Maths for Brain Imaging: Lecture 7

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

November 21, 2006

1 Contents

Making inferences about models

- Bayes factors
- Evidence for Bayesian GLMs
- Multimodal imaging
- Bayesian Model Averaging
- Nonlinear M/EEG source localisation.

2 Bayes Factors

Bayes rule for data y and 'model' or 'hypothesis' i

$$p(m = i|y) = \frac{p(y|m = i)p(m = i)}{p(y)}$$

In this context, $p(y|m = i)$, is known as the evidence for model i. Similarly for model j

$$p(m = j|y) = \frac{p(y|m = j)p(m = j)}{p(y)}$$

Dividing one by the other gives

$$\frac{p(m = i|y)}{p(m = j|y)} = \frac{p(y|m = i)}{p(y|m = j)} \times \frac{p(m = i)}{p(m = j)}$$

This is the fundamental relationship

$$PosteriorOdds = BayesFactor \times PriorOdds$$

The Bayes factor is a ratio of model evidences. It tells you how the odds have changed. It can be written BF_{ij}.

2.1 Inferring cognitive processes

Poldrack[3] considers the relationship between engagement of cognitive processes, m, and activation of brain regions y. For example, using the BrainMap database, the frequency of language studies, L, that give rise to Broca activations (20mm ROI at $x = -37, y = 18, z = 18$mm) can be used to estimate

$$\begin{aligned} p(y = B|m = L) &= \frac{p(y = B, m = L)}{p(m = L)} \\ &= \frac{166}{869} = 0.191 \end{aligned} \tag{1}$$

Simiarly, given the number of non-language studies, $\bar{L}$, that also activate Broca's area

$$\begin{aligned} p(y = B|m = \bar{L}) &= \frac{p(y = B, m = \bar{L})}{p(m = \bar{L})} \\ &= \frac{199}{2353} = 0.085 \end{aligned} \tag{2}$$

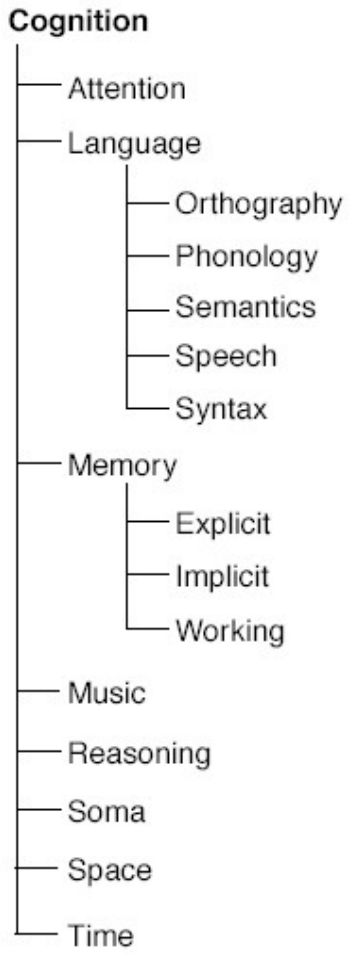

Figure 1: Cognitive processes, m, described in BrainMap database.

This gives rise to a Bayes factor

$$\begin{aligned} BF_{L,\bar{L}} &= \frac{p(y = B|m = L)}{p(y = B|m = \bar{L})} \\ &= \frac{0.191}{0.085} = 2.3 \end{aligned} \tag{3}$$

That is, after seeing a Broca activation, the odds that a language process has been engaged are larger by a factor 2.3.

For equal prior odds $p(m = L) = p(m = \bar{L}) = 0.5$, the posterior probability of language processes given a Broca activation is

$$\begin{aligned} p(m = L|y = B) &= \frac{p(y = B|m = L)p(m = L)}{p(y = B|m = L)p(m = L) + p(y = B|m = \bar{L})p(m = \bar{L})} \\ &= \frac{0.191}{0.191 + 0.085} = 0.69 \end{aligned}$$

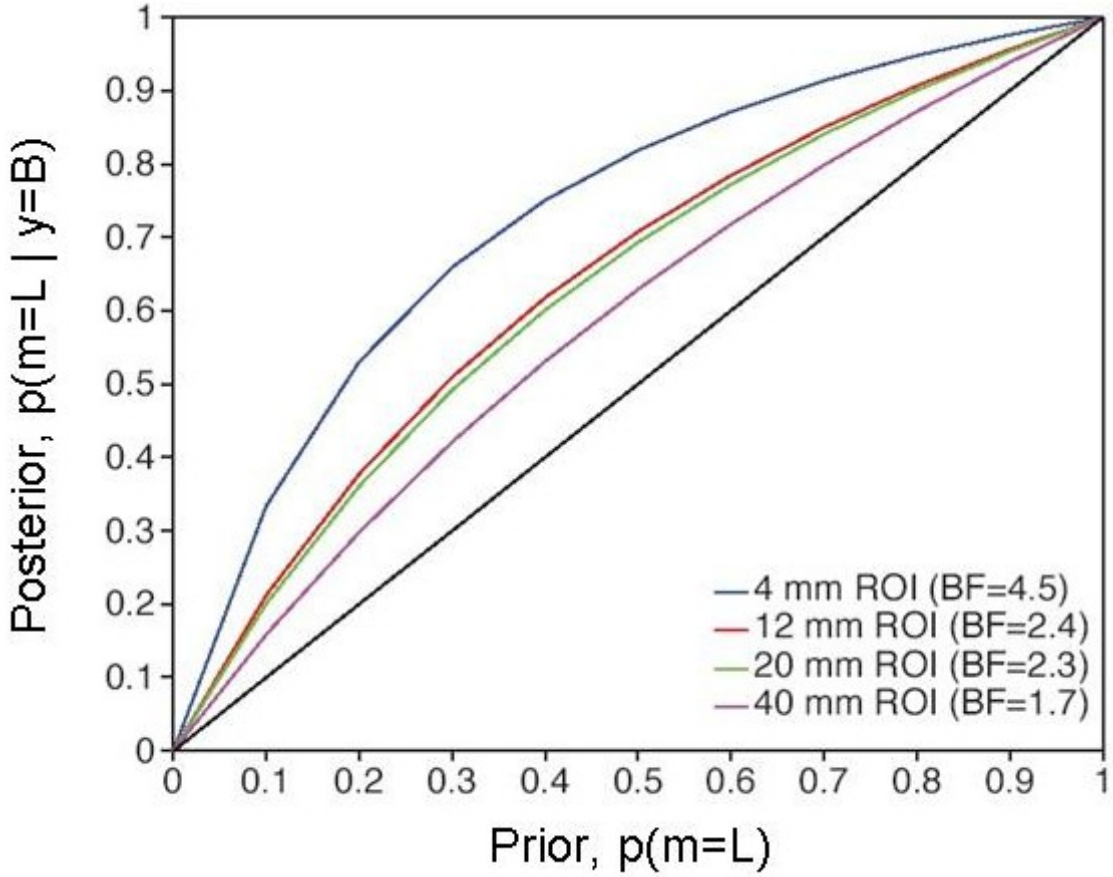

Figure 2: Effect of ROI size on posterior probability. Power of reverse inference is increased using smaller, more selective regions.

3 Making inferences about models

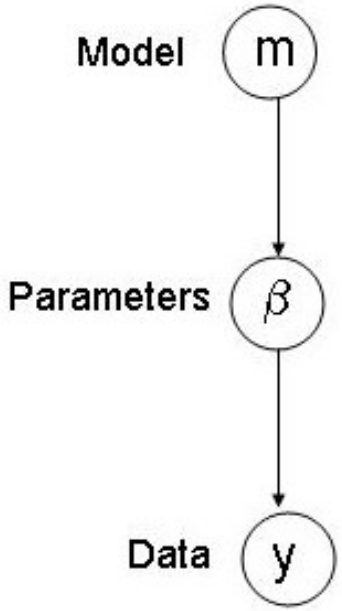

Figure 3: *Hierarchical generative model in which members of a model class, indexed by m, are considered as part of the hierarchy. Typically, m indexes the structure of the model. This might be the connectivity pattern in a dynamic causal model or set of anatomical or functional constraints in a source reconstruction model. Once a model has been chosen from the distribution $p(m)$, its parameters are generated from the parameter prior $p(\theta|m)$ and finally data is generated from the likelihood $p(y|\theta, m)$.*

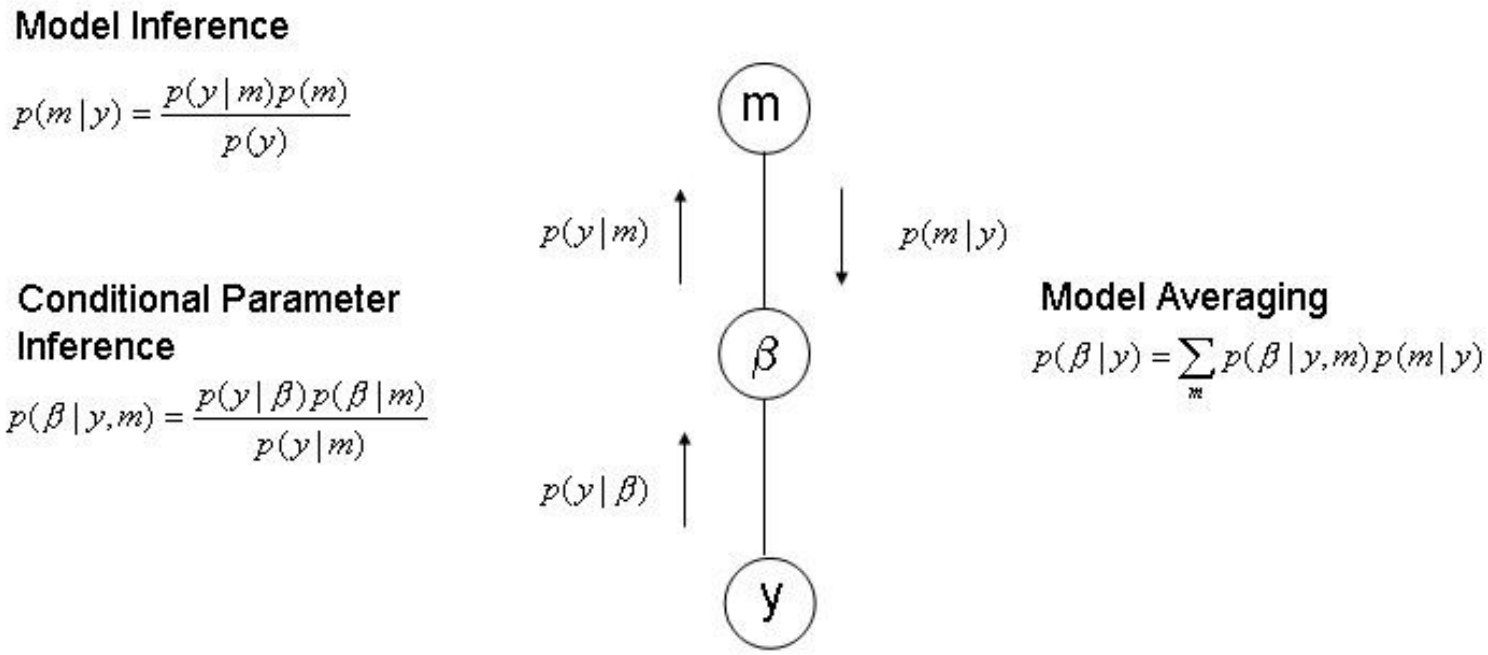

Figure 4: *In Bayesian Model Selection (BMS), the posterior model probability $p(m|y)$, is used to select a single 'best' model. In Bayesian Model Averaging (BMA), inferences are based on all models and $p(m|y)$ is used as a weighting factor. Only in BMA, are parameter inferences based on the correct marginal density $p(\theta|y)$.*

4 Evidence for Bayesian GLMs

For a Bayesian GLM

$$\begin{aligned} y &= X\beta + e_1 \\ \beta &= \mu + e_2 \end{aligned} \tag{4}$$

with linear covariance constraints

$$\begin{aligned} C_1 &= \sum_i \lambda_i Q_i \\ C_2 &= \sum_j \lambda_j Q_j \end{aligned} \tag{5}$$

From lecture 6 we know that the posterior distribution over regression coefficients is

$$\begin{aligned} \Sigma^{-1} &= \bar{X}^T V^{-1} \bar{X} \\ \hat{\beta} &= \Sigma(\bar{X}^T V^{-1} \bar{y}) \end{aligned} \tag{6}$$

where

$$\bar{X} = \begin{bmatrix} X \\ I \end{bmatrix} \tag{7}$$

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

where we've augmented the data matrix with prior expectations.

We'll assume that we've run PEB and so have estimated parameters, $\hat{\beta}$, and hyperparameters, $\hat{\lambda}$. We now wish to compute the model evidence $p(y|m)$.

From lecture 5 we know that

$$
\begin{aligned}
p(y|\lambda, m) &= (2\pi)^{-N/2}|V|^{-1/2} \\
&\times \exp\left(\frac{1}{2}(\bar{y} - \bar{X}\hat{\beta})^T V^{-1}(\bar{y} - \bar{X}\hat{\beta})\right) \\
&\times |\bar{X}^T V^{-1}\bar{X}|^{-1/2}
\end{aligned} \tag{8}
$$

By substuting in the expressions for V, $\bar{y}$ and $\bar{X}$ and taking logs we can write the log evidence as

$$\log p(y|\lambda, m) = Accuracy(m) - Complexity(m) \quad (9)$$

where

$$\begin{aligned} Accuracy(m) &= -\frac{1}{2}\log|C_1| - \frac{1}{2}(y - X\hat{\beta})^T C_1^{-1}(y - X\hat{\beta}) \\ Complexity(m) &= \frac{1}{2}\log|C_2| - \frac{1}{2}\log|\Sigma| + \frac{1}{2}(\mu - \hat{\beta})^T C_2^{-1}(\mu - \hat{\beta}) \end{aligned}$$

The second term is referred to as 'complexity' because eg. the quadratic term scales with the number of parameters in the model. A model with high evidence must therefore provide a good trade-off between accuracy and complexity.

This trade-off is also employed in other more ad-hoc model selection schemes eg. AIC and BIC have complexity terms embodying fixed costs for each parameter of 1 (AIC) and $\frac{1}{2}\log N$. See eg. [2] for more details.

4.1 Integrating out hyperparameters

To get the evidence $p(y|m)$ we must integrate out the uncertainty in the hyperparameters λ.

$$p(y|m) = \int p(y|\lambda, m) d\lambda \tag{10}$$

To do this we'll assume that the hyperparameters have a Gaussian distribution about their estimated value, $\hat{\lambda}$. As the hyperparameters must be positive we'll assume that this distribution is in log space. If we have a single hyperparameter then

$$p(y|\lambda, m) = p(y|\hat{\lambda}, m) \exp\left(-\frac{(\log\lambda - \log\hat{\lambda})^2}{2\sigma^2_{\log\lambda}}\right) \tag{11}$$

where $\sigma^2_{\log\lambda}$ is our uncertainty (variance) in the (log) estimated hyperparamater. We can then evaluate the integral to give

$$p(y|m) = p(y|\hat{\lambda}, m)(2\pi)^{1/2}\sigma_{\log\lambda} \tag{12}$$

The last terms are just the normalising constant for the Gaussian density. This expression for the evidence takes

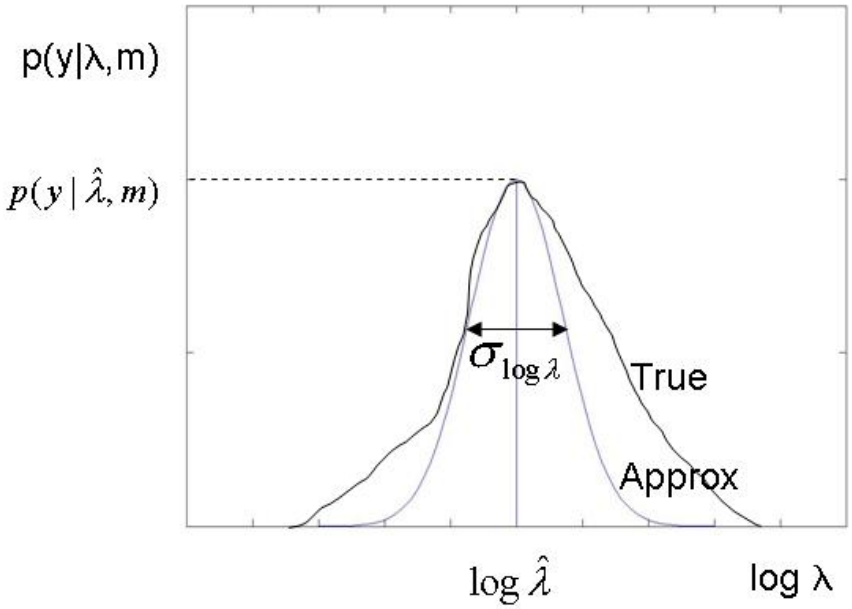

Figure 5: *Approximating the hyperparameter uncertainty with a Gaussian in log space.*

into account uncertainty in the estimation of the hypeparameters. If we have H hyperparameters then we get

$$p(y|m) = p(y|\hat{\lambda}, m)(2\pi)^{H/2} \prod_{h=1}^{H} \sigma_{\log \lambda_h} \tag{13}$$

5 Multimodal Imaging

Source reconstruction of EEG using fMRI location priors [1]. To 'reconstruct' EEG data at a *single time point* use the model

$$\begin{aligned} y &= X\beta + e_1 \\ \beta &= \mu + e_2 \end{aligned} \tag{14}$$

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. We use $\mu = 0$.

$$\begin{aligned} C_1 &= \sum_i \lambda_i Q_i \\ C_2 &= \sum_j \lambda_j Q_j \end{aligned} \tag{15}$$

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' or 'smoothness' matrix set up so that we expect the squared difference between neighboring voxels to be λ_j. Also consider extra Q_j's to incorporate valid and invalid location priors from eg. fMRI [1].

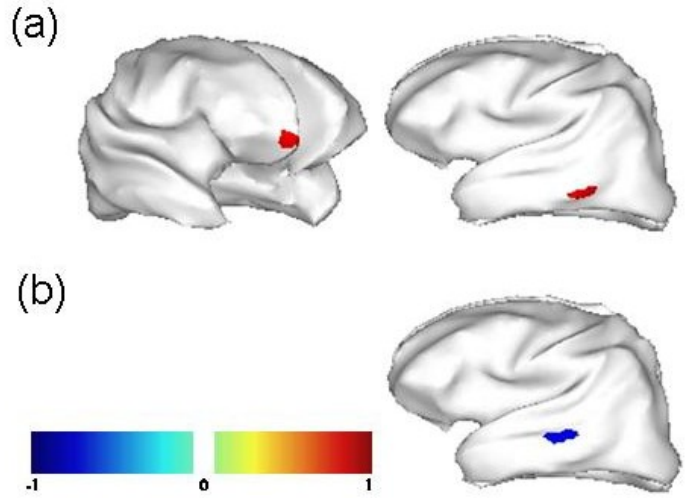

Figure 6: *Inflated cortical representation of (a) two simulated source locations ('valid' prior) and (b) 'invalid' prior location.*

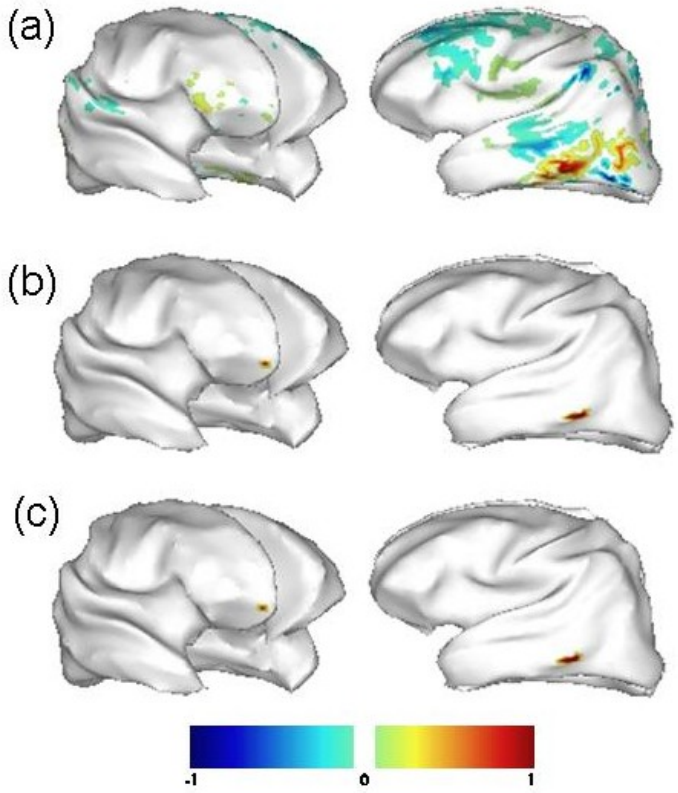

Figure 7: *Inflated cortical representation of representative source reconstructions using (a) smoothness prior, (b) smoothness and valid priors and (c) smoothness, valid and invalid priors. The reconstructed values have been normalised between -1 and 1.*

6 Nonlinear source reconstruction

Trujillo-Barreto et al. [4] describe a nonlinear source reconstruction algorithm based on combining reconstructions from a very large number of different models $m = 1..M$, using Bayesian Model Averaging (BMA)

$$p(\beta|y) = \sum_m p(\beta|y,m)p(m|y) \tag{16}$$

where $p(\beta|y,m)$ is the estimated CSD from model m and $p(m|y)$ is the posterior probability of model m. If all models are equilikely apriori then $p(m|y) = p(y|m)$. We therefore need to

- Fit model m to get CSD estimates
- Estimate model evidence $p(y|m)$
- Search model space M

Model space contains $M = 2^{71}$ models. There's no point fitting models that will have a low evidence. Use a greedy search strategy where eg. at search iteration i our model

contains regions 13, 40-45 and 62. Add/delete a region chosen uniformly at random and select it for iteration $i+1$ if evidence is higher. Keep all models with evidence greater than 1/20th of max so far - this is Occam's window of models.

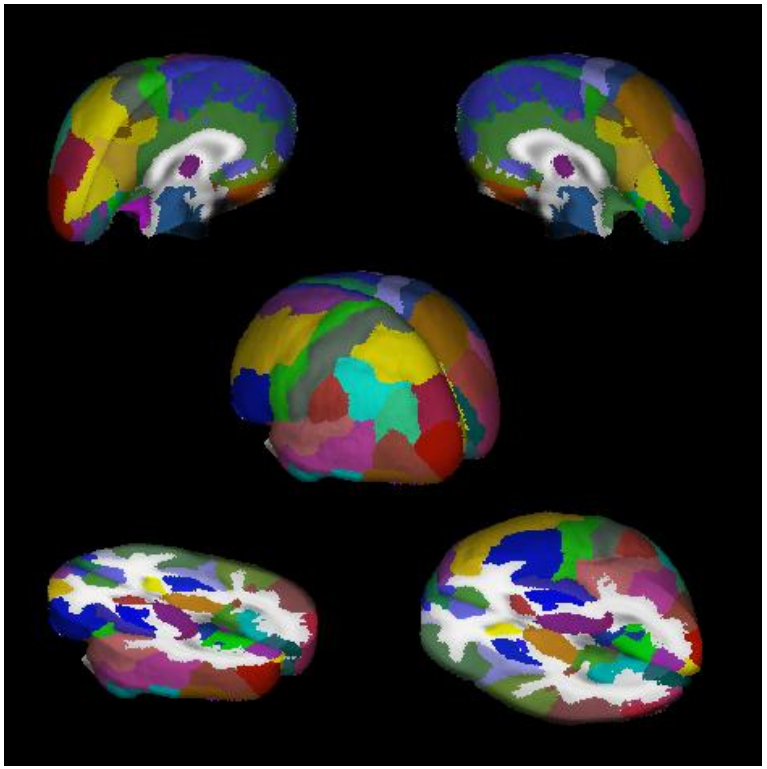

Figure 8: 3D segmentation of 71 structures of the Probabilistic MRI Atlas developed at the Montreal Neurological Institute. As shown in the color scale, brain areas belonging to different hemispheres were segmented separately.

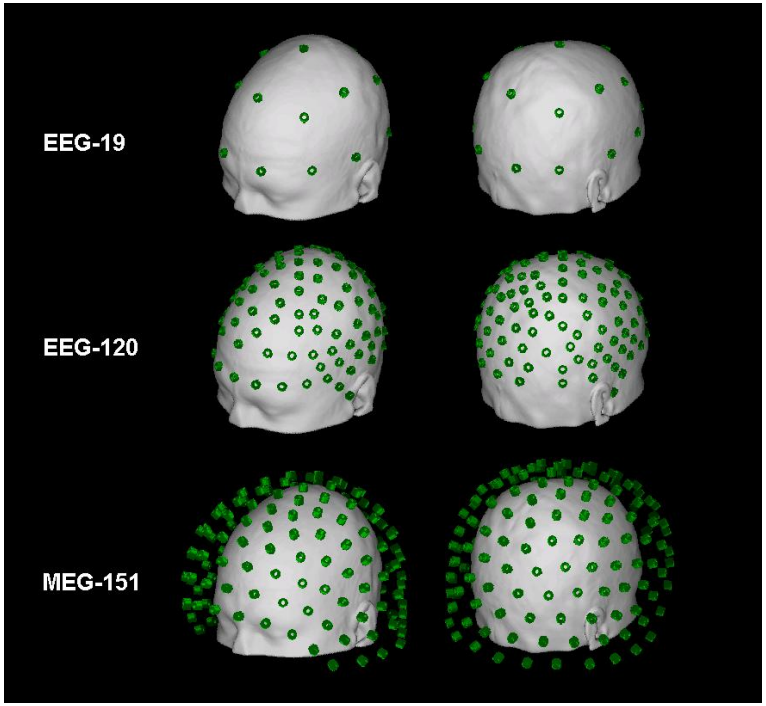

Figure 9: Different arrays of sensors used in the simulations. EEG-19 represents the 10/20 electrode system; EEG-120 is obtained by extending and refining the 10/20 system; and MEG-151 corresponds to the spatial configuration of MEG sensors in the helmet of the CTF System Inc.

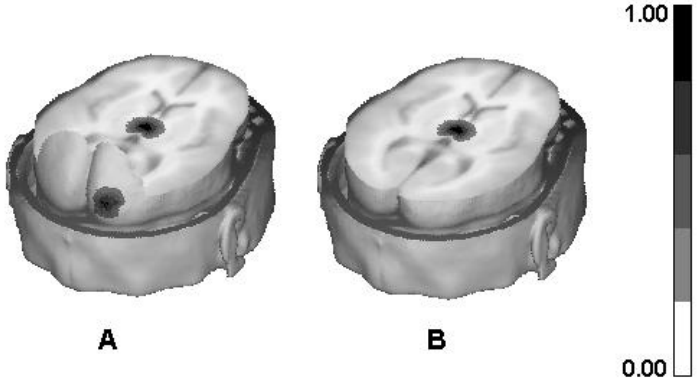

Figure 10: Spatial distributions of the simulated primary current densities. A) Simultaneous activation of two sources at different depths: one in the right Occipital Pole and the other in the Thalamus (OPR+TH). B) Simulation of a single source in the Thalamus (TH).

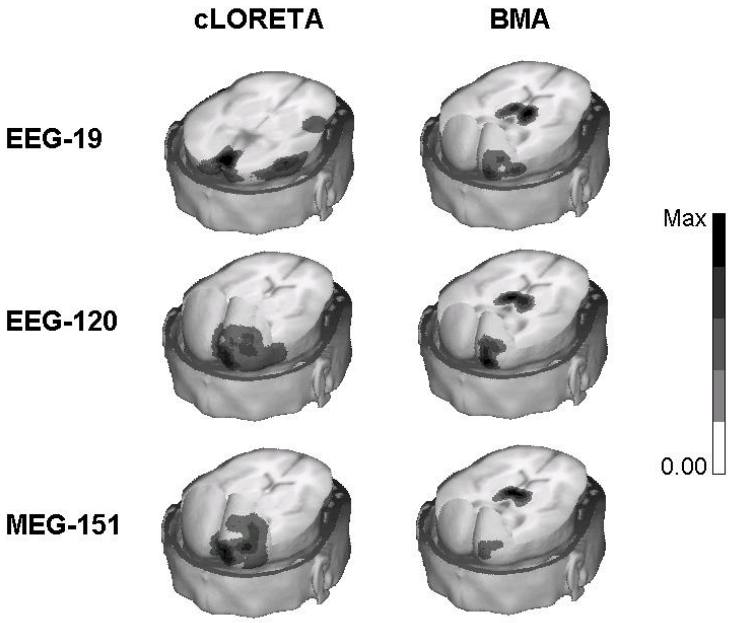

Figure 11: 3D reconstructions of the absolute values of BMA and cLORETA solutions for the OPR+TH source case. The first column indicates the array of sensors used in each simulated data set. The maximum of the scale is different for each case.

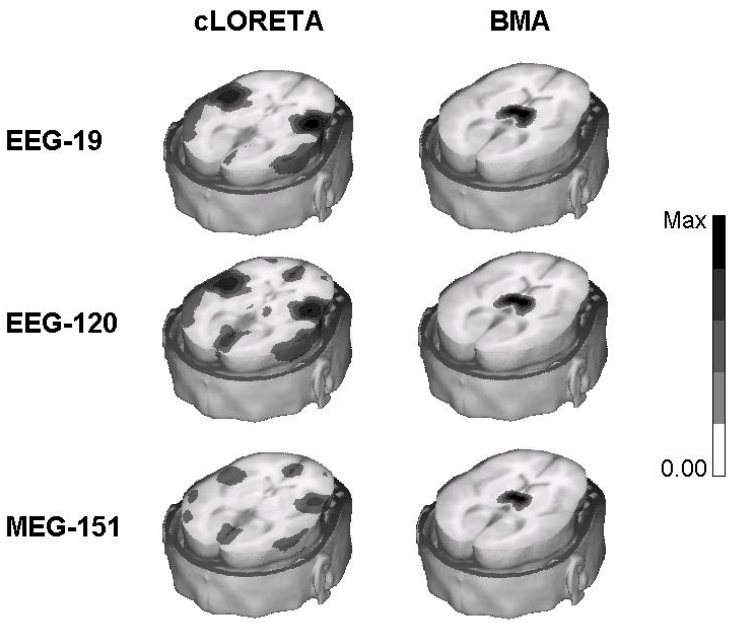

Figure 12: 3D reconstructions of the absolute values of BMA and cLORETA solutions for the TH source case. The first column indicates the array of sensors used in each simulated data set.

References

[1] J. Mattout, C. Phillips, W.D. Penny, M. Rugg, and K.J. Friston. Meg source localisation under multiple constraints: an extended Bayesian framework. *NeuroImage*, 2005.

[2] W.D. Penny, K.E. Stephan, A. Mechelli, and K.J. Friston. Comparing Dynamic Causal Models. *NeuroImage*, 22(3):1157–1172, 2004.

[3] R. Poldrack. Can cognitive processes be inferred from neuroimaging data ? *Trends in Cognitive Sciences*, 2006. In Press.

[4] N. Trujillo-Barreto, E. Aubert-Vazquez, and P. Valdes-Sosa. Bayesian model averaging in EEG/MEG imaging. *Neuroimage*, 21:1300–1319, 2004. In Press.