Expectation-Maximisation

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

March 19, 2007

1 Contents

- Kullback-Liebler divergence
- Variational Bayes
- Expectation Maximisation
- Mixture models
- Bayes rule for Gaussians and GLMs
- Parametric Empirical Bayes
- M/EEG source reconstruction

2 Kullback-Liebler divergence

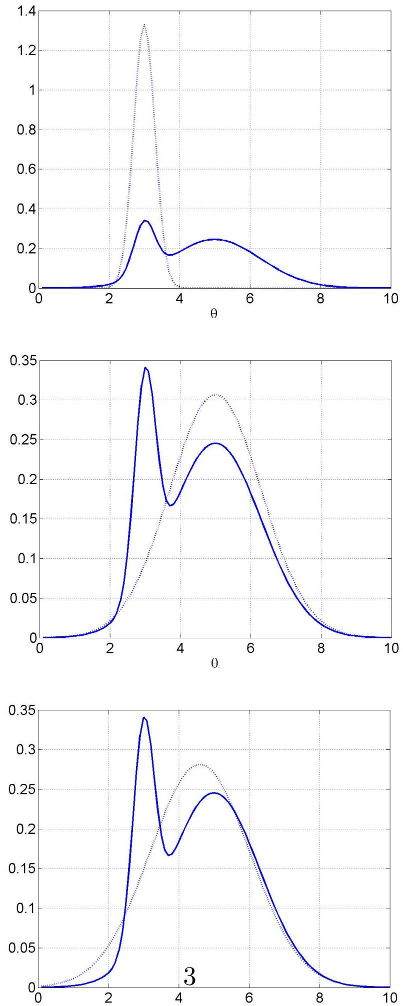
For densities $q(H)$ and $p(H)$ the Relative Entropy or Kullback-Liebler (KL) divergence from q to p is

$$KL[q||p] = \int q(H) \log \frac{q(H)}{p(H)} dH \tag{1}$$

The KL-divergence satisfies the Gibb's inequality

$$KL[q||p] \geq 0 \tag{2}$$

with equality only if $q = p$. In general $KL[q||p] \neq KL[p||q]$, so KL is not a distance measure.



(a)

(b)

(c)

Figure 1: *Probability densities* $p(H)$ *(solid lines) and* $q(H)$ *(dashed lines) for a Gaussian mixture* $p(H) = 0.2 \times \mathsf{N}(m_1, \sigma_1^2) + 0.8 \times \mathsf{N}(m_2, \sigma_2^2)$ *with* $m_1 = 3, m_2 = 5, \sigma_1 = 0.3, \sigma_2 = 1.3$*, and a single Gaussian* $q(H) = \mathsf{N}(\mu, \sigma^2)$ *with (a)* $\mu = \mu_1, \sigma = \sigma_1$ *which fits the first mode, (b)* $\mu = \mu_2, \sigma = \sigma_2$ *which fits the second mode and (c)* $\mu = 4.6, \sigma = 1.4$ *which is moment-matched to* $p(H)$*.*

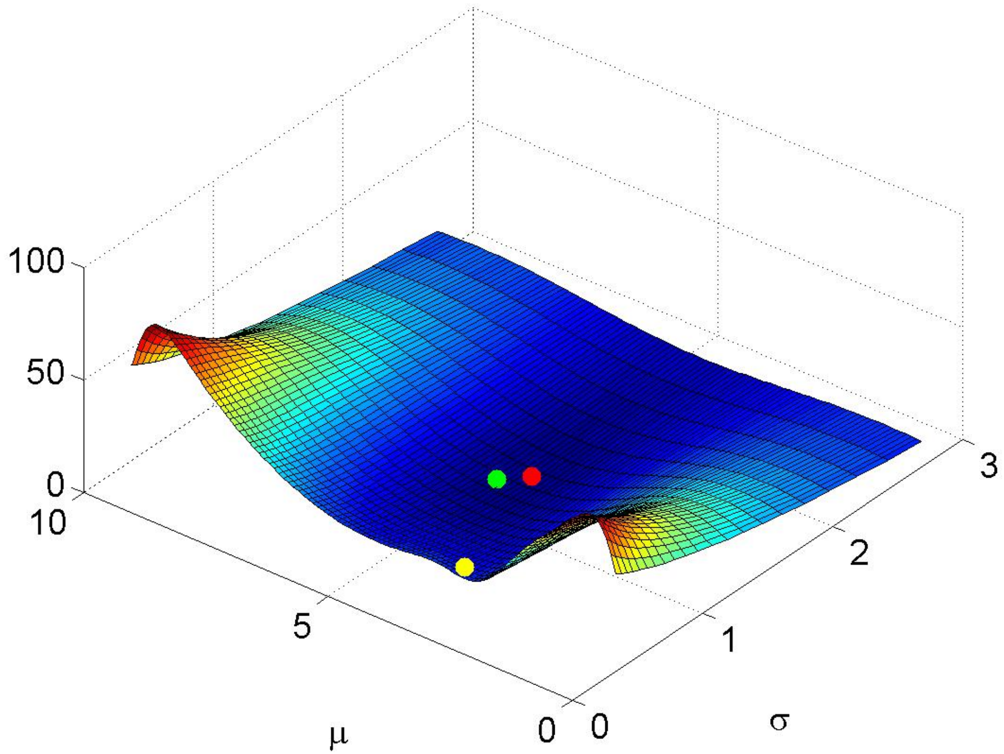


Figure 2: *KL-divergence,* $KL(q||p)$ *for* p *as defined in Figure 1 and* q *being a Gaussian with mean* μ *and standard deviation* σ*. The KL-divergences of the approximations in Figure 1 are (a) 11.73 for the first mode (yellow ball), (b) 0.93 for the second mode (green ball) and (c) 0.71 for the moment-matched solution (red ball).*

3 Variational Bayes

Given a probabilistic model of some data, the log of the 'evidence' or 'marginal likelihood' can be written as

$$
\begin{aligned}
\log p(Y) &= \int q(H) \log p(Y) dH \\
&= \int q(H) \log \frac{p(Y,H)}{p(H|Y)} dH \\
&= \int q(H) \log \left[\frac{p(Y,H)q(H)}{q(H)p(H|Y)} \right] dH \\
&= F + KL(q(H)||p(H|Y)) \qquad (3)
\end{aligned}
$$

where $q(H)$ is considered, for the moment, as an arbitrary density. We have

$$
F = \int q(H) \log \frac{p(Y,H)}{q(H)} dH, \qquad (4)
$$

which in statistical physics is known as the *negative* variational free energy. The second term in equation 3 is the KL-divergence between the density $q(H)$ and the true posterior $p(H|Y)$. Equation 3 is the fundamental equation of the VB-framework and is shown graphically in Figure 3. Because KL is always positive, due to the Gibbs inequality, F provides a lower bound on the model evidence. Moreover, because KL is zero when two densities are the same, F will become equal to the model evidence when $q(H)$ is equal to the true posterior. For this reason $q(H)$ can be viewed as an *approximate posterior*.

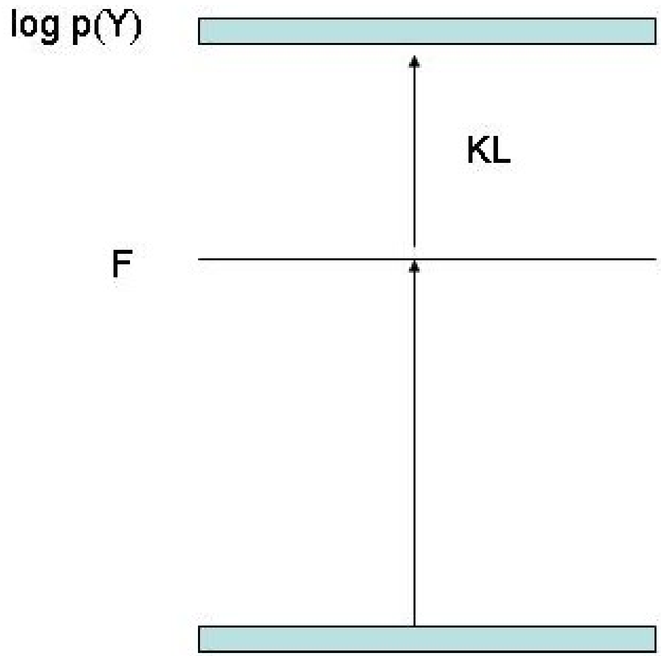


Figure 3: *The negative variational free energy, F, provides a lower bound on the log-evidence of the model with equality when the approximate posterior equals the true posterior.*

4 Mixture models

4.1 EM for mixture models

In this context EM is a maximum-likelihood algorithm for models with observed variables Y and hidden variables H. Hidden variable denotes which Gaussian is used to generate a data point. Select Gaussian k with probability k. That Gaussian has parameters μ_k and Σ_k.

Now, repeat 'VB derivation' but with eveything conditioned on parameters $\beta = \{\mu_k, \Sigma_k, \pi_k\}$. This gives

$$\log p(Y|\beta) = F_{EM} + KL[q(H)||p(H|Y,\beta)] \tag{5}$$

where

$$F_{EM} = \int q(H) \log \frac{p(H,Y|\beta)}{q(H)} dH \tag{6}$$

This gives rise to the following algorithm.

- E-Step: Set $q(H) = p(H|Y,\beta)$. This sets the KL term to zero. This can be done by letting

$$q(h_n) \;=\; p(h_n|y_n,\beta) \tag{7}$$

$$= \frac{p(y_n|h_n, \beta)p(h_n|\beta)}{p(y_n|\beta)} \quad (8)$$

for all data points n. This is just Bayes rule. Write $\gamma_n^k = q(h_n = k)$, the responsibilies ie. the probability that data point n was generated from the kth Gaussian.

- M-step: Now, as $KL = 0$, $F_{EM} = \log p(Y|\beta)$, so we can maximise the likelihood wrt. β by maximising F_{EM} wrt. β. We have

$$\begin{aligned} F_{EM} &= \sum_k \sum_n \gamma_k^n \log p(y_n|h_n = k)p(h_n = k) \quad (9) \\ &= \sum_k \sum_n \gamma_k^n \log p(y_n|h_n = k) + \sum_k \sum_n \gamma_k^n p(h_n = k) \end{aligned}$$

Setting the derivatives $dF_{EM}/d\beta$ to zero gives the following updates

$$\mu_k = \frac{\sum_n \gamma_n^k y_n}{\sum_n \gamma_n^k} \quad (10)$$

$$
\begin{aligned}
\Sigma_k &= \frac{\sum_n \gamma_n^k (y_n - \mu_k)(y_n - \mu_k)^T}{\sum_n \gamma_n^k} \\
\pi_k &= \frac{\sum_n \gamma_n^k}{N}
\end{aligned}
$$

See netlab demo `demgmm1.m`.

5 Bayes rule for 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.

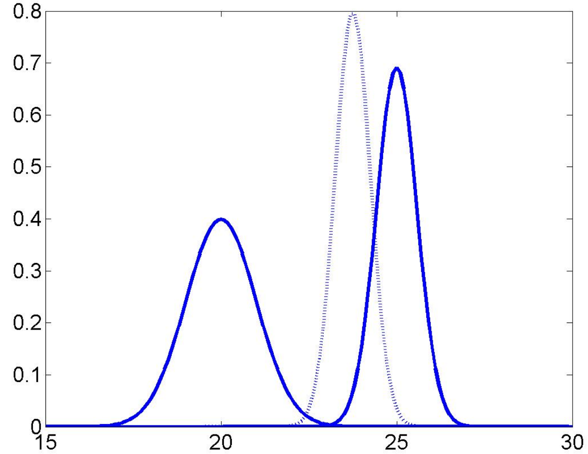


Figure 4: *Bayes rule for univariate Gaussians. The two solid curves show the probability densities for the prior* $m_0 = 20$, $p_0 = 1$ *and the likelihood* $m_D = 25$ *and* $p_D = 3$. *The dotted curve shows the posterior distribution with* $m = 23.75$ *and* $p = 4$. *The posterior is closer to the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

6 Bayesian GLM

A Bayesian GLM is defined as

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

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

$$\begin{aligned} p(y|\beta) &\propto \exp\left(-\tfrac{1}{2}(y - X\beta)^T C_1^{-1}(y - X\beta)\right) \\ p(\beta) &\propto \exp\left(-\tfrac{1}{2}(\beta - \mu)^T C_2^{-1}(\beta - \mu)\right) \end{aligned} \tag{12}$$

The posterior distribution is then

$$\begin{aligned} p(\beta|y) &= \mathsf{N}(m, \Sigma) \\ \Sigma^{-1} &= X^T C_1^{-1} X + C_2^{-1} \\ m &= \Sigma(X^T C_1^{-1} y + C_2^{-1}\mu) \end{aligned} \tag{13}$$

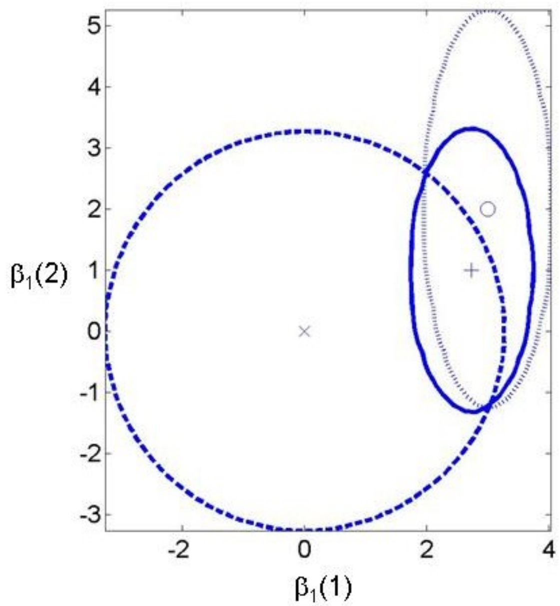


Figure 5: GLMs with two parameters. The prior (dashed line) has mean $\mu = [0,0]^T$ (cross) and precision $C_1^{-1} = \mathsf{diag}([1,1])$. The likelihood (dotted line) has mean $X^T y = [3,2]^T$ (circle) and precision $(X^T C_1^{-1} X)^{-1} = \mathsf{diag}([10,1])$. The posterior (solid line) has 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.

6.1 Augmented Form

From before

$$
\begin{aligned}
p(\beta|y) &= \mathsf{N}(m, \Sigma) \\
\Sigma^{-1} &= X^T C_1^{-1} X + C_2^{-1} \\
m &= \Sigma(X^T C_1^{-1} y + C_2^{-1} \mu)
\end{aligned} \tag{14}
$$

This can also be written as

$$
\begin{aligned}
\Sigma^{-1} &= \bar{X}^T V^{-1} \bar{X} \\
m &= \Sigma(\bar{X}^T V^{-1} \bar{y})
\end{aligned} \tag{15}
$$

where

$$
\begin{aligned}
\bar{X} &= \begin{bmatrix} X \\ I \end{bmatrix} \\
V &= \begin{bmatrix} C_1 & 0 \\ 0 & C_2 \end{bmatrix} \\
\bar{y} &= \begin{bmatrix} y \\ \mu \end{bmatrix}
\end{aligned} \tag{16}
$$

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.

7 Parametric Empirical Bayes

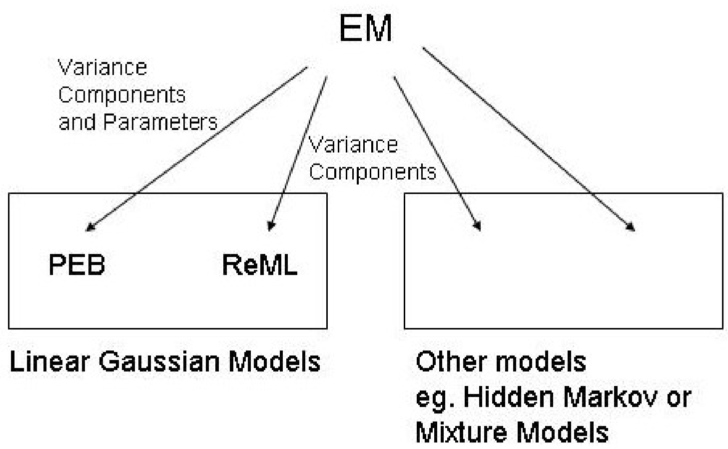
For a Bayesian GLM

$$\begin{aligned} y &= X\beta + e_1 \\ \beta &= \mu + e_2 \end{aligned} \tag{17}$$

with linear covariance constraints

$$\begin{aligned} C_1 &= \sum_i \lambda_i Q_i \\ C_2 &= \sum_j \lambda_j Q_j \end{aligned} \tag{18}$$

PEB is a special case of an Expectation-Maximisation (EM) algorithm where (i) E-Step: estimate posterior dis-



tribution over β's (ii) M-Step: update λ's. PEB is 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

- E-Step: Update distribution over parameters β

$$\begin{aligned} \Sigma^{-1} &= \bar{X}^T V^{-1} \bar{X} \\ m &= \Sigma(\bar{X}^T V^{-1} \bar{y}) \end{aligned} \tag{19}$$

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

$$\begin{aligned} r &= \bar{y} - \bar{X} m \\ g_i &= -\frac{1}{2} Tr(V^{-1} Q_i) + \frac{1}{2} Tr(\Sigma \bar{X}^T V^{-1} Q_i V^{-1} \bar{X}) \\ &+ \frac{1}{2} r^T V^{-1} Q_i V^{-1} r \end{aligned} \tag{20}$$

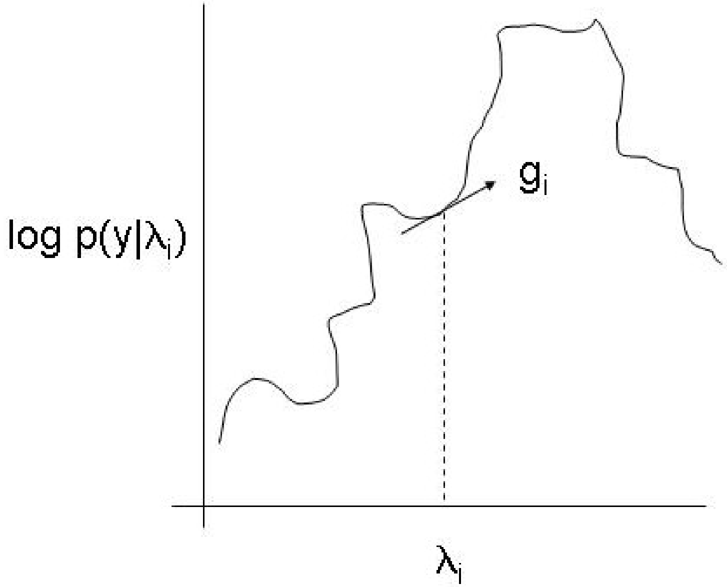


Figure 6: *EM and ReML estimate hyperparameters λ_i by following the gradient to the (local) maximum.*

7.1 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.

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

$$(23)$$

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 *Rik Henson*.

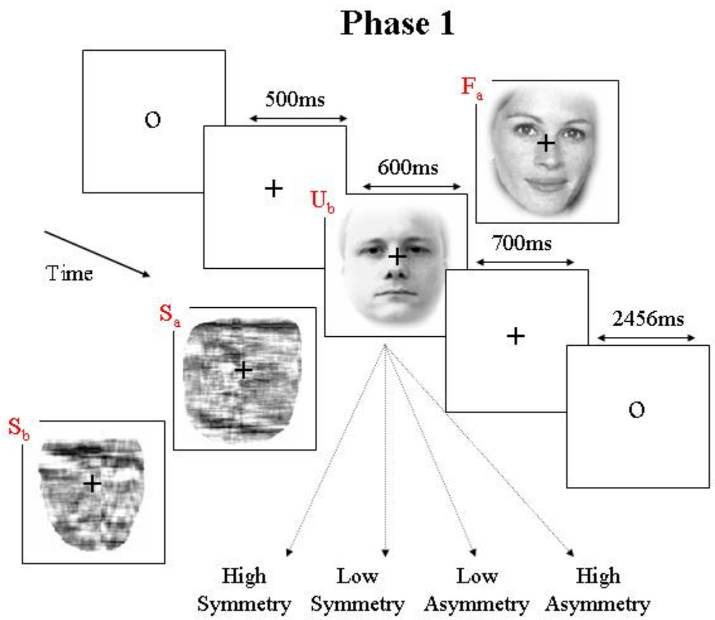


Figure 7: *Subjects are presented images of faces and scrambled faces and are asked to make symmetry judgements.*

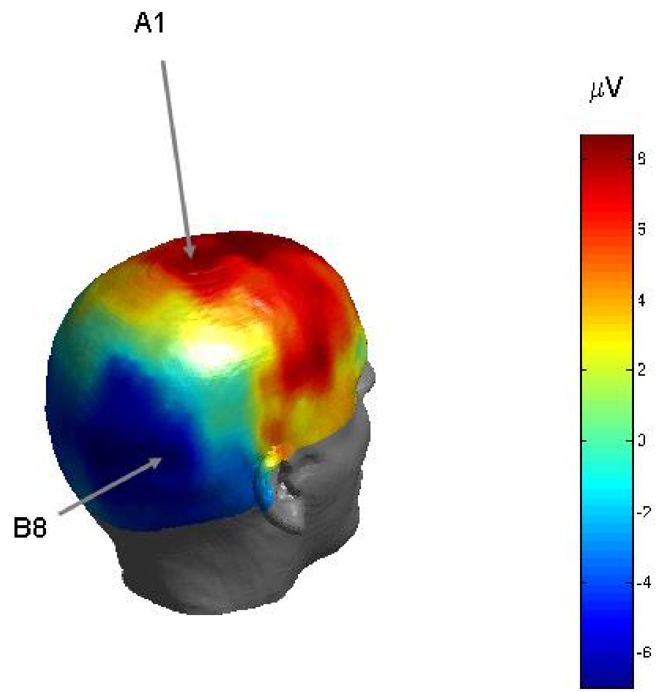


Figure 8: *Electrode voltages at 160ms post-stimulus, y. This is an Event-Related Potential (ERP), the result of averaging the responses to many (86) trials.*

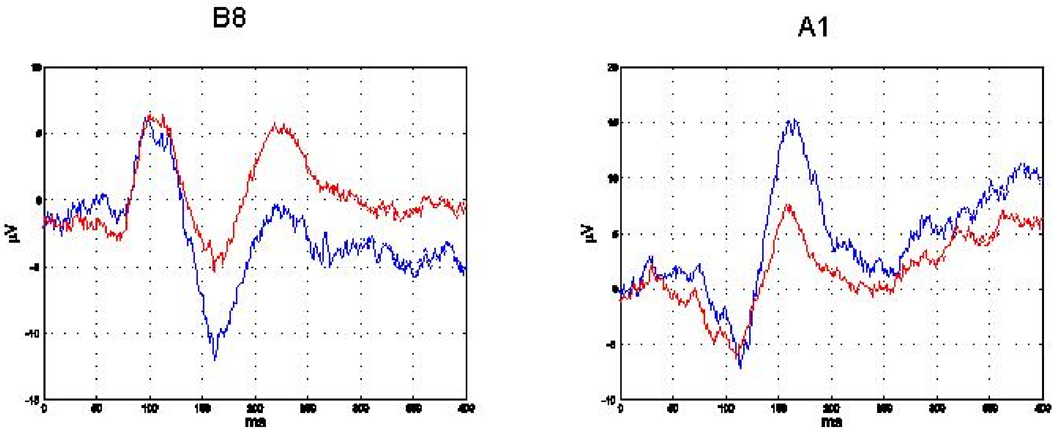


Figure 9: *Voltages at two different electrodes for faces (blue) and scrambled faces (red). These are Event-Related Potentials (ERPs), the result of averaging the responses to many (86) trials.*

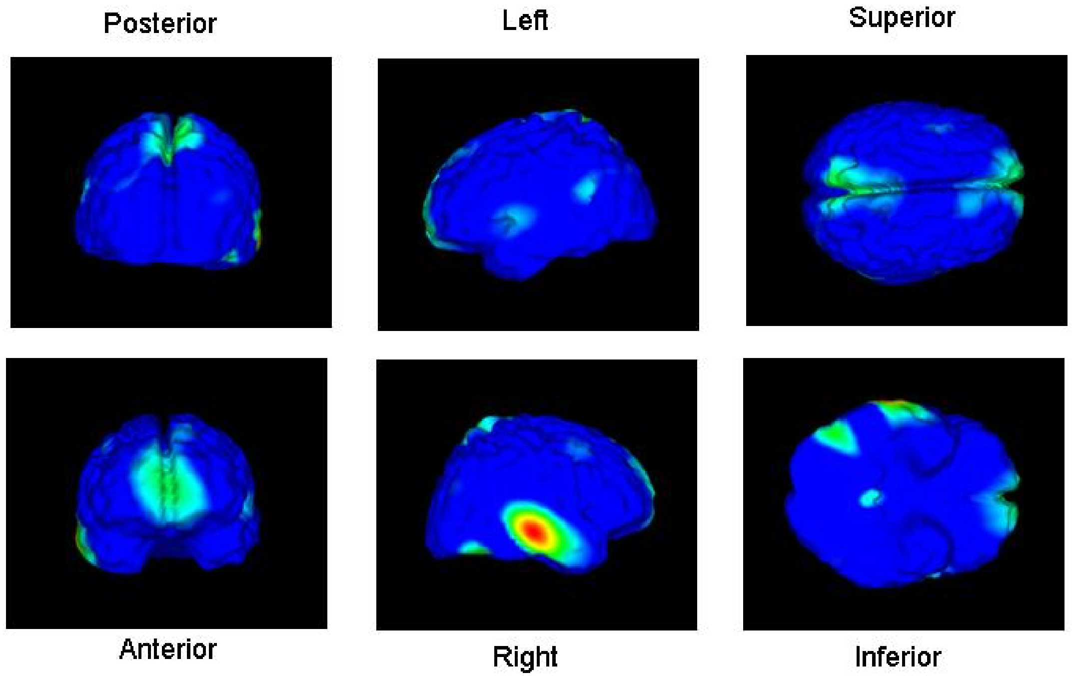


Figure 10: Estimate of CSD, β. Computed as the CSD difference for faces minus scrambled faces.