Maths for Brain Imaging: Lecture 6

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

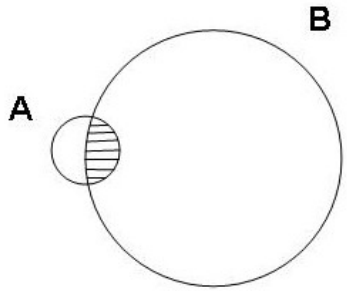
November 15, 2006

1 Contents

Bayes rule for

- Gaussians
- General Linear Models

and Parametric Empirical Bayes (PEB). Application to M/EEG source localisation.



Given probabilities $p(A)$, $p(B)$, and the joint probability $p(A, B)$, we can write the conditional probabilities

$$\begin{aligned} p(B|A) &= \frac{p(A,B)}{p(A)} \\ p(A|B) &= \frac{p(A,B)}{p(B)} \end{aligned}$$

Eliminating $p(A, B)$ gives Bayes rule

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)}$$

1.1 Gaussians

'Precision' is inverse variance eg. variance of 0.1 is precision of 10.

For a Gaussian prior with mean m_0 and precision p_0, and a Gaussian likelihood with mean m_D and precision p_D the posterior is Gaussian with

$$\begin{aligned} p &= p_0 + p_D \\ m &= \frac{p_0}{p} m_0 + \frac{p_D}{p} m_D \end{aligned}$$

So, (1) precisions add and (2) the posterior mean is the sum of the prior and data means, but each weighted by their relative precision.

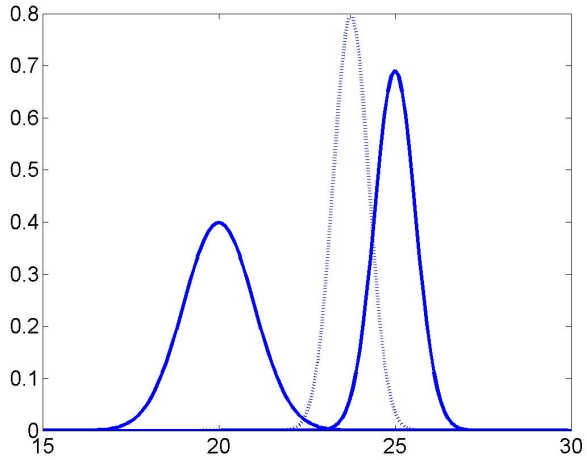


Figure 1: *Bayes rule for univariate Gaussians. The two solid curves show the probability densities for the prior* $m_0 = 20$, $p_0 = 1$ *and the likelihood* $m_D = 25$ *and* $p_D = 3$. *The dotted curve shows the posterior distribution with* $m = 23.75$ *and* $p = 4$. *The posterior is closer to the likelihood because the likelihood has higher precision.*

1.2 Bayesian GLM

If $p(x) = \mathsf{N}(m, \Sigma)$ then

$$p(x) \propto \exp\left(-\frac{1}{2}(x-m)^T\Sigma^{-1}(x-m)\right) \tag{1}$$

A Bayesian GLM is defined as

$$\begin{aligned} y &= X\beta + e_1 \\ \beta &= \mu + e_2 \end{aligned} \tag{2}$$

where the errors are zero mean Gaussian with covariances $\mathsf{Cov}[e_1] = C_1$ and $\mathsf{Cov}[e_2] = C_2$.

$$\begin{aligned} p(y|\beta) &\propto \exp\left(-\tfrac{1}{2}(y-X\beta)^T C_1^{-1}(y-X\beta)\right) \\ p(\beta) &\propto \exp\left(-\tfrac{1}{2}(\beta-\mu)^T C_2^{-1}(\beta-\mu)\right) \end{aligned} \tag{3}$$

The posterior distribution is then

$$p(\beta|y) \propto p(y|\beta)p(\beta) \tag{4}$$

Taking logs and keeping only those terms that depend on β gives

$$\begin{aligned}
\log p(\beta|y) &= -\frac{1}{2}(y - X\beta)^T C_1^{-1}(y - X\beta) \qquad (5)\\
&- \frac{1}{2}(\beta - \mu)^T C_2^{-1}(\beta - \mu) + ..\\
&= -\frac{1}{2}\beta^T (X^T C_1^{-1} X + C_2^{-1})\beta\\
&+ \beta^T (X^T C_1^{-1} y + C_2^{-1}\mu) + ..
\end{aligned}$$

Taking logs of the Gaussian density $p(x)$ in equation 2 and keeping only those terms that depend on x gives

$$\log p(x) = -\frac{1}{2}x^T \Sigma^{-1} x + x^T \Sigma^{-1} m + .. \qquad (6)$$

Comparing equation 5 with terms in the above equation shows that

$$\begin{aligned}
p(\beta|y) &= \mathsf{N}(m, \Sigma) \qquad (7)\\
\Sigma^{-1} &= X^T C_1^{-1} X + C_2^{-1}\\
m &= \Sigma(X^T C_1^{-1} y + C_2^{-1}\mu)
\end{aligned}$$

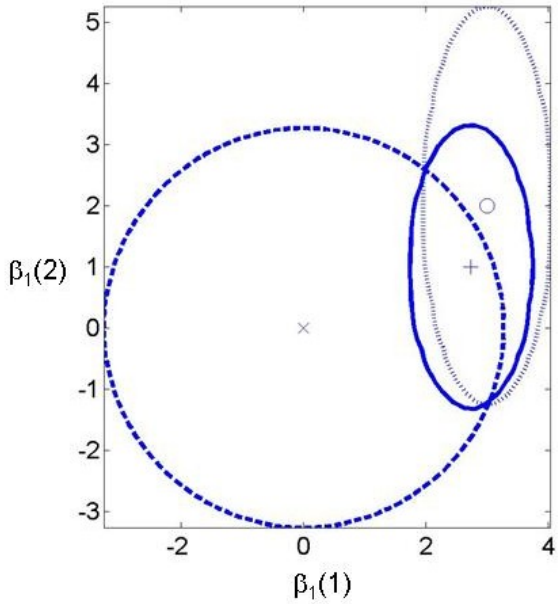


Figure 2: GLMs with two parameters. The prior (dashed line) has mean $\mu = [0, 0]^T$ (cross) and precision $C_1^{-1} = \mathsf{diag}([1, 1])$. The likelihood (dotted line) has mean $X^T y = [3, 2]^T$ (circle) and precision $(X^T C_1^{-1} X)^{-1} = \mathsf{diag}([10, 1])$. The posterior (solid line) has mean $m = [2.73, 1]^T$ (cross) and precision $\Sigma^{-1} = \mathsf{diag}([11, 2])$. In this example, the measurements are more informative about $\beta(1)$ than $\beta(2)$. This is reflected in the posterior distribution.

1.3 Augmented Form

From before

$$\begin{aligned} p(\beta|y) &= \mathsf{N}(m, \Sigma) \\ \Sigma^{-1} &= X^T C_1^{-1} X + C_2^{-1} \\ m &= \Sigma(X^T C_1^{-1} y + C_2^{-1} \mu) \end{aligned} \tag{8}$$

This can also be written as

$$\begin{aligned} \Sigma^{-1} &= \bar{X}^T V^{-1} \bar{X} \\ m &= \Sigma(\bar{X}^T V^{-1} \bar{y}) \end{aligned} \tag{9}$$

where

$$\begin{aligned} \bar{X} &= \begin{bmatrix} X \\ I \end{bmatrix} \\ V &= \begin{bmatrix} C_1 & 0 \\ 0 & C_2 \end{bmatrix} \\ \bar{y} &= \begin{bmatrix} y \\ \mu \end{bmatrix} \end{aligned} \tag{10}$$

where we've augmented the data matrix with prior expectations. Estimation in a Bayesian GLM is therefore equivalent to Maximum Likelihood estimation (ie. for IID covariances this is the same as Weighted Least Squares) with *augmented* data. Our prior beliefs can be thought of as extra data points.

2 Parametric Empirical Bayes

For a Bayesian GLM

$$\begin{aligned} y &= X\beta + e_1 \\ \beta &= \mu + e_2 \end{aligned} \tag{11}$$

with linear covariance constraints

$$\begin{aligned} C_1 &= \sum_i \lambda_i Q_i \\ C_2 &= \sum_j \lambda_j Q_j \end{aligned} \tag{12}$$

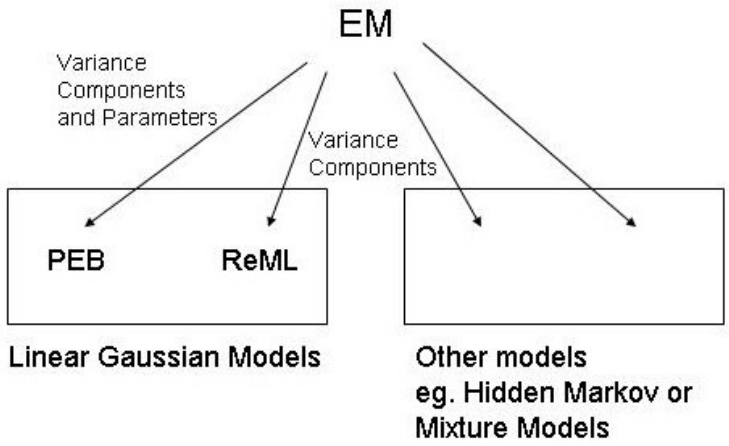
the covariance components can be estimated using ReML (last lecture). We can then make inferences about inter-

mediate level parameters eg. β using Bayes rule (earlier in this lecture).

Also, the ReML algorithm can be reformulated into two steps (i) estimation the posterior distribution over β's and (ii) hyperparameter estimation (λ's). This reformulation is known as Parametric Empirical Bayes (PEB). The difference is that, in ReML, step (i) is embedded into step (ii). For ReML the goal is to estimate variance components, for PEB the goal is to estimate (intermediate level) parameters.

PEB is a special case of an Expectation-Maximisation (EM) algorithm where (i) E-Step: estimate posterior distribution over β's (ii) M-Step: update λ's. PEB/ReML are specific to linear Gaussian models but EM is generic, ie. there is an EM algorithm for mixture models, hidden Markov models etc.

For hierarchical linear models the PEB/EM algorithm is


EM
Variance
Components
and Parameters
Variance
Components
PEB
ReML
Linear Gaussian Models
Other models
eg. Hidden Markov or
Mixture Models

- E-Step: Update distribution over parameters β

$$\begin{aligned} \Sigma^{-1} &= \bar{X}^T V^{-1} \bar{X} \\ m &= \Sigma(\bar{X}^T V^{-1} \bar{y}) \end{aligned} \tag{13}$$

- M-Step: Update hyperparameters λ_i (and therefore V) by following gradient g_i

$$\begin{aligned} r &= \bar{y} - \bar{X}m \\ g_i &= -\frac{1}{2} Tr(V^{-1} Q_i) + \frac{1}{2} Tr(\Sigma \bar{X}^T V^{-1} Q_i V^{-1} \bar{X}) \\ &+ \frac{1}{2} r^T V^{-1} Q_i V^{-1} r \end{aligned} \tag{14}$$

The M-Step is identical to ReML (last lecture) as the gradient can be expressed as

$$\begin{aligned} g_i &= -\frac{1}{2} Tr(PQ_i) + \frac{1}{2} y^T P^T Q_i P y \\ P &= V^{-1} - V^{-1} \bar{X} (\bar{X}^T V^{-1} \bar{X})^{-1} \bar{X}^T V^{-1} \end{aligned} \tag{15}$$

Whether or not EM or ReML is more computationally efficient for estimating variance components depends on

the sparsity of the covariance constraints Q_i. For more details (and Fisher scoring implementation) see [3].

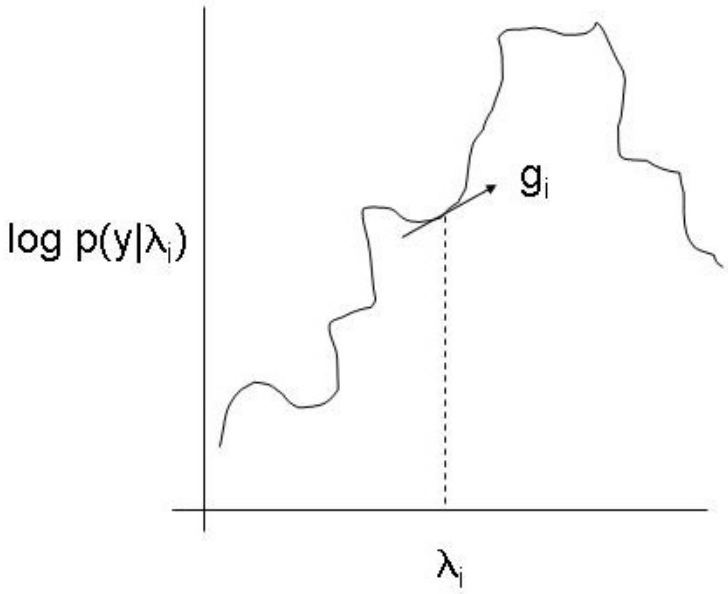


Figure 3: *EM and ReML estimate hyperparameters λ_i by following the gradient to the (local) maximum.*

2.1 Global Shrinkage Priors

Used in eg. fMRI analysis [2]. Special case of hierarchical model

$$\begin{aligned} y &= X\beta + e_1 \\ \beta &= \mu + e_2 \end{aligned} \tag{16}$$

with 20 voxels and 10 data points per voxel

$$X = I_{20} \otimes 1_{10} \tag{17}$$

$$\begin{aligned} C_1 &= \sum_{i=1}^{20} \frac{1}{v_i} Q_i \\ C_2 &= \frac{1}{\alpha} I_{20} \end{aligned} \tag{18}$$

The parameter $\beta(i)$ encodes the effect size at voxel i. This model assumes that across the brain (i) average effect size is zero, $\mu = 0$, and (ii) the variability of responses follows a Gaussian with precision α. Hyperparameters are $\lambda = \{v_i, \alpha\}$.

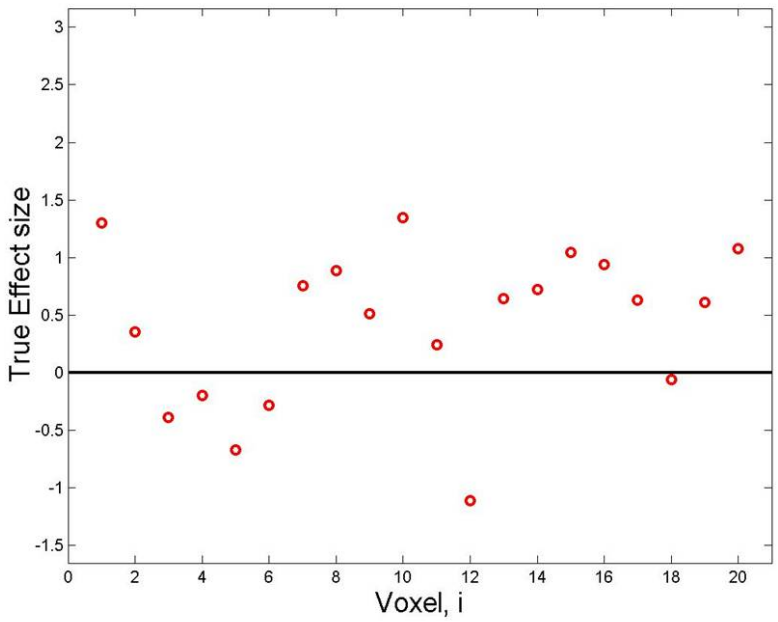


Figure 4: *Across the 20-voxel brain (i) average effect size is zero, $\mu = 0$, the variability of responses follows a Gaussian with precision α. True effect sizes (red circles).*

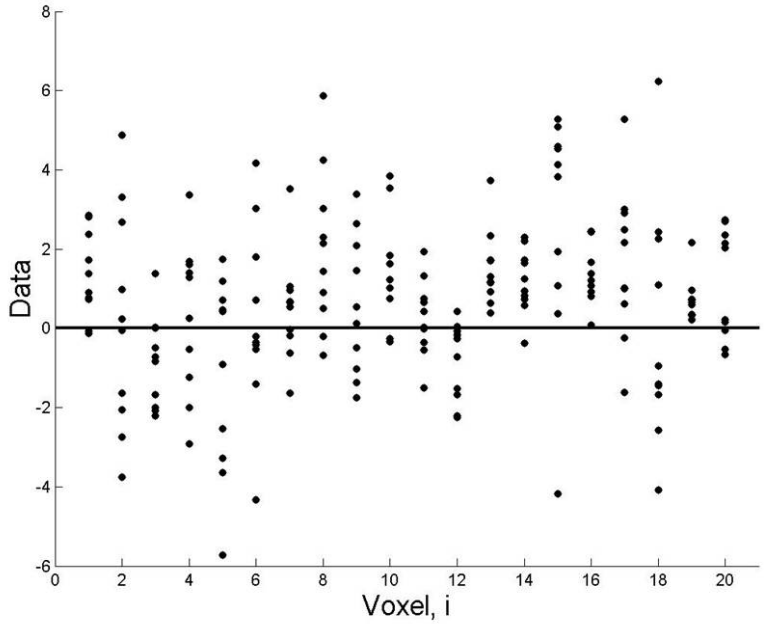


Figure 5: *Data at each voxel are normally distributed about the effect size at that voxel with precision λ_i eg. voxels 2, 5 and 15 have noisier data than others.*

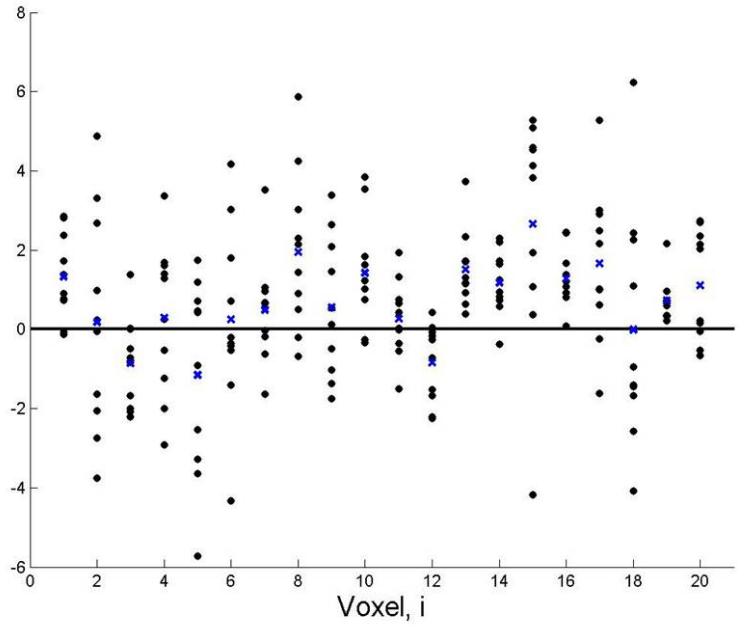


Figure 6: *Previous graph but with sample means (blue crosses) also at each voxel.*

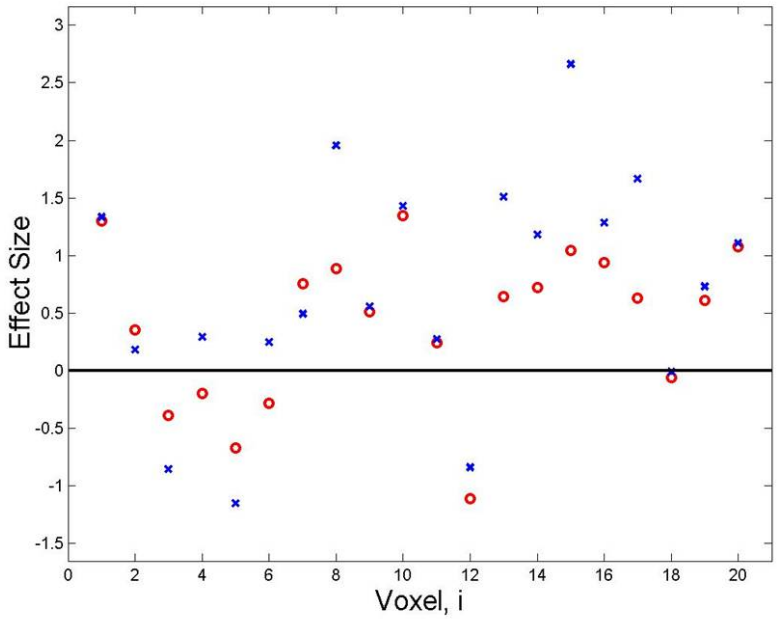


Figure 7: Sample means (also ML estimates - blue crosses) and true effect sizes (red circles). Estimation error =0.71.

For this model the PEB algorithm has a simple form. By setting the gradients g_i to zero we can get the following updates for the hyperparameters $\lambda = \{v_i, \alpha\}$.

$$\begin{aligned} \beta(i) &= \frac{\gamma_i}{N} \sum_{n=1}^{N} y_{in} \\ \frac{1}{v_i} &= \frac{1}{N - \gamma_i} \sum_{n=1}^{N} (y_{in} - \beta(i))^2 \\ \gamma_i &= \frac{N v_i}{N v_i + \alpha} \\ \frac{1}{\alpha} &= \frac{1}{\sum_i \gamma_i} \sum_{i=1}^{V} \beta(i)^2 \end{aligned} \tag{19}$$

where y_{in} is the nth scan at the ith voxel, γ_i is the ratio of the data precision to the posterior precision.

Without a prior, $\gamma_i = 1$ we get

$$\frac{1}{v_i} = \frac{1}{N-1} \sum_{n=1}^{N} (y_{in} - \beta(i))^2 \tag{20}$$

This is the familiar 'unbiased' estimate, if we only have

to estimate variance components at a single level. The PEB updates partition the total degrees of freedom N into those used to estimate first or second level hyperparameters.

See code `em1.m`.

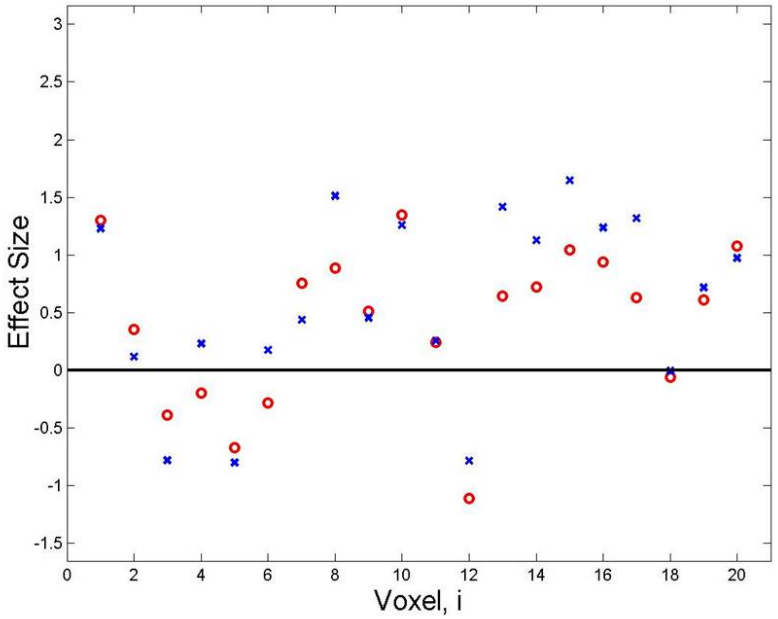


Figure 8: *After PEB iteration 3*

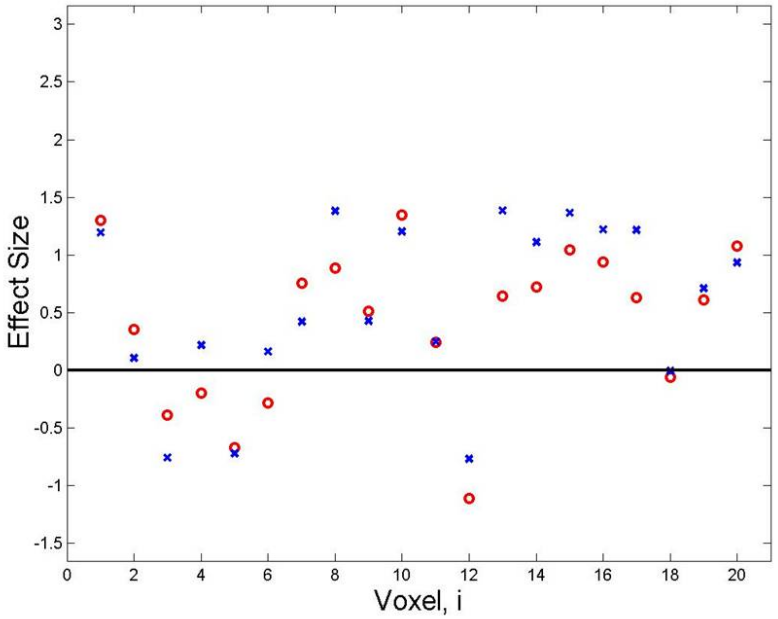


Figure 9: *After PEB iteration 7. Estimation error =0.34.*

On average, across the brain, PEB is more accurate than ML. It does better at most voxels at the expense of being worse at a minority eg. voxel 2.

For most voxels we have $\gamma_i = 0.9$, but for the noisy voxels 2, 15 and 18 we have $\gamma_i = 0.5$. PEB thus relies more on prior information where data are unreliable.

2.2 EEG Source Reconstruction

To 'reconstruct' EEG data at a *single time point* use the model

$$\begin{aligned} y &= X\beta + e_1 \qquad (21) \\ \beta &= \mu + e_2 \end{aligned}$$

where X is a lead-field matrix transforming Current Source Density (CSD) β at V voxels in brain space into EEG voltages y at S electrodes. For more on this see eg. [1].

$$\begin{aligned} C_1 &= \sum_i \lambda_i Q_i \qquad (22) \\ C_2 &= \sum_j \lambda_j Q_j \end{aligned}$$

$$(23)$$

where Q_i defines structure of sensor noise, and Q_j source noise ie. uncertainty in sources. In the application that follows we use $Q_i = I$ and $Q_j = L$, a 'Laplacian' matrix set up so that we expect the squared difference between

neighboring voxels to be λ_j ie. this enforces a smoothness constraint.

The data in this analysis is from [4].

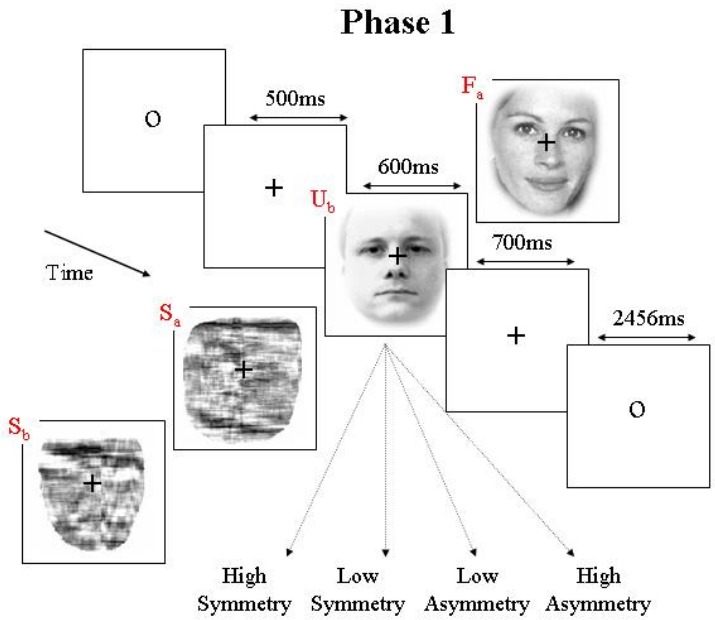


Figure 10: *Subjects are presented images of faces and scrambled faces and are asked to make symmetry judgements.*

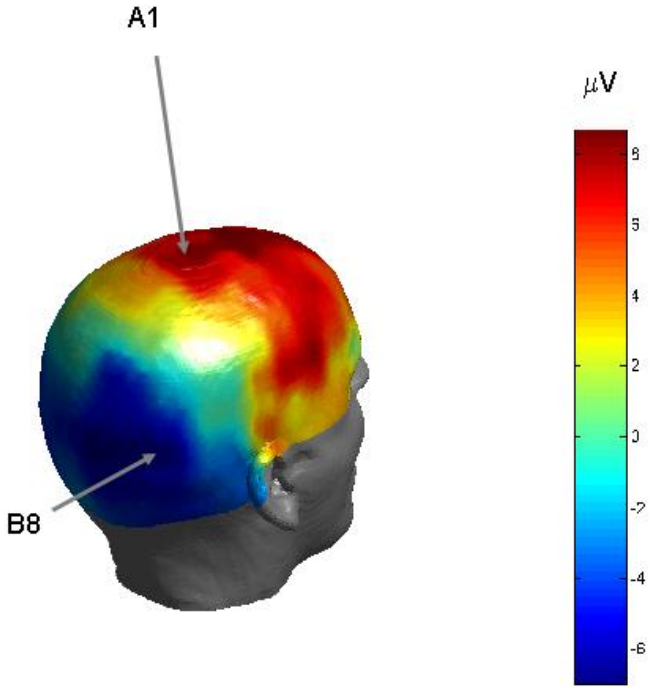


Figure 11: *Electrode voltages at 160ms post-stimulus, y. This is an Event-Related Potential (ERP), the result of averaging the responses to many (86) trials.*

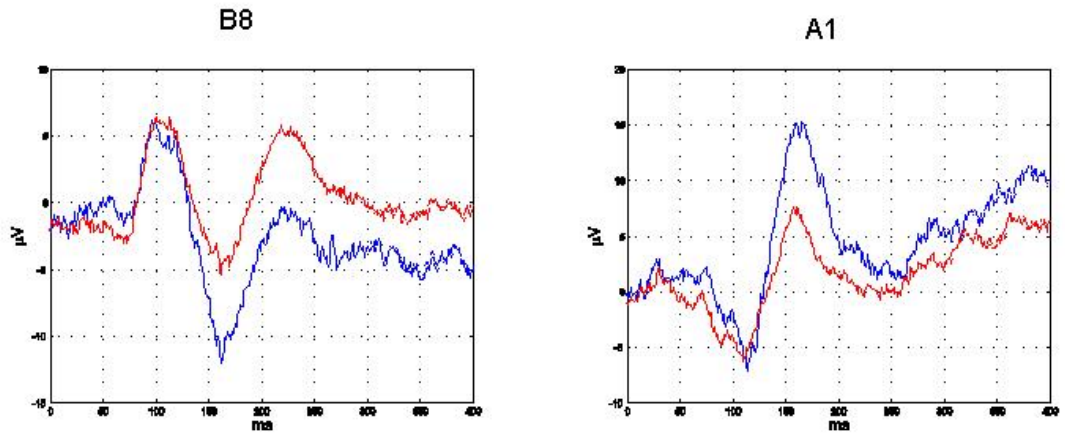


Figure 12: *Voltages at two different electrodes for faces (blue) and scrambled faces (red). These are Event-Related Potentials (ERPs), the result of averaging the responses to many (86) trials.*

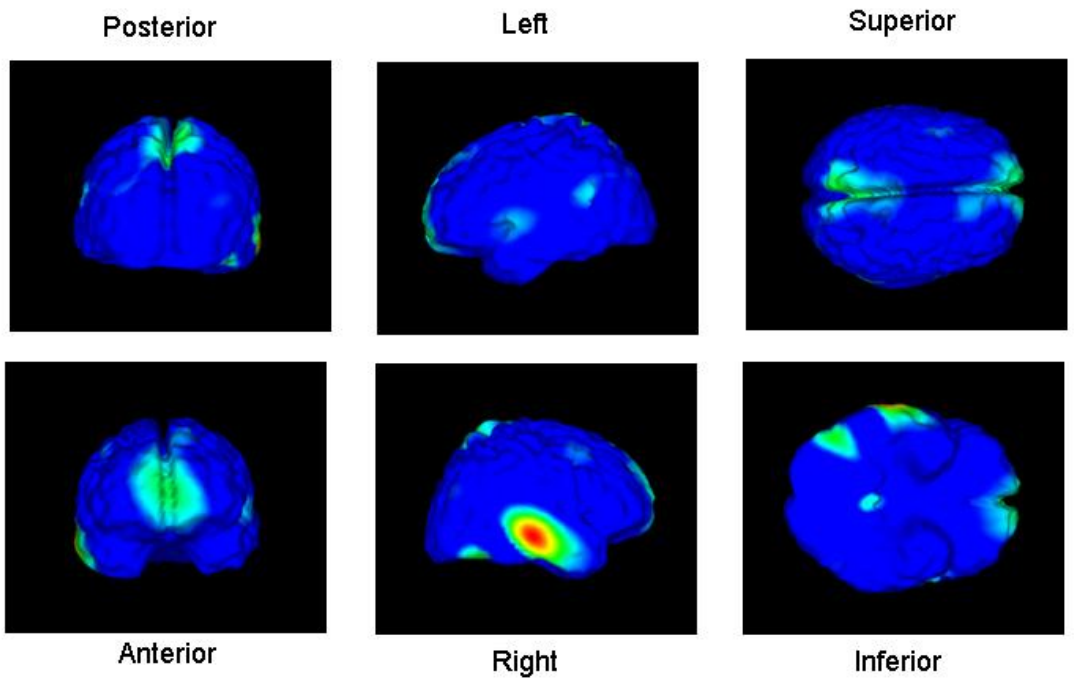


Figure 13: Estimate of CSD, β. Computed as the CSD difference for faces minus scrambled faces.

References

[1] S. Baillet, J.C. Mosher, and R.M. Leahy. Electromagnetic Brain Mapping. *IEEE Signal Processing Magazine*, pages 14–30, November 2001.

[2] K.J. Friston, D.E. Glaser, R.N.A. Henson, S.J. Kiebel, C. Phillips, and J. Ashburner. Classical and Bayesian inference in neuroimaging: Applications. *NeuroImage*, 16:484–512, 2002.

[3] K.J. Friston, W.D. Penny, C. Phillips, S.J. Kiebel, G. Hinton, and J. Ashburner. Classical and Bayesian inference in neuroimaging: Theory. *NeuroImage*, 16:465–483, 2002.

[4] R.N.A. Henson, Y. Goshen-Gottstein, T. Ganel, L.J. Otten, A. Quayle, and M.D. Rugg. Electrophysiological and hemodynamic correlates of face perception, recognition and priming. *Cerebral Cortex*, 13:793–805, 2003.