab/`.

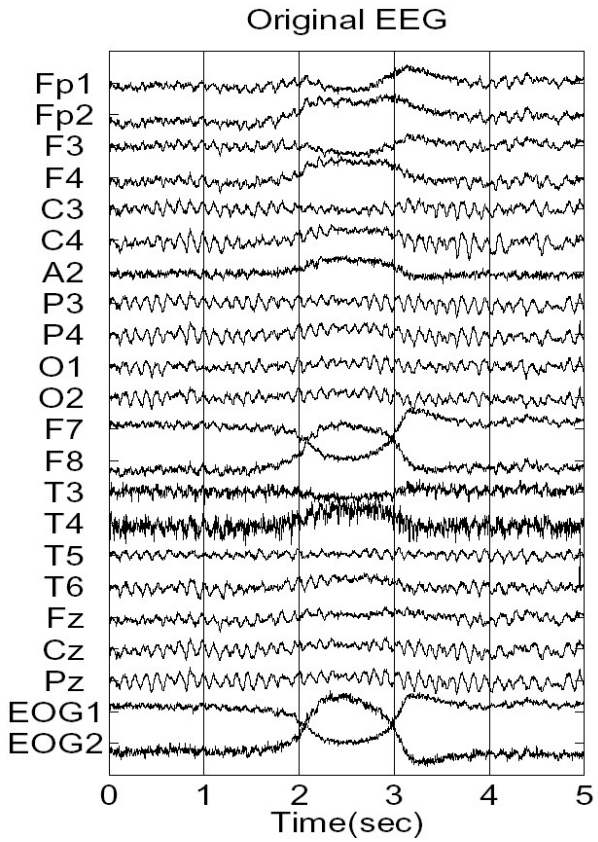


Figure 6: Original 5-second EEG record, containing prominent slow eye-movement (seconds 2 to 4).

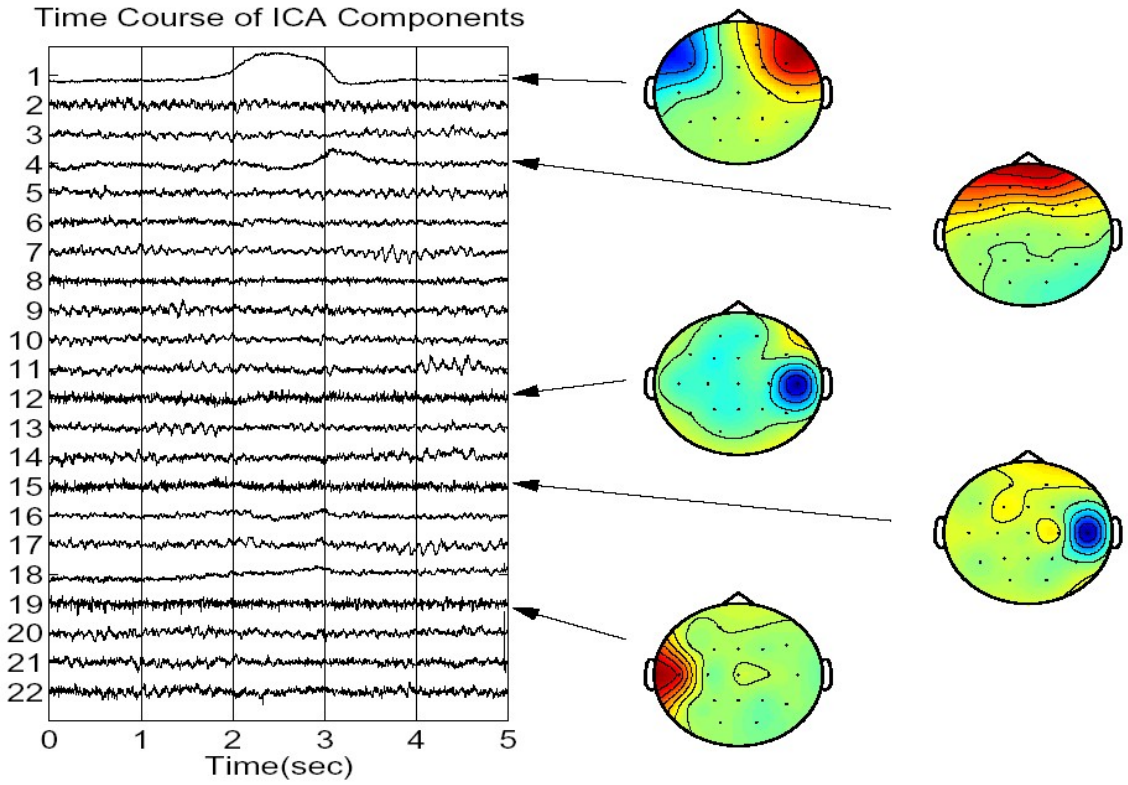


Figure 7: Left panel: Time course of source estimates β_{ni} for n=1..N samples (N=5 x 256), and i=1..22 sources. Right panel: Spatial topographies (rows of mixing matrix X) for 5 selected components. The top two components account for eye movement and the bottom three for muscle activity over fronto-temporal regions.

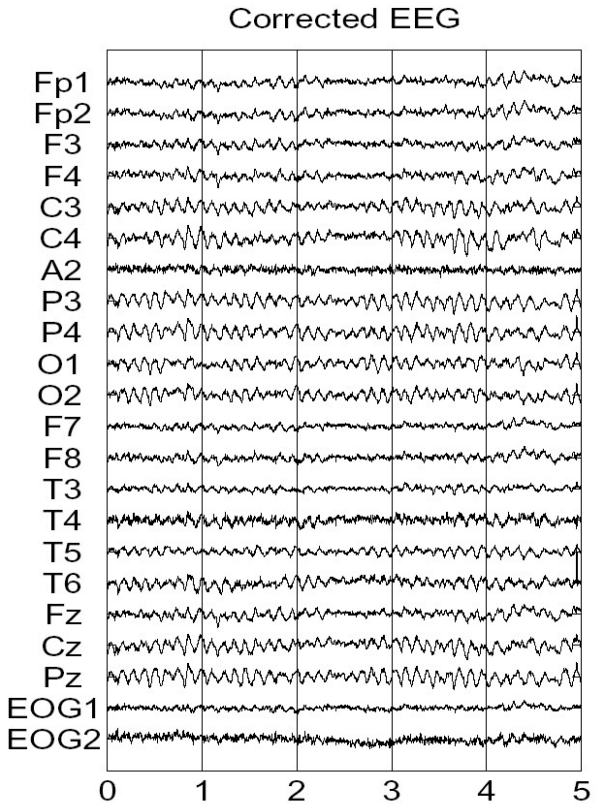


Figure 8: Corrected EEG formed by subtracting five selected components from original data. This data is free from EOG and muscle artifacts. We can now see activity in T3/T4 that was previously masked by muscle artifact.

4 Discriminant analysis

4.1 Linear decision boundary

The aim of discriminant analysis is to estimate class label $y = \{1, 2\}$ given multivariate data x. This could be eg. $y = 1$ for patients and $y = 2$ for controls. One approach is to use labelled data to form a likelihood model for each class, $p(x|y)$. New data points are then assigned to the class with the highest likelihood. Another way of saying this is to form the Likelihood Ratio (LR)

$$LR_{12} = \frac{p(x|y = 1)}{p(x|y = 2)} \tag{11}$$

and assign to class 1, if LR_{12} is greater than one. According to the Neymann-Pearson Lemma (see eg. [3]) this test has the highest sensitivity, for any given level of specificity. Any monotonic function of LR_{12} will provide as good a test as the likelihood, and the logarithm is often used.

If, additionally, we have prior probabilities for each

category, $p(y)$ then the optimal decision is to assign to the class with the highest posterior probability. For class 1 we have

$$p(y=1|x) = \frac{p(x|y=1)p(y=1)}{p(x|y=1)p(y=1) + p(x|y=2)p(y=2)} \tag{12}$$

For equal priors this reduces to an LR test.

A simple likelihood model for each class is a Gaussian. The above posterior probability is then the same as the 'responsibilty' in a Gaussian mixture model (see last lecture). It can be re-written as

$$\begin{aligned} p(y=1|x) &= \frac{1}{1 + \frac{p(x|y=2)p(y=2)}{p(x|y=1)p(y=1)}} \\ &= \frac{1}{1 + exp(-a)} \\ &= g(a) \end{aligned} \tag{13}$$

where $g(a)$ is the 'sigmoid' or 'logistic' function and

$$a = \log\left(\frac{p(x|y=1)p(y=1)}{p(x|y=2)p(y=2)}\right) \tag{14}$$

For Gaussians with *equal covariances* $\Sigma_1 = \Sigma_2 = \Sigma$ we have

$$\begin{aligned} \log p(x|y=1) &= \frac{d}{2}\log 2\pi + \frac{1}{2}\log|\Sigma| \\ &- \frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) \\ \log p(x|y=2) &= \frac{d}{2}\log 2\pi + \frac{1}{2}\log|\Sigma| \\ &- \frac{1}{2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2) \end{aligned} \tag{15}$$

This gives

$$a = w^T x + w_0 \tag{16}$$

where

$$w = \Sigma^{-1}(\mu_1 - \mu_2) \tag{17}$$

$$w_0 = -\frac{1}{2}\mu_1^T\Sigma\mu_1 + \frac{1}{2}\mu_2^T\Sigma\mu_2 + \log\frac{p(y=1)}{p(y=2)}$$

This approach is known as logistic discrimination or logistic classification. The decision boundary is given by $a = 0$.

4.2 Nonlinear decision boundary

If the Gaussians do not have equal covariance then the decision boundary becomes quadratic. If Gaussians are not good models of the class probability densities then another approach is required eg. Nearest Neighbour classifiers, or Multi-Layer Perceptrons (MLPs). An MLP comprises nested logistic functions.

A two-layer MLP is given by

$$\begin{aligned} p(y=1|x) &= g\left(\sum_{h=1}^{H} w_h^{(2)} z_h\right) \\ z_h &= g\left(\sum_{d=1}^{D} w_{hd}^{(1)} x_d)\right) \end{aligned} \tag{18}$$

with D is the dimension of the input x, H is the number of 'hidden units' in the 'first layer', and z_h is the output of the hth unit. Superscripts 1 and 2 denote 1st and 2nd layer weights. This allows for classification using arbitrary decision boundaries. There is no closed form for the parameters w, but they can be estimated using an optimisation algorithm as described in [2]. This is an example of an Artificial Neural Network (ANN).

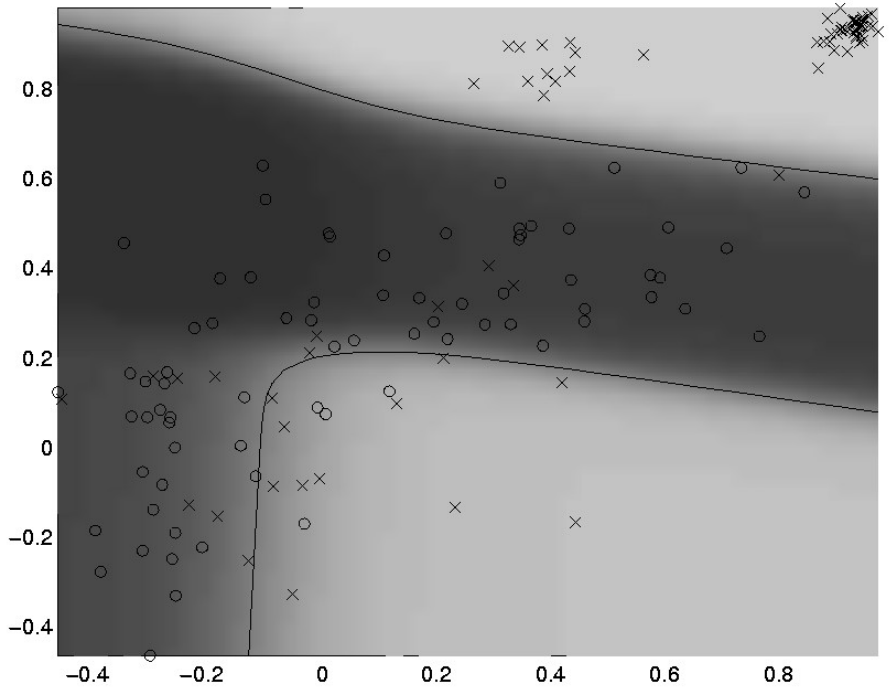


Figure 9: Tremor data. Crosses represent data points from patients $y = 1$, circles data points from normal subjects, $y = 2$. The solid line shows the overall decision boundary formed by an MLP with three hidden units. The shade of gray codes the output, $p(y = 1|x)$ and the features, x, are from autoregressive modelling of EMG data.

5 Estimating perceptual state from fMRI

Haynes and Rees [4] used Linear Discriminant Analysis (LDA) to classify perceptual state during binocular rivalry from fMRI data.

Retinotopic mapping and functional localisers (reversing checkerboard stimuli) were used to identify the V1, V2, V3 and V5 regions of visual cortex. The 50 most visually responsive voxels in each region were then selected for subsequent analysis.

Subjects then viewed rivalrous stimuli, and pressed buttons to indicate perceptual state, $y_t = 1$ for red percept and $y_t = 2$ for blue percept. Activity in selected voxels x_t were then used to estimate y_t. The labels y_t were time-shifted to accomodate the delay in the hemodynamic response.

Estimates of perceptual state were then formed using

$$\begin{aligned} \hat{y}_t &= w^T x_t \\ w &= \Sigma^{-1}(\mu_{red} - \mu_{blue}) \end{aligned} \tag{19}$$

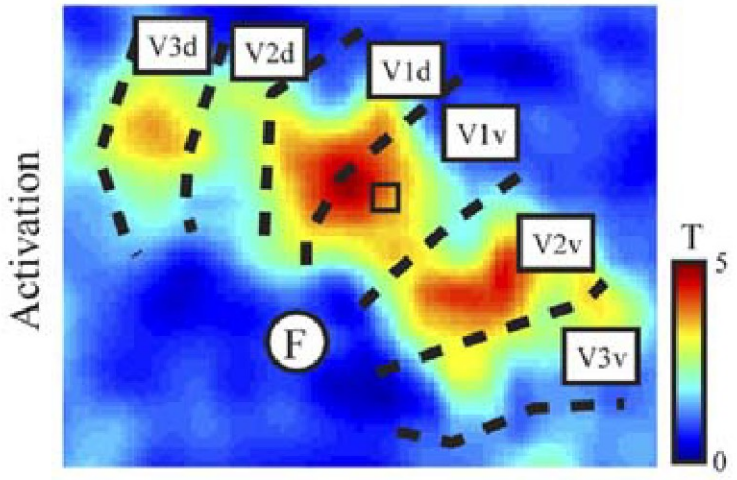


Figure 10: Functional localiser (4Hz reversing checkerboard) is used to select 50 visually responsive voxels in each region.

where Σ is the within group sample covariance (estimated from both red and blue fMRI samples) and m_{red} and m_{blue} are the mean fMRI vectors for each condition. These estimates were then time-shifted, by convolving with a 'Canonical HRF' before comparison with true values. Cross-validation was used to assess accuracy.

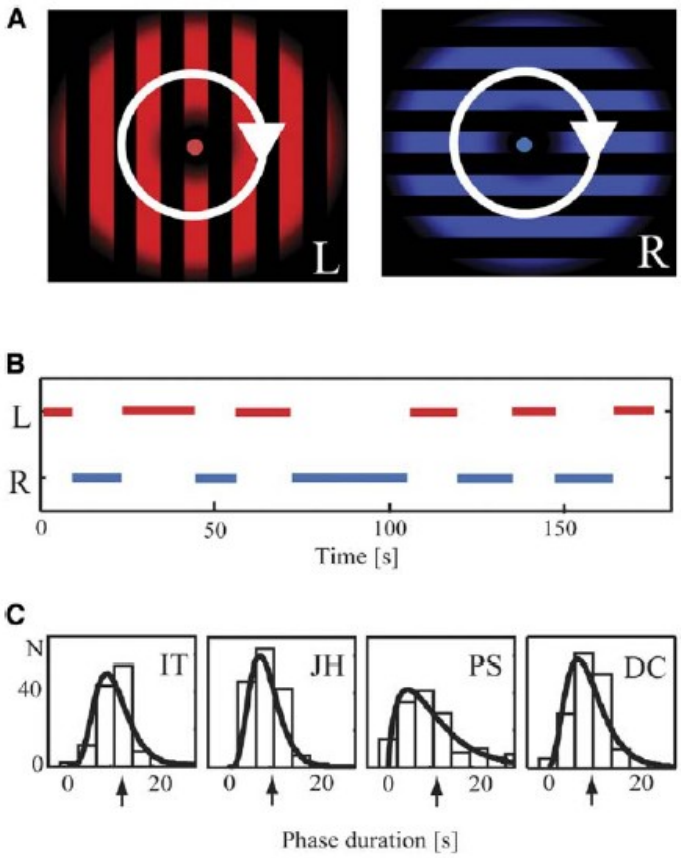


Figure 11: A: Superimposed gratings viewed through red/blue filtering glasses. B: Subjects pressed buttons indicating perceptual state. C: Percept durations for four subjects.

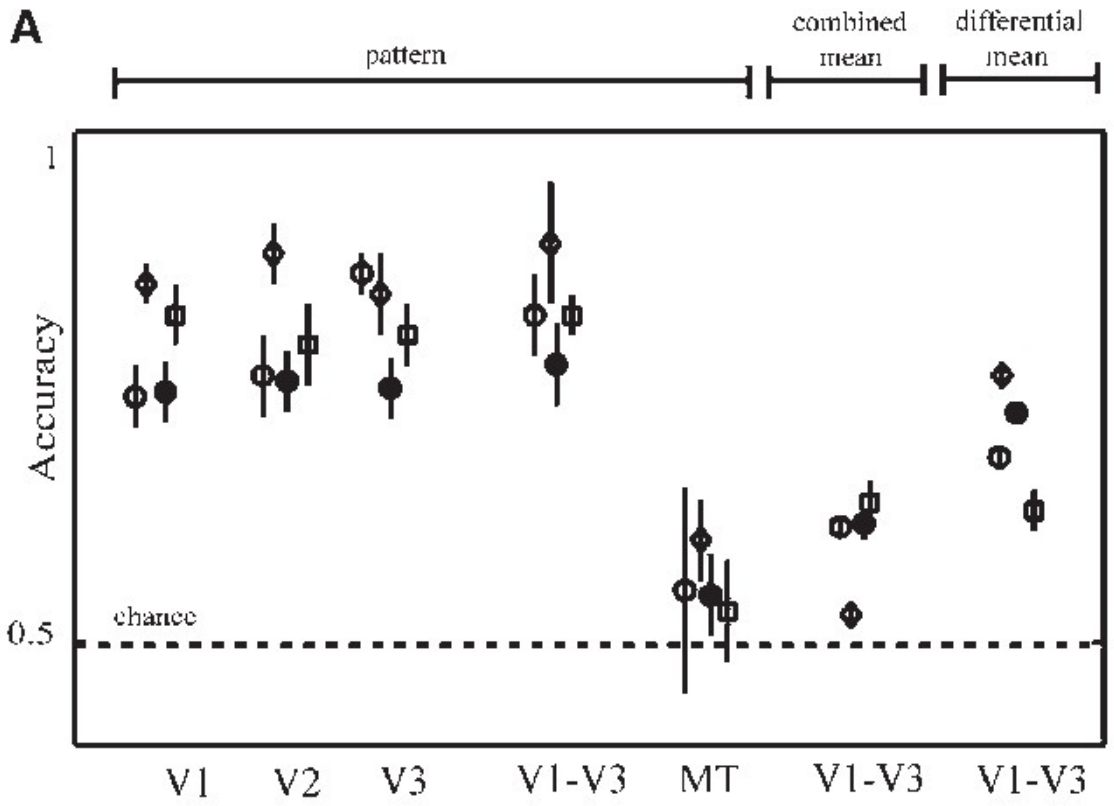


Figure 12: Accuracy by region assessed using cross-validation.

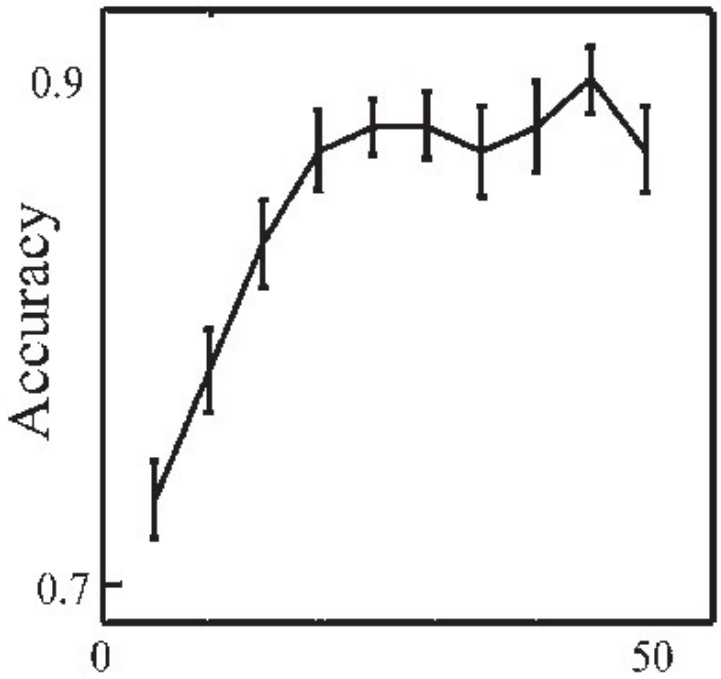


Figure 13: Accuracy by number of voxels assessed using cross-validation.

References

[1] A.J. Bell and T.J. Sejnowski. An information maximisation approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1004–1034, 1995.

[2] C.M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, 1995.

[3] P. Dayan and L.F. Abbott. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. MIT Press, 2001.

[4] J.D. Haynes and G. Rees. Predicting the stream of consciousness from activity in human visual cortex. *Current Biology*, 15:1301–1307, 2005.

[5] T. Jung, S. Makeig, C. Humphries, T. Lee, M. J. McKeown, V. Iragui, and T. J. Sejnowski. Removing Electroencephalographic Artifacts by Blind Source Separation. *Psychophysiology*, 1999.

[6] D.D. Wackerley, W. Mendenhall, and R.L. Scheaffer. *Mathematical statistics with applications*. Duxbury Press, 1996.