Chapter 35: Bayesian model selection and averaging

W.D. Penny, J.Mattout and N. Trujillo-Barreto

May 10, 2006

Introduction

In Chapter 11 we described how Bayesian inference can be applied to hierarchical models. In this chapter we show how the members of a model class, indexed by m, can also be considered as part of a hierarchy. Model classes might be GLMs where m indexes different choices for the design matrix, DCMs where m indexes connectivity or input patterns, or source reconstruction models where m indexes functional or anatomical constraints. Explicitly including model structure in this way will allow us to make inferences about that structure.

Figure 1 shows the generative model we have in mind. First, a member of the model class is chosen. Then model parameters θ and finally the data y are generated. Bayesian inference for hierarchical models can be implemented using the belief propagation algorithm. Figure 2 shows how this can be applied for model selection and averaging. It comprises three stages that we will refer to as (i) conditional parameter inference, (ii) model inference and (iii) model averaging. These stages can be implemented using the equations shown in Figure 2.

Conditional parameter inference is based on Bayes rule whereby, after observing data y, prior beliefs about model parameters are updated to posterior beliefs. This update requires the likelihood $p(y|\theta, m)$. It allows one to compute the density $p(\theta|y, m)$. The term conditional is used to highlight the fact that these inferences are based on model m. Of course, being a posterior density, it is also conditional on the data y.

Model inference is based on Bayes rule whereby, after observing data y, prior beliefs about model structure are updated to posterior beliefs. This update requires the evidence $p(y|m)$. Model selection is then implemented by picking the model that maximises the posterior probability $p(m|y)$. If the model priors $p(m)$ are uniform then this is equivalent to picking the model with the highest evidence. Pairwise model comparisons are based on Bayes factors, which are ratios of evidences.

Model averaging, as depicted in Figure 2, also allows for inferences to be made about parameters. But these inferences are based on the distribution $p(\theta|y)$, rather than $p(\theta|y, m)$, and so are free from assumptions about model structure.

This chapter comprises theoretical and empirical sections. In the theory sections we describe (i) conditional parameter inference for linear and nonlinear models (ii) model inference, including a review of three different ways to approximate model evidence and pairwise model comparisons based on Bayes factors and (iii) model averaging, with a focus on how to search the model space using 'Occam's window'. The empirical sections show how these principles can be applied to DCMs and source reconstruction models. We finish with a discussion.

Notation

We use upper-case letters to denote matrices and lower-case to denote vectors. $\mathsf{N}(m, \Sigma)$ denotes a uni/multivariate Gaussian with mean m and variance/covariance Σ. I_K denotes the $K \times K$ identity matrix, 1_K is a $1 \times K$ vector of ones, 0_K is a $1 \times K$ vector of zeros. If X is a matrix, X_{ij} denotes the i, jth element, X^T denotes the matrix transpose and $\mathsf{vec}(X)$ returns a column vector comprising its columns, $\mathsf{diag}(x)$ returns a diagonal matrix with leading diagonal elements given by the vector x, $\otimes$ denotes the Kronecker product and $\log x$ denotes the natural logarithm.

Conditional Parameter Inference

Readers requiring a more basic introduction to Bayesian modelling are referred to [Gelman et al. 1995], and chapter 11.

Linear models

For linear models

$$y = X\theta + e \tag{1}$$

with data y, parameters θ, Gaussian errors e and design matrix X, the likelihood can be written

$$p(y|\theta, m) = \mathsf{N}(X\theta, C_e) \tag{2}$$

where C_e is the error covariance matrix. If our prior beliefs can be specified using the Gaussian distribution

$$p(\theta|m) = \mathsf{N}(\mu_p, C_p) \tag{3}$$

where μ_p is the prior mean and C_p is the prior covariance, then the posterior distribution is [Lee 1997]

$$p(\theta|y, m) = \mathsf{N}(\mu, C) \tag{4}$$

where

$$\begin{aligned} C^{-1} &= X^T C_e^{-1} X + C_p^{-1} \\ \mu &= C(X^T C_e^{-1} y + C_p^{-1} \mu_p) \end{aligned} \tag{5}$$

As in Chapter 11, it is often useful to refer to precision matrices, C^{-1}, rather than covariance matrices, C. This is because the posterior precision, C^{-1}, is equal to the sum of the prior precision, C_p^{-1}, plus the data precision, $X^T C_e^{-1} X$. The posterior mean, μ, is given by the sum of the prior mean plus the data mean, but where each is weighted according to their relative precision. This linear Gaussian framework is used for the source reconstruction methods described later in the chapter. Here X is the lead-field matrix which transforms measurements from source space to sensor space [Baillet et al. 2001].

Our model assumptions, m, are typically embodied in different choices for the design or prior covariance matrices. These allow for the specification of GLMs with different regressors or different covariance components.

Variance components

Bayesian estimation, as described in the previous section, assumed that we knew the prior covariance, C_p, and error covariance, C_e. This information is, however, rarely available. In [Friston et al. 2002] these covariances are expressed as

$$\begin{aligned} C_p &= \sum_i \lambda_i Q_i \\ C_e &= \sum_j \lambda_j Q_j \end{aligned} \tag{6}$$

where Q_i and Q_j are known as 'covariance components' and λ_i, λ_j are hyperparameters. Chapter 24 and [Friston et al. 2002] show how these hyperparameters can be estimated using Parametric Empirical Bayes (PEB). It is also possible to represent precision matrices, rather than covariance matrices, using a linear expansion as shown in Chapter 47.

Nonlinear models

For nonlinear models, we have

$$y = h(\theta) + e \tag{7}$$

where $h(\theta)$ is a nonlinear function of parameter vector θ. We assume Gaussian prior and likelihood distributions

$$\begin{aligned} p(\theta|m) &= \mathsf{N}(\mu_p, C_p) \\ p(y|\theta, m) &= \mathsf{N}(h(\theta), C_e) \end{aligned} \tag{8}$$

where m indexes model structure, θ_p is the prior mean, C_p the prior covariance and C_e is the error covariance.

The linear framework described in the previous section can be applied by locally linearizing the nonlinearity, about a 'current' estimate μ_i, using a first order Taylor series expansion

$$h(\theta) = h(\mu_i) + \frac{\partial h(\mu_i)}{\partial \theta}(\theta - \mu_i) \tag{9}$$

Substituting this into 7 and defining $r \equiv y - h(\mu_i)$, $J \equiv \frac{\partial h(\mu_i)}{\partial \theta}$ and $\Delta\theta \equiv \theta - \mu_i$ gives

$$r = J\Delta\theta + e \tag{10}$$

which now conforms to a GLM (cf. equation 1). The 'prior' (based on starting estimate μ_i), likelihood and posterior are now given by

$$\begin{aligned} p(\Delta\theta|m) &= \mathsf{N}(\mu_p - \mu_i, C_p) \\ p(r|\Delta\theta, m) &= \mathsf{N}(J\Delta\theta, C_e) \\ p(\Delta\theta|r, m) &= \mathsf{N}(\Delta\mu, C_{i+1}) \end{aligned} \tag{11}$$

The quantities $\Delta\mu$ and C_{i+1} can be found using the result for the linear case (substitute r for y and J for X in equation 5). If we define our 'new' parameter estimate as $\mu_{i+1} = \mu_i + \Delta\mu$ then

$$\begin{aligned} C_{i+1}^{-1} &= J^T C_e^{-1} J + C_p^{-1} \\ \mu_{i+1} &= \mu_i + C_{i+1}(J^T C_e^{-1} r + C_p^{-1}(\mu_p - \mu_i)) \end{aligned} \tag{12}$$

This update is applied iteratively, in that the estimate μ_{i+1} becomes the starting point for a new Taylor series expansion. It can also be combined with hyperparameter estimates, to characterise C_p and C_e, as described in [Friston 2002]. This then corresponds to the PEB algorithm described in Chapter 22. This algorithm is used, for example, to estimate parameters of Dynamic Causal Models. For DCM, the nonlinearity $h(\theta)$ corresponds to the integration of a dynamic system.

As described, in chapter 24 this PEB algorithm is a special case of Variational Bayes with a fixed-form full-covariance Gaussian ensemble. When the algorithm has converged it provides an estimate of the posterior density

$$p(\theta|y, m) = \mathsf{N}(\mu_{PEB}, C_{PEB}) \tag{13}$$

which can then be used for parameter inference and model selection.

The above algorithm can also be viewed as the E-step of an EM algorithm, described in section 3.1 of [Friston 2002] and Chapter 46 in the appendices. The M-step of this algorithm, which we have not described, updates the hyperparameters. This E-step can also be viewed as a Gauss-Newton optimisation whereby parameter estimates are updated in the direction of the gradient of the log-posterior by an amount proportional to its curvature (see e.g. [Press et al. 1992]).

Model Inference

Given a particular model class, we let the variable m index members of that class. Model classes might be GLMs where m indexes design matrices, DCMs where m indexes connectivity or input patterns, or source reconstruction models where m indexes functional or anatomical constraints. Explicitly including model structure in this way will allow us to make inferences about model structure.

We may, for example, have prior beliefs $p(m)$. In the abscence of any genuine prior information here, a uniform distribution will suffice. We can then use Bayes rule which, in light of observed data y, will update these model priors into model posteriors

$$p(m|y) = \frac{p(y|m)p(m)}{p(y)} \tag{14}$$

Model inference can then proceed based on this distribution. This will allow for Bayesian Model Comparisons (BMCs). In Bayesian Model *Selection* (BMS), a model is selected which maximises this probability

$$m_{MP} = \underset{m}{\mathrm{argmax}}[p(m|y)]$$

If the prior is uniform, $p(m) = 1/M$ then this is equivalent to picking the model with the highest evidence

$$m_{ME} = \underset{m}{\mathrm{argmax}}[p(y|m)]$$

If we have uniform priors then BMC can be implemented with Bayes factors. Before covering this in more detail we emphasise that all of these model inferences require computation of the model evidence. This is given by

$$p(y|m) = \int p(y|\theta, m)p(\theta|m)d\theta$$

The model evidence is simply the normalisation term from parameter inference, as shown in Figure 2. This is the 'message' that is passed up the hierachy during belief propagation, as shown in Figure 2. For linear Gaussian models, the evidence can be expressed analytically. For non-linear models there are various approximations which are discussed in later subsections.

Bayes factors

Given models $m = i$ and $m = j$ the Bayes factor comparing model i to model j is defined as [Kass and Raftery 1993, Kass and Raftery 1995]

$$B_{ij} = \frac{p(y|m = i)}{p(y|m = j)} \tag{15}$$

where $p(y|m = j)$ is the evidence for model j. When $B_{ij} > 1$, the data favour model i over model j, and when $B_{ij} < 1$ the data favour model j. If there are more than two models to compare then we choose one of them as a reference model and calculate Bayes factors relative to that reference. When model i is an alternate model and model j a null model, B_{ij} is the likelihood ratio upon which classical statistics are based (see Chapter 44).

A classic example here is the analysis of variance for factorially designed experiments, described in chapter 13. To see if there is a main effect of a factor, one compares two models. One in which the levels of the factor are described by (i) a single variable or (ii) separate variables. Evidence in favour of model (ii) allows one to infer that there is a main effect.

In this chapter we will use Bayes factors to compare Dynamic Causal Models. In these applications, often the most important inference is on model space. For example, whether or not experimental effects are mediated by changes in feedforward or feedback pathways. This particular topic is dealt with in greater detail in Chapter 43.

The Bayes factor is a summary of the evidence provided by the data in favour of one scientific theory, represented by a statistical model, as opposed to another. Raftery [Raftery 1995] presents an interpretation of Bayes factors shown in Table 1. Jefferys [Jefferys 1935] presents a similar grading for the comparison of scientific theories. These partitionings are somewhat arbitrary but do provide descriptive statements.

Table 1 also shows the equivalent posterior probability of hypothesis i

$$p(m = i|y) = \frac{p(y|m = i)p(m = i)}{p(y|m = i)p(m = i) + p(y|m = j)p(m = j)} \tag{16}$$

assuming equal model priors $p(m = i) = p(m = j) = 0.5$.

If we define the 'prior odds ratio' as $p(m = i)/p(m = j)$ and the 'posterior odds ratio' as

$$O_{ij} = \frac{p(m = i|y)}{p(m = j|y)} \tag{17}$$

then the posterior odds is given by the prior odds multiplied by the Bayes factor. For prior odds of unity the posterior odds is therefore equal to the Bayes factor. Here, a Bayes factor of $B_{ij} = 100$, for example, corresponds to odds of 100-to-1. In betting shop parlance this is 100-to-1 'on'. A value of $B_{ij} = 0.01$ is 100-to-1 'against'.

Bayes factors in Bayesian statistics play a similar role to p-values in classical statistics. In [Raftery 1995], however, Raftery argues that p-values can give misleading results, especially in large samples. The background to this assertion is that Fisher originally suggested the use of significance levels (the p-values beyond which a result is deemed significant) $\alpha = 0.05$ or 0.01 based on his experience with small agricultural experiments having between 30 and 200 data points. Subsequent advice, notably from Neyman and Pearson, was that power and significance should be balanced when choosing α. This essentially corresponds to reducing α for large samples (but they did'nt say *how* α should be reduced). Bayes factors provide a principled way to do this.

The relation between p-values and Bayes factors is well illustrated by the following example [Raftery 1995]. For linear regression models one can use Bayes factors or p-values to decide whether to include an extra regressor. For a sample size of $N_s = 50$, positive evidence in favour of inclusion (say, $B_{12} = 3$) corresponds to a p-value of 0.019. For $N_s = 100$ and 1000 the corresponding p-values reduce to 0.01 and 0.003. If one wishes to decide whether to include multiple extra regressors the corresponding p-values drop more quickly.

Importantly, unlike p-values, Bayes factors can be used to compare models that cannot be nested [1]. This provides an optimal inference framework that can, for example, be applied to determine which hemodynamic basis functions are appropriate for fMRI [Penny et al. 2006]. They also allow one to quantify evidence in favour of a null hypothesis.

Computing the model evidence

This section shows how the model evidence can be computed for nonlinear models. The evidence for linear models is then given as a special case. The

[1] Model selection using classical inference requires nested models. Inference is made using step-down procedures and the 'extra sum of squares' principle, as described in Chapter 8.

Table 1. Interpretation of Bayes factors. Bayes factors can be interpreted as follows. Given candidate hypotheses i and j a Bayes factor of 20 corresponds to a belief of 95% in the statement 'hypothesis i is true'. This corresponds to strong evidence in favour of i.

B_{ij}	$p(m = i \vert y)(\%)$	Evidence in favour of model i
1 to 3	50-75	Weak
3 to 20	75-95	Positive
20 to 150	95-99	Strong
$\geq$ 150	$\geq$ 99	Very Strong

prior and likelihood of the nonlinear model can be expanded as

$$\begin{aligned} p(\theta|m) &= (2\pi)^{-p/2}|C_p|^{-1/2}\exp(-\frac{1}{2}e(\theta)^T C_p^{-1} e(\theta)) \\ p(y|\theta, m) &= (2\pi)^{-N_s/2}|C_e|^{-1/2}\exp(-\frac{1}{2}r(\theta)^T C_e^{-1} r(\theta)) \end{aligned} \quad (18)$$

where

$$\begin{aligned} e(\theta) &= \theta - \theta_p \\ r(\theta) &= y - h(\theta) \end{aligned} \quad (19)$$

are the 'parameter errors' and 'prediction errors'.

Substituting these expressions into equation 15 and re-arranging allows the evidence to be expressed as

$$p(y|m) = (2\pi)^{-p/2}|C_p|^{-1/2}(2\pi)^{-N_s/2}|C_e|^{-1/2} I(\theta) \quad (20)$$

where

$$I(\theta) = \int \exp(-\frac{1}{2}r(\theta)^T C_e^{-1} r(\theta) - \frac{1}{2}e(\theta)^T C_p^{-1} e(\theta)) d\theta \quad (21)$$

For linear models this integral can be expressed analytically. For nonlinear models it can be estimated using a Laplace approximation.

Laplace approximation

The Laplace approximation was introduced in Chapter 24. It makes use of the first order Taylor series approximation referred to in equation 9, but this time placed around the solution, θ_L, found by an optimisation algorithm.

Usually, the term 'Laplace approximation' refers to an expansion around the Maximum a Posterior (MAP) solution

$$\theta_{MAP} = \underset{\theta}{\operatorname{argmax}} \left[p(y|\theta, m) p(\theta|m) \right] \quad (22)$$

Thus $\theta_L = \theta_{MAP}$.

But more generally one can make an expansion around any solution, for example the one provided by PEB. In this case $\theta_L = \mu_{PEB}$. As we have described in Chapter 24, PEB is a special case of VB with a fixed-form Gaussian ensemble, and so does not deliver the MAP solution. Rather, PEB maximises the negative free energy and so implicitly minimises the KL-divergence between the true posterior and a full-covariance Gaussian approximation to it. This difference is discussed in Chapter 24.

Whatever the expansion point, the model nonlinearity is approximated using

$$h(\theta) = h(\theta_L) + J(\theta - \theta_L) \tag{23}$$

where $J = \frac{\partial h(\theta_L)}{\partial \theta}$. We also make use of the knowledge that the posterior covariance is given by

$$C_L^{-1} = J^T C_e^{-1} J + C_p^{-1} \tag{24}$$

For $C_L = C_{PEB}$ this follows directly from equation 12.

By using the substitutions $e(\theta) = (\theta - \theta_L) + (\theta_L - \theta_p)$ and $r(\theta) = (y - h(\theta_L)) + (h(\theta_L) - h(\theta))$, making use of the above two expressions, and removing terms not dependent on θ, we can write

$$\begin{aligned} I(\theta) &= \left[\int \exp(-\frac{1}{2}(\theta - \theta_L)^T C_L^{-1}(\theta - \theta_L))d\theta\right] & (25) \\ &\times \left[\exp(-\frac{1}{2}r(\theta_L)^T C_e^{-1} r(\theta_L) - \frac{1}{2}e(\theta_L)^T C_p^{-1} e(\theta_L))\right] & (26) \end{aligned}$$

where the first factor is the normalising term of the multivariate Gaussian density. The algebraic steps involved in the above substitutions are detailed in [Stephan et al. 2005]. Hence

$$\begin{aligned} I(\theta) &= (2\pi)^{p/2}|C_L|^{1/2} \exp(-\frac{1}{2}r(\theta_L)^T C_e^{-1} r(\theta_L) & (27) \\ &- \frac{1}{2}e(\theta_L)^T C_p^{-1} e(\theta_L)) \end{aligned}$$

Substituting this expression into 20 and taking logs gives the Laplace approximation to the log-evidence

$$\begin{aligned} \log p(y|m)_L &= -\frac{N_s}{2}\log 2\pi - \frac{1}{2}\log|C_e| - \frac{1}{2}\log|C_p| + \frac{1}{2}\log|C_L| & (28) \\ &- \frac{1}{2}r(\theta_L)^T C_e^{-1} r(\theta_L) - \frac{1}{2}e(\theta_L)^T C_p^{-1} e(\theta_L) \end{aligned}$$

When comparing the evidence for different models we can ignore the first term as it will be the same for all models. Dropping the first term and rearranging gives

$$\log p(y|m)_L = Accuracy(m) - Complexity(m) \tag{29}$$

where

$$Accuracy(m) = -\frac{1}{2}\log|C_e| - \frac{1}{2}r(\theta_L)^T C_e^{-1} r(\theta_L) \tag{30}$$

$$Complexity(m) = \frac{1}{2}\log|C_p| - \frac{1}{2}\log|C_L| + \frac{1}{2}e(\theta_L)^T C_p^{-1} e(\theta_L)$$

Use of base-e or base-2 logarithms leads to the log-evidence being measured in 'nats' or 'bits' respectively. Models with high evidence optimally trade-off two conflicting requirements of a good model, that it fits the data and be as simple as possible.

The complexity term depends on the prior covariance, C_p, which determines the 'cost' of parameters. This dependence is worrisome if the prior covariances are fixed a-priori, as the parameter cost will also be fixed a-priori. This will lead to biases in the resulting model comparisons. For example, if the prior (co)variances are set to large values, model comparison will consistently favour models that are less complex than the true model.

In DCM for fMRI [Friston et al. 2003], prior variances are set to fixed values so as to enforce dynamic stability, with high probability. Use of the Laplace approximation in this context could therefore lead to biases in model comparison. A second issue in this context is that, to enforce dynamic stability, models with different numbers of connections will employ different prior variances. Therefore the priors change from model to model. This means that model comparison entails a comparison of the priors.

To overcome these potential problems with DCM for fMRI, alternative approximations to the model evidence are used instead. These are the BIC and AIC introduced below. They also use fixed parameter costs, but they are fixed between models and are different for BIC than AIC. It is suggested in [Penny et al. 2004], that if the two measures provide consistent evidence, a model selection can be made.

Finally, we note that if prior covariances are estimated from data then the parameter cost will also have been estimated from data, and this source of bias in model comparison is removed. In this case, the model evidence also includes terms which account for uncertainty in the variance component estimation, as described in Chapter 10 of [Bishop 1995].

Bayesian Information Criterion

An alternative approximation to the model evidence is given by the Bayesian Information Criterion [Schwarz 1978]. This is a special case of the Laplace approximation which drops all terms that don't scale with the number of data points, and can be derived as follows.

Substituting Eq. 27 into Eq. 20 gives

$$p(y|m)_L = p(y|\theta_L, m)p(\theta_L|m)(2\pi)^{p/2}|C_L|^{1/2} \tag{31}$$

Taking logs gives

$$\log p(y|m)_L = \log p(y|\theta_L, m) + \log p(\theta_L|m) + \frac{p}{2}\log 2\pi + \frac{1}{2}\log|C_L| \tag{32}$$

The dependence of the first three terms on the number of data points is $O(N_s)$, $O(1)$ and $O(1)$. For the 4th term, entries in the posterior covariance scale linearly with N_s^{-1}

$$\begin{aligned}\lim_{N_s \to \infty} \frac{1}{2} \log |C_L| &= \frac{1}{2} \log |\frac{C_L(0)}{N_s}| \\ &= -\frac{p}{2} \log N_s + \frac{1}{2} \log |C_L(0)|\end{aligned} \tag{33}$$

where $C_L(0)$ is the posterior covariance based on $N_s = 0$ data points (ie. the prior covariance). This last term therefore scales as $O(1)$. Schwarz [Schwarz 1978] notes that in the limit of large N_s equation 32 therefore reduces to

$$\begin{aligned}BIC &= \lim_{N_s \to \infty} \log p(y|m)_L \\ &= \log p(y|\theta_L, m) - \frac{p}{2} \log N_s\end{aligned} \tag{34}$$

This can be re-written as

$$BIC = Accuracy(m) - \frac{p}{2} \log N_s \tag{35}$$

where p is the number of parameters in the model. In BIC, the cost of a parameter, $-0.5 \log N_s$ bits, therefore reduces with an increasing number of data points.

Akaike's Information Criterion

The second criterion we use is Akaike's Information Criterion (AIC)[2] [Akaike 1973]. AIC is maximised when the approximating likelihood of a novel data point is closest to the true likelihood, as measured by the Kullback-Liebler divergence (this is shown in [Ripley 1995]). The AIC is given by

$$AIC = Accuracy(m) - p \tag{36}$$

Though not originally motivated from a Bayesian perspective, model comparisons based on AIC are asymptotically equivalent (ie. as $N_s \to \infty$) to those based on Bayes factors [Akaike 1983], ie. AIC approximates the model evidence.

Empirically, BIC is biased towards simple models and AIC to complex models [Kass and Raftery 1993]. Indeed, inspection of Equations 35 and 36 shows that for values appropriate for eg. DCM for fMRI, where $p \approx 10$ and $N_s \approx 200$, BIC pays a heavier parameter penalty than AIC.

Model averaging

The parameter inferences referred to in previous sections are based on the distribution $p(\theta|y,m)$. That m appears as a dependent variable, makes it explicit that these inferences are contingent on assumptions about model structure. More generally, however, if inferences about model parameters are paramount one would use a BMA approach. Here, inferences are based on the distribution

$$p(\theta|y) = \sum_m p(\theta|y,m) p(m|y) \tag{37}$$

[2] Strictly, AIC should be referred to as An Information Criterion.

where $p(m|y)$ is the posterior probability of model m.

$$p(m|y) = \frac{p(y|m)p(m)}{p(y)} \tag{38}$$

As shown in Figure 2, only when these 'messages', $p(m|y)$, have been passed back down the hierarchy is belief propagation complete. Only then do we have the true marginal density $p(\theta|y)$. Thus, BMA allows for correct Bayesian inferences, whereas what we have previously described as 'parameter inferences' are conditional on model structure. Of course, if our model space comprises just one model there is no distribution.

BMA accounts for uncertainty in the model selection process, something which classical statistical analysis neglects. By averaging over competing models, BMA incorporates model uncertainty into conclusions about parameters. BMA has been successfully applied to many statistical model classes including linear regression, generalised linear models, and discrete graphical models, in all cases improving predictive performance. See [Hoeting et al. 1999] for a review [3]. In this Chapter we describe the application of BMA to EEG source reconstruction.

There are, however, several practical difficulties with expression 37 when the number of models and numbers of variables in each model are large. In neuroimaging, models can have tens of thousands of parameters. This issue has been widely treated in the literature [Draper 1995], and the general consensus has been to construct search strategies to find a set of models that are 'worth taking into account'. One of these strategies is to generate a Markov chain to explore the model space and then approximate equation 37 using samples from the posterior $p(m|y)$ [Madigan 1992]. But this is computationally very expensive.

In this Chapter we will instead use the Occam's Window procedure for nested models described in [Madigan 1994]. First, a model that is N_0 times less likely a posteriori than the maximum posterior model is removed (in this Chapter we use $N_0 = 20$). Second, complex models with posterior probabilities smaller than their simpler counterparts are also excluded. The remaining models fall in Occam's window. This leads to the following approximation to the posterior density

$$p(\theta|y) = \sum_{m \epsilon C} p(\theta|y,m)p(m|y) \tag{39}$$

where the set C identifies 'Occam's Window'. Models falling in this window can be identified using the search strategy defined in [Madigan 1994].

Dynamic Causal Models

The term 'causal' in DCM arises because the brain is treated as a deterministic dynamical system (see eg. section 1.1 in [Friston et al. 2003]) in which external inputs cause changes in neuronal activity which in turn cause changes in the resulting fMRI, MEG or EEG signal. DCMs for fMRI comprise a bilinear model for the neurodynamics and an extended Balloon model [Friston 2002,

[3] Software is also available from $http://www.research.att.com/\ volinsky/bma.html$.

Buxton 1998] for the hemodynamics. These are described in detail in Chapter 41.

The effective connectivity in DCM is characterised by a set of 'intrinsic connections', that specify which regions are connected and whether these connections are unidirectional or bidirectional. We also define a set of input connections that specify which inputs are connected to which regions, and a set of modulatory connections that specify which intrinsic connections can be changed by which inputs. The overall specification of input, intrinsic and modulatory connectivity comprise our assumptions about model structure. This in turn represents a scientific hypothesis about the structure of the large-scale neuronal network mediating the underlying cognitive function. Examples of DCMs are shown in Figure 5.

Attention to Visual Motion

In previous work we have established that attention modulates connectivity in a distributed system of cortical regions that subtend visual motion processing [Buchel and Friston 1997, Friston and Buchel 2000]. These findings were based on data acquired using the following experimental paradigm. Subjects viewed a computer screen which displayed either a fixation point, stationary dots or dots moving radially outward at a fixed velocity. For the purpose of our analysis we can consider three experimental variables. The 'photic stimulation' variable indicates when dots were on the screen, the 'motion' variable indicates that the dots were moving and the 'attention' variable indicates that the subject was attending to possible velocity changes. These are the three input variables that we use in our DCM analyses and are shown in Figure 3.

In this paper we model the activity in three regions V1, V5 and superior parietal cortex (SPC). The original 360-scan time series were extracted from the data set of a single subject using a local eigendecomposition and are shown in Figure 4.

We initially set up three DCMs, each embodying different assumptions about how attention modulates connections to V5. Model 1 assumes that attention modulates the forward connection from V1 to V5, model 2 assumes that attention modulates the backward connection from SPC to V5 and model 3 assumes attention modulates both connections. These models are shown in Figure 5. Each model assumes that the effect of motion is to modulate the connection from V1 to V5 and uses the same reciprocal hierarchical intrinsic connectivity.

We fitted the models and computed Bayes factors shown in Table 2. We did not use the Laplace approximation to the model evidence, as DCM for fMRI uses fixed prior variances which compound model comparison, as described in section 3.2.1. Instead, we computed both AIC and BIC and made an inference only if the two resulting Bayes factors were consistent [Penny et al. 2004].

Table 2 shows that the data provide consistent evidence in favour of the hypothesis embodied in model 1, that attention modulates solely the forward connection from V1 to V5.

We now look more closely at the comparison of model 1 to model 2. The estimated connection strengths of the attentional modulation were 0.23 for the forward connection in model 1 and 0.55 for the backward connection in model 2. This shows that attentional modulation of the backwards connection is stronger than the forwards connection. However, a breakdown of the Bayes factor B_{12}

Table 2. Attention Data - comparing modulatory connectivities Bayes factors provide consistent evidence in favour of the hypothesis embodied in model 1, that attention modulates (solely) the bottom-up connection from V1 to V5. Model 1 is preferred to models 2 and 3. Models 1 and 2 have the same number of connections so AIC and BIC give identical values.

	B_{12}	B_{13}	B_{32}
AIC	3.56	2.81	1.27
BIC	3.56	19.62	0.18

in table 3 shows that the reason model 1 is favoured over model 2 is because it is more accurate. In particular, it predicts SPC activity much more accurately. Thus, although model 2 does show a significant modulation of the SPC-V5 connection, the required change in its prediction of SPC activity is sufficient to compromise the overall fit of the model. If we assume models 1 and 2 are equally likely apriori then our posterior belief in model 1 is 0.78 (from 3.56/(3.56+1)). Thus, model 1 is the favoured model even though the effect of attentional modulation is weaker.

This example makes an important point. Two models can only be compared by computing the evidence for each model. It is not sufficient to compare values of single connections. This is because changing a single connection changes overall network dynamics and each hypothesis is assessed (in part) by how well it predicts the data, and the relevant data are the activities in a distributed network.

We now focus on model 3 that has *both* modulation of forward and backward connections. Firstly, we make a statistical inference to see if, within model 3, modulation of the forward connection is larger than modulation of the backward connection. For these data the posterior distribution of estimated parameters tells us that this is the case with probability 0.75. This is a different sort of inference to that made above. Instead of inferring which is more likely, modulation of a forward or backward connection, we are making an inference about which effect is stronger when both are assumed present.

However, this inference is contingent on the assumption that model 3 is a good model. It is based on the density $p(\theta|y, m = 3)$. The Bayes factors in Table 2, however, show that the data provide consistent evidence in favour of the hypothesis embodied in model 1, that attention modulates *only* the forward connection. Table 4 shows a breakdown of B_{13}. Here the largest contribution to the Bayes factor (somewhere between 2.72 and 18.97) is the increased parameter cost for model 3.

The combined use of Bayes factors and DCM provides us with a formal method for evaluating competing scientific theories about the forms of large-scale neural networks and the changes in them that mediate perception and cognition. These issues are pursued in Chapter 43 in which DCMs are compared so as to make inferences about inter-hemispheric integration from fMRI data.

Source reconstruction

A comprehensive introduction to source reconstruction is provided in [Baillet et al. 2001]. For more recent developments see [Michel et al. 2004] and Chapters 28 to 30.

Table 3. Attention Data: Breakdown of contributions to the Bayes factor for model 1 versus model 2. The largest single contribution to the Bayes factor is the increased model accuracy in region SPC, where 8.38 fewer bits are required to code the prediction errors. The overall Bayes factor B_{12} of 3.56 provides consistent evidence in favour of model 1.

Source	Model 1 vs. Model 2 Relative Cost (bits)	Bayes Factor B_{12}
V1 accuracy	7.32	0.01
V5 accuracy	-0.77	1.70
SPC accuracy	-8.38	333.36
Complexity (AIC)	0.00	1.00
Complexity (BIC)	0.00	1.00
Overall (AIC)	-1.83	3.56
Overall (BIC)	-1.83	3.56

Table 4. Attention Data: Breakdown of contributions to the Bayes factor for model 1 versus model 3. The largest single contribution to the Bayes factor is the cost of coding the parameters. The table indicates that both models are similarly accurate but model 1 is more parsimonious. The overall Bayes factor B_{13} provides consistent evidence in favour of the (solely) bottom-up model.

Source	Model 1 vs. Model 3 Relative Cost (bits)	Bayes Factor B_{13}
V1 accuracy	-0.01	1.01
V5 accuracy	0.02	0.99
SPC accuracy	-0.05	1.04
Complexity (AIC)	-1.44	2.72
Complexity (BIC)	-4.25	18.97
Overall (AIC)	-1.49	2.81
Overall (BIC)	-4.29	19.62

The aim of source reconstruction is to estimate sources, θ, from sensors, y, where

$$y = X\theta + e \tag{40}$$

e is an error vector and X defines a lead-field matrix. Distributed source solutions usually assume a Gaussian prior for

$$p(\theta) = \mathsf{N}(\mu_p, C_p) \tag{41}$$

Parameter inference for source reconstruction can then be implemented as described in the section above on linear models. Model inference can be implemented using the expression in equation 29. For the numerical results in this paper we augmented this expression to account for uncertainty in the estimation of the hyperparameters. The full expression for the log-evidence of hyperparameterised models under the Laplace approximation is described in [Trujillo-Barreto et al. 2004] and Chapter 47.

Multiple constraints

This section considers source reconstruction with multiple constraints. This topic is covered in greater detail and from a different perspective in Chapters 29 and 30. The constaints are implemented using a decomposition of the prior covariance into distinct components

$$C_p = \sum_i \lambda_i Q_i \tag{42}$$

The first type of constraint is a smoothness constraint, Q_{sc}, based on the usual L^2-norm. The second is an intrinsic functional constraint, Q_{int}, based on Multivariate Source Prelocalisation (MSP) [Mattout et al. 2005]. This provides an estimate, based on a multivariate characterisation of the M/EEG data itself. Thirdly, we used extrinsic functional constraints which were considered as 'valid', Q^v_{ext}, or 'invalid', Q^i_{ext}. These extrinsic constraints are derived from other imaging modalities such as fMRI. We used invalid constraints to test the robustness of the source reconstructions.

To test the approach, we generated simulated sources from the locations shown in Figure 6a. Temporal activity followed a half-period sine function with a period of 30ms. This activity was projected onto 130 virtual MEG sensors and Gaussian noise was then added. Further details on the simulations are given in [Mattout et al. 2006].

We then reconstructed the sources using all combinations of the various constraints. Figure 7 shows a sample of source reconstructions. Table 5 shows the evidence for each model which we computed using the Laplace approximation (which is exact for these linear Gaussian models). As expected, the model with the single valid location prior had the highest evidence.

Further, any model which contains the valid location prior has high evidence. The table also shows that any model which contains both valid and invalid location priors does not show a dramatic decrease in evidence, compared to the same model without the invalid location prior. These trends can be assessed more formally by computing the relevant Bayes factors, as shown in table 6. This shows significantly enhanced evidence in favor of models including valid location priors. It also suggests that the smoothness and intrinsic location priors can ameliorate the misleading effect of invalid priors.

Model averaging

In this section we consider source localisations with anatomical constraints. A class of source reconstruction models is defined where, for each model, activity is assumed to derive from a particular anatomical 'compartment' or combination of compartments. Anatomical compartments are defined by taking 71 brain regions, obtained from a 3D segmentation of the Probabilistic MRI Atlas (PMA) [Evans et al. 1993] shown in Figure 8. These compartments preserve the hemispheric symmetry of the brain, and include deep areas like thalamus, basal ganglia and brain stem. Simple activations may be localised to single compartments and more complex activations to combinations of compartments. These combinations define a nested family of source reconstruction models which can be searched using the Occam's window approach described in section 4.

The source space consists of a 3D-grid of points that represent the possible generators of the EEG/MEG inside the brain, while the measurement space is defined by the array of sensors where the EEG/MEG is recorded. We used a 4.25 mm grid spacing and different arrays of electrodes/coils are placed in registration with the PMA. The 3D-grid is further clipped by the gray matter, which consists of all brain regions segmented and shown in figure 8.

Three arrays of sensors were used and are depicted in figure 9. For EEG simulations a first set of 19 electrodes (EEG-19) from the 10/20 system is chosen. A second configuration of 120 electrodes (EEG-120) is also used in order to investigate the dependence of the results on the number of sensors. Here, electrode positions were determined by extending and refining the 10/20 system. For MEG simulations, a dense array of 151 sensors were used (MEG-151). The physical models constructed in this way, allow us to compute the electric/magnetic lead field matrices that relate the Primary Current Density (PCD) inside the head, to the voltage/magnetic field measured at the sensors.

We now present the results of two simulation studies. In the first study two distributed sources were simulated. One source was located in the right occipital pole, and the other in the thalamus. This simulation is referred to as 'OPR+TH'. The spatial distribution of PCD (ie. the true θ vector) was generated using two narrow Gaussian functions of the same amplitude shown in figure 10A.

The temporal dynamics were specified using a linear combination of sine functions with frequency components evenly spaced in the alpha band (8-12Hz). The amplitude of the oscillation as a function of frequencies is a narrow Gaussian peaked at 10Hz. That is, activity is given by

$$j(t) = \sum_{i=1}^{N} \exp(-8(f_i - 10)^2) \sin(2\pi f_i t) \tag{43}$$

where $8 \leq f_i \leq 12$Hz. Here, f_i is the frequency and t denotes time. These same settings are then used for the second simulation study, in which only the thalamic (TH) source was used (see figure 10B). This second simulation is referred to as 'TH'. In both cases the measurements were generated with a Signal to Noise Ratio (SNR) of 10.

The simulated data were then analysed using Bayesian Model Averaging (BMA) in order to reconstruct the sources. We searched through model space using the Occam's window approach described in section 4. For comparison, we also applied the constrained Low Resolution Tomography (cLORETA) algorithm. This method constrains the solution to gray matter and again uses the usual L^2-norm. The cLORETA model is included in the model class used for BMA, and corresponds to a model comprising all 71 anatomical compartments.

The absolute values of the BMA and cLORETA solutions for the OPR+TH example, and for the three arrays of sensors used, are depicted in figure 11. In all cases, cLORETA is unable to recover the TH source and the OPR source estimate is overly dispersed. For BMA, the spatial localizations of both cortical and subcortical sources are recovered with reasonable accuracy in all cases. These results suggest that the EEG/MEG contains enough information for estimating deep sources, even in cases where such generators might be hidden by cortical activations.

The reconstructed sources shown in figure 12 for the TH case show that cLORETA suffers from a 'depth biasing' problem. That is, deep sources are misattributed to superficial sources. This biasing is not due to masking effects, since no cortical source is present in this set of simulations. Again, BMA gives significantly better estimates of the PCD.

Figures 11 and 12 also show that the reconstructed sources become more concentrated and clearer, as the number of sensors increases. Tables 7 and 8 show the number of models in Occam's window for each simulation study. The number of models reduces with increasing number of sensors. This is natural since more precise measurements imply more information available about the underlying phenomena, and then narrower and sharper model distributions are obtained. Consequently, as shown in the table, the probability and hence, the rank of the true model in the Occam's Window increases for dense arrays of sensors.

Tables 7 and 8 also show that the model with the highest probability is not always the true one. This fact supports the use of BMA instead of using the maximum posterior or maximum evidence model. In the present simulations, this is not critical, since the examples analyzed are quite simple. But it becomes a determining factor when analyzing more complex data, as is the case with some real experimental conditions [Trujillo-Barreto et al. 2004].

An obvious question then arises. Why is cLORETA unable to fully exploit the information contained in the M/EEG? The answer given by Bayesian inference is simply that cLORETA, which assumes activity is distributed over all of gray matter, is not a good model. In the model averaging framework, the cLORETA model was always rejected due to its low posterior probability, placing it outside Occam's window.

Discussion

Chapter 11 showed how Bayesian inference in hierarchical models can be implemented using the belief propagation algorithm. This involves passing messages up and down the hierarchy, the upward messages being likelihoods and evidences and the downward messages being posterior probabilities.

In this Chapter we have shown how belief propagation can be used to make inferences about members of a model class. Three stages were identified in this process: (i) conditional parameter inference, (ii) model inference and (iii) model averaging. Only at the model averaging stage is belief propagation complete. Only then will parameter inferences be based on the correct marginal density.

We have described how this process can be implemented for linear and nonlinear models and applied to domains such as Dynamic Causal Modelling and M/EEG source reconstruction. In DCM, often the most important inference to be made is a model inference. This can be implemented using Bayes factors and allows one to make inferences about the structure of large scale neural networks that mediate cognitive and perceptual processing. This issue is taken further in Chapter 43 which considers inter-hemispheric integration.

The application of model averaging to M/EEG source reconstruction results in the solution of an outstanding problem in the field. That is, how to detect deep sources. Simulations show that a standard method (cLORETA) is simply not a good model and that model averaging can combine the estimates of better models to make veridical source estimates.

The use of Bayes factors for model comparison is somewhat analagous to the use of F-tests in the General Linear Model. Whereas t-tests are used to assess individual effects, F-tests allow one to assess the significance of a set of effects. This is achieved by comparing models with and without the set of effects of interest. The smaller model is 'nested' within the larger one. Bayes factors play a similar role but additionally allow inferences to be constrained by prior knowledge. Moreover, it is possible to simultaneously entertain a number of hypotheses and compare them using the model evidence. Importantly, these hypotheses are not constrained to be nested.

References

[Akaike 1973] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B.N. Petrox and F. Caski, editors, *Second international symposium on information theory*, page 267, 1973. Budapest: Akademiai Kiado.

[Akaike 1983] H. Akaike. Information Measures and Model Selection. Bulletin of the International Statistical Institute, 50,277–290, 1983.

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

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

[Buchel and Friston 1997] C. Buchel and K.J. Friston. Modulation of connectivity in visual pathways by attention: Cortical interactions evaluated with structural equation modelling and fMRI. *Cerebral Cortex*, 7:768–778, 1997.

[Buxton 1998] R.B. Buxton, E.C. Wong, and L.R. Frank. Dynamics of blood flow and oxygenation changes during brain activation: The Balloon Model. *Magnetic Resonance in Medicine*, 39:855–864, 1998.

[Draper 1995] D. Draper. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society Series B*, 57:45–97, 1995.

[Evans et al. 1993] A. Evans, D. Collins, S. Mills, E. Brown, R. Kelly, and T. Peters. 3D statistical neuroanatomical models from 305 mri volumes. In *Proceedings IEEE Nuclear Science Symposium and Medical Imaging Conference*, 1993.

[Friston 2002] K.J. Friston. Bayesian estimation of dynamical systems: An application to fMRI. *NeuroImage*, 16:513–530, 2002.

[Friston and Buchel 2000] K.J. Friston and C. Buchel. Attentional modulation of effective connectivity from V2 to V5/MT in humans. *Proc. Natl. Acad. Sci. USA*, 97(13):7591–7596, 2000.

[Friston et al. 2002] 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.

[Friston et al. 2003] K.J. Friston, L. Harrison, and W.D. Penny. Dynamic Causal Modelling. *NeuroImage*, 19(4):1273–1302, 2003.

[Gelman et al. 1995] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis.* Chapman and Hall, Boca Raton, 1995.

[Hoeting et al. 1999] J.A. Hoeting, D. Madigan, A.E. Raftery, and C.T. Volinsky. Bayesian Model Averaging: A Tutorial. *Statistical Science*, 14(4):382–417, 1999.

[Jefferys 1935] H. Jefferys. Some tests of significance, treated by the theory of probability. *Proceedings of the Cambridge Philosophical Society*, 31:203–222, 1935.

[Kass and Raftery 1993] R.E. Kass and A.E. Raftery. Bayes factors and model uncertainty. Technical Report 254, University of Washington, 1993. http://www.stat.washington.edu/tech.reports/tr254.ps.

[Kass and Raftery 1995] R.E. Kass and A.E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.

[Lee 1997] P. M. Lee. *Bayesian Statistics: An Introduction.* Arnold, 2 edition, 1997.

[Madigan 1994] D. Madigan and A. Raftery. Model selection and accounting for uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, 89:1535–1546, 1994.

[Madigan 1992] D. Madigan and J. York. Bayesian graphical models for discrete data. Technical report, Department of Statistics, University of Washington, 1992. Report number 259.

[Mattout et al. 2005] J. Mattout, M. Pelegrini-Isaac, L. Garnero, and H. Benali. Multivariate Source Prelocalisation: use of functionally informed basis functions for better conditioning the MEG inverse problem. *Neuroimage*, 26:356–373, 2005.

[Mattout et al. 2006] J. Mattout, C. Phillips, W.D. Penny, M. Rugg, and K.J. Friston. MEG source localisation under multiple constraints: an extended Bayesian framework. *NeuroImage*, 30(3):753-767, 2006.

[Michel et al. 2004] C.M. Michel, M.M. Marraya, G. Lantza, S. Gonzalez, L. Spinelli, and R. Grave de Peralta. EEG source imaging. *Clinical Neurophysiology*, 115:2195–2222, 2004.

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

[Penny et al. 2006] W.D. Penny, G. Flandin, and N. Trujillo-Barreto. Bayesian Comparison of Spatially Regularised General Linear Models. *Human Brain Mapping*, 2006. Accepted for publication.

[Press et al. 1992] W. H. Press, S.A. Teukolsky, W.T. Vetterling, and B.V.P. Flannery. *Numerical Recipes in C.* Cambridge, 1992.

[Raftery 1995] A.E. Raftery. Bayesian model selection in social research. In P.V. Marsden, editor, *Sociological Methodology*, pages 111–196. Cambridge, Mass., 1995.

[Ripley 1995] B. Ripley. *Pattern Recognition and Neural Networks.* Cambridge University Press, Cambridge, 1995.

[Schwarz 1978] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[Stephan et al. 2005] K.E. Stephan, K.J. Friston, and W.D. Penny. Computing the objective function in DCM. Technical report, Wellcome Department of Imaging Neuroscience, ION, UCL, 2005.

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

Log-evidence		
1 constraint	Q_{sc}	205.2
	Q_{int}	208.4
	Q^v_{ext}	**215.6**
	Q^i_{ext}	131.5
2 constraints	Q_{sc}, Q_{int}	207.4
	Q_{sc}, Q^v_{ext}	214.1
	Q_{sc}, Q^i_{ext}	204.9
	Q_{int}, Q^v_{ext}	214.9
	Q_{int}, Q^i_{ext}	207.4
	Q^v_{ext}, Q^i_{ext}	213.2
3 constraints	$Q_{sc}, Q_{int}, Q^v_{ext}$	211.5
	$Q_{sc}, Q_{int}, Q^i_{ext}$	207.2
	$Q_{sc}, Q^v_{ext}, Q^i_{ext}$	214.7
	$Q_{int}, Q^v_{ext}, Q^i_{ext}$	212.7
4 constraints	$Q_{sc}, Q_{int}, Q^v_{ext}, Q^i_{ext}$	211.3

Table 5. Log-evidence of models with different combinations of smoothness constraints, Q_{sc}, intrinsic constraints, Q_{int}, valid, Q^v_{ext} and invalid, Q^i_{ext}, extrinsic constraints.

Bayes factor				
Model 1	Model 2	Model 3	$\mathcal{B}_{21}$	$\mathcal{B}_{31}$
Q_{sc}	Q_{sc}, Q^v_{ext}	Q_{sc}, Q^i_{ext}	7047	0.8
Q_{int}	Q_{int}, Q^v_{ext}	Q_{int}, Q^i_{ext}	655	0.4
Q_{sc}, Q_{int}	$Q_{sc}, Q_{int}, Q^v_{ext}$	$Q_{sc}, Q_{int}, Q^i_{ext}$	60	0.8

Table 6. Bayes factors for models with and without valid location priors, B_{21}, and with and without invalid location priors, B_{31}. Valid location priors make the models significantly better, wheras invalid location priors do not make them significantly worse.

Sensors	Number of models	Min	Max	Prob True model
EEG-19	15	0.02	0.30	0.11 (3)
EEG-120	2	0.49	0.51	0.49 (2)
MEG-151	1	1	1	1

Table 7. BMA results for the 'Opr +Th' simulation study. The second, third and fourth columns show the number of models, and minimum and maximum probabilities, in Occam's window. In the last column, the number in parenthesis indicates the position of the true model when all models in Occam's window are ranked by probability.

Sensors	Number of models	Min	Max	Prob True model
EEG-19	3	0.30	0.37	0.30 (3)
EEG-120	1	1	1	1
MEG-151	1	1	1	1

Table 8. BMA results for the 'Th' simulation study. The second, third and fourth columns show the number of models, and minimum and maximum probabilities, in Occam's window. In the last column, the number in parenthesis indicates the position of the true model when all models in Occam's window are ranked by probability.

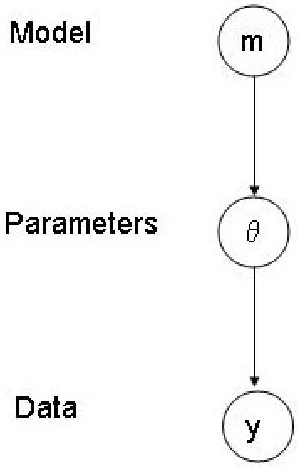

Figure 1. *Hierarchical generative model in which members of a model class, indexed by m, are considered as part of the hierarchy. Typically, m indexes the structure of the model. This might be the connectivity pattern in a dynamic causal model or set of anatomical or functional constraints in a source 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 $p(y|\theta, m)$.*

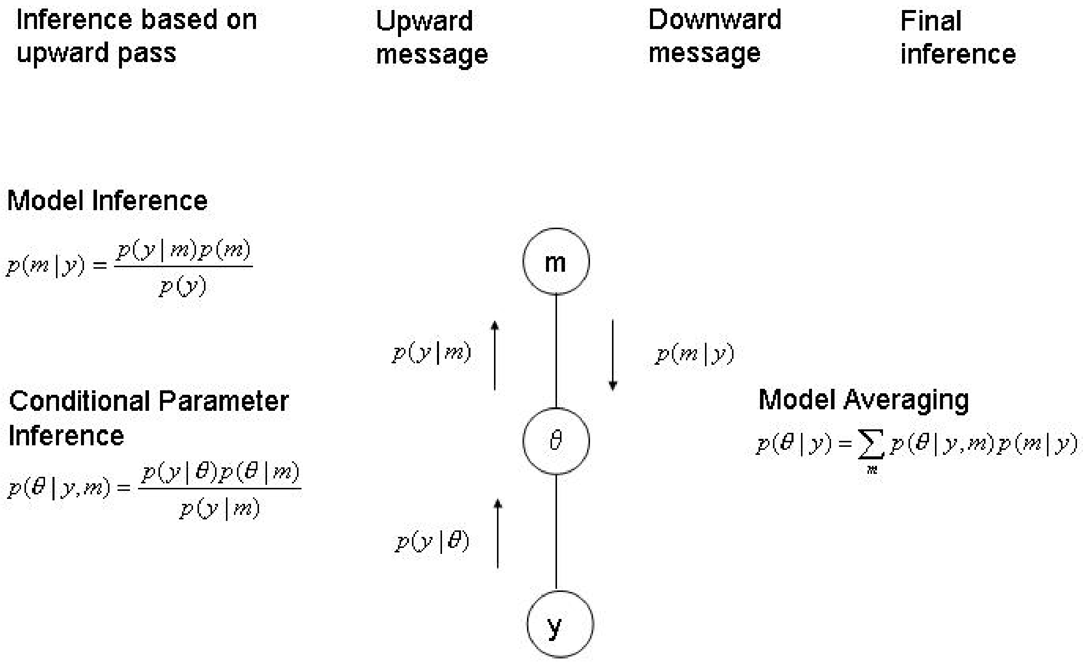
Inference based on upward pass
Upward message
Downward message
Final inference
Model Inference
$p(m \mid y) = \frac{p(y \mid m) p(m)}{p(y)}$
m
$p(y \mid m)$
$p(m \mid y)$
Conditional Parameter Inference
$p(\theta \mid y, m) = \frac{p(y \mid \theta) p(\theta \mid m)}{p(y \mid m)}$
θ
Model Averaging
$p(\theta \mid y) = \sum_m p(\theta \mid y, m) p(m \mid y)$
$p(y \mid \theta)$
y

Figure 2. *Figure 5 in chaper 11 describes the belief propagation alorithm for implementing Bayesian inference in hierarchical models. This figure shows a special case of belief propagation for Bayesian Model Selection (BMS) and Bayesian Model Averaging (BMA). In BMS, the posterior model probability $p(m|y)$, is used to select a single 'best' model. In 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 $p(\theta|y)$.*

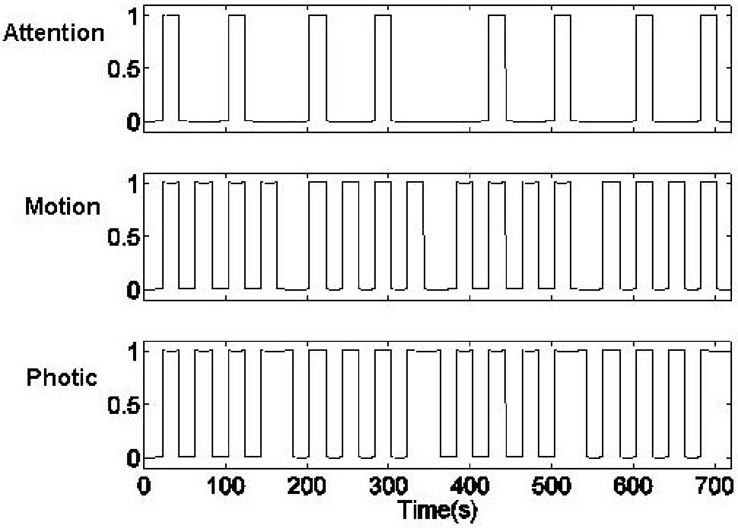

Figure 3. The 'Photic', 'Motion' and 'Attention' variables used in the DCM analysis of the Attention to Visual Motion data (see Figures 4 and 5).

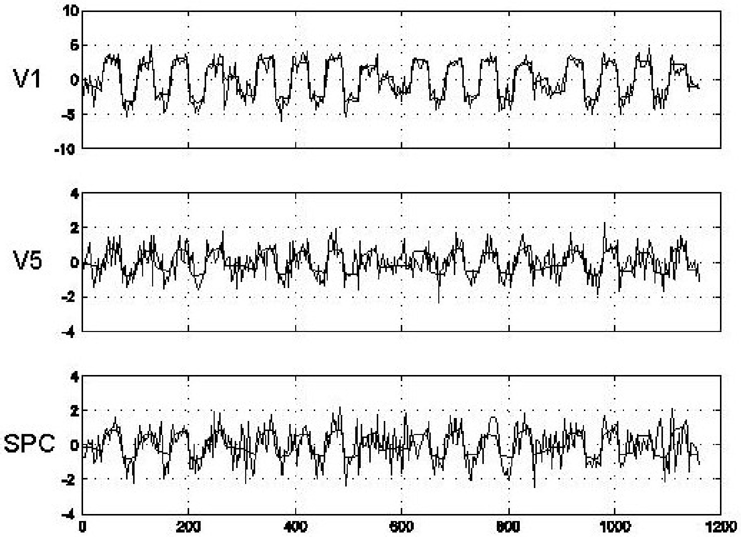

Figure 4. **Attention data.** fMRI time series (rough solid lines) from regions V1, V5 and SPC and the corresponding estimates from DCM model 1 (smooth solid lines).

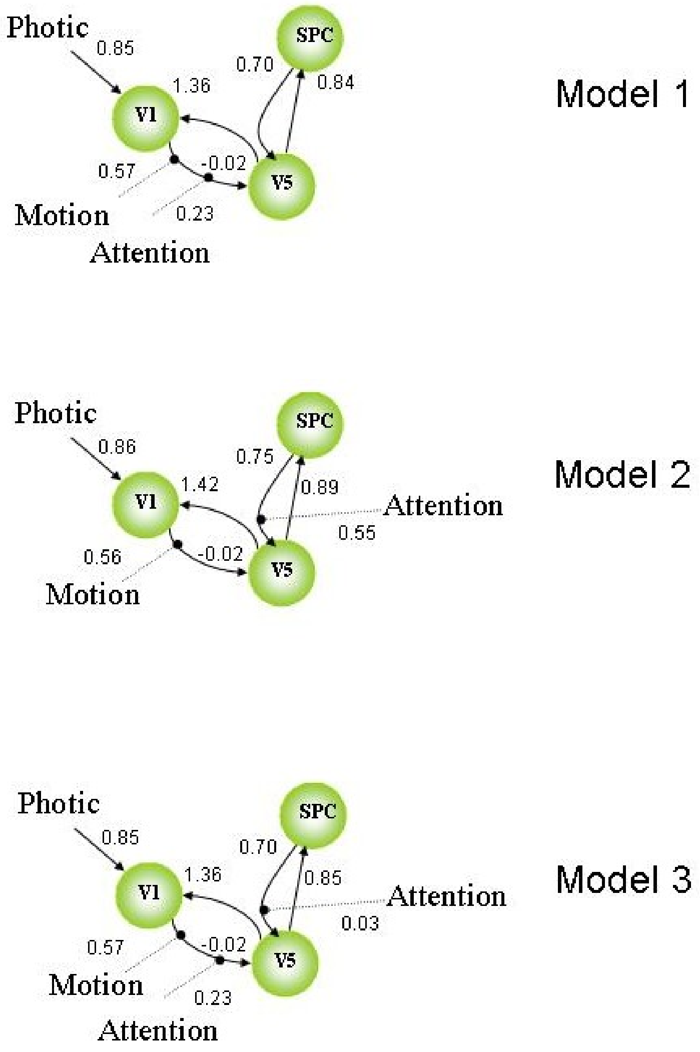

Figure 5. Attention models. In all models photic stimulation enters V1 and the motion variable modulates the connection from V1 to V5. Models 1, 2 and 3 have reciprocal and hierarchically organised intrinsic connectitivty. They differ in how attention modulates the connectivity to V5, with model 1 assuming modulation of the forward connection, model 2 assuming modulation of the backward connection and model 3 assuming both. Solid arrows indicate input and intrinsic connections and dotted lines indicate modulatory connections.

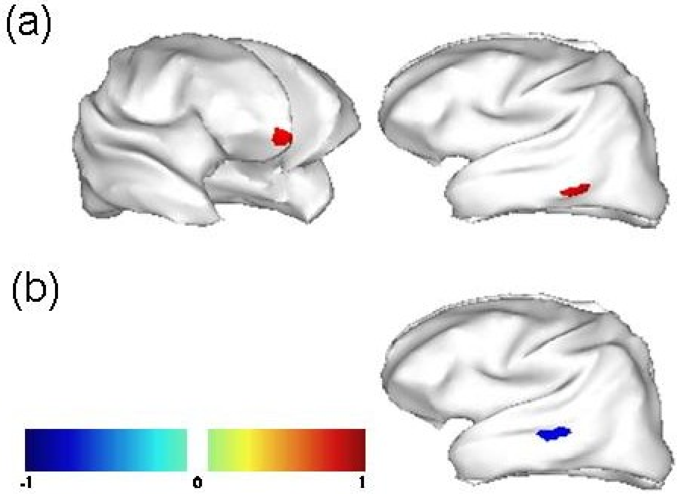

Figure 6. *Inflated cortical representation of (a) two simulated source locations ('valid' prior) and (b) 'invalid' prior location.*

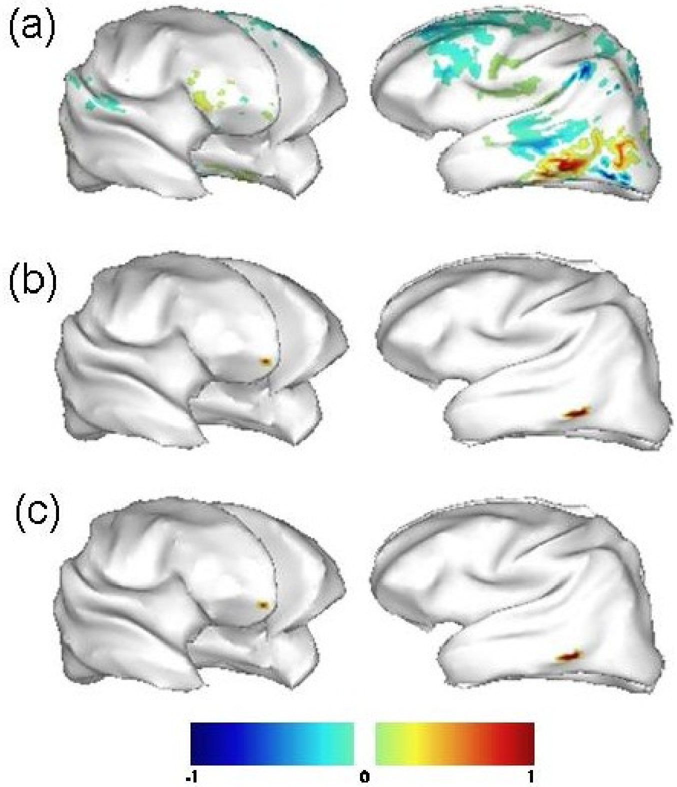

Figure 7. *Inflated cortical representation of representative source reconstructions using (a) smoothness prior, (b) smoothness and valid priors and (c) smoothness, valid and invalid priors. The reconstructed values have been normalised between -1 and 1.*

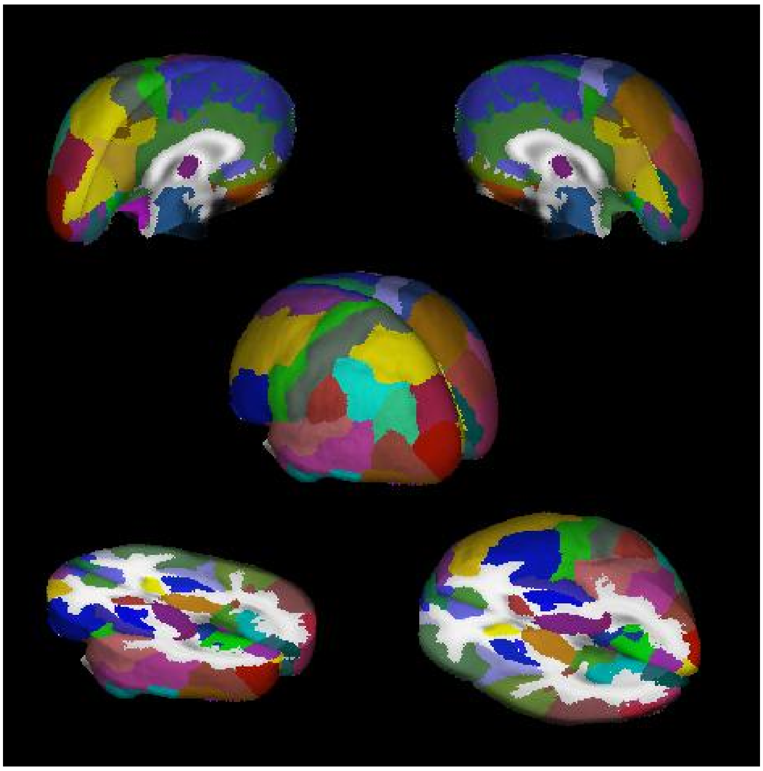

Figure 8. 3D segmentation of 71 structures of the Probabilistic MRI Atlas developed at the Montreal Neurological Institute. As shown in the color scale, brain areas belonging to different hemispheres were segmented separately.

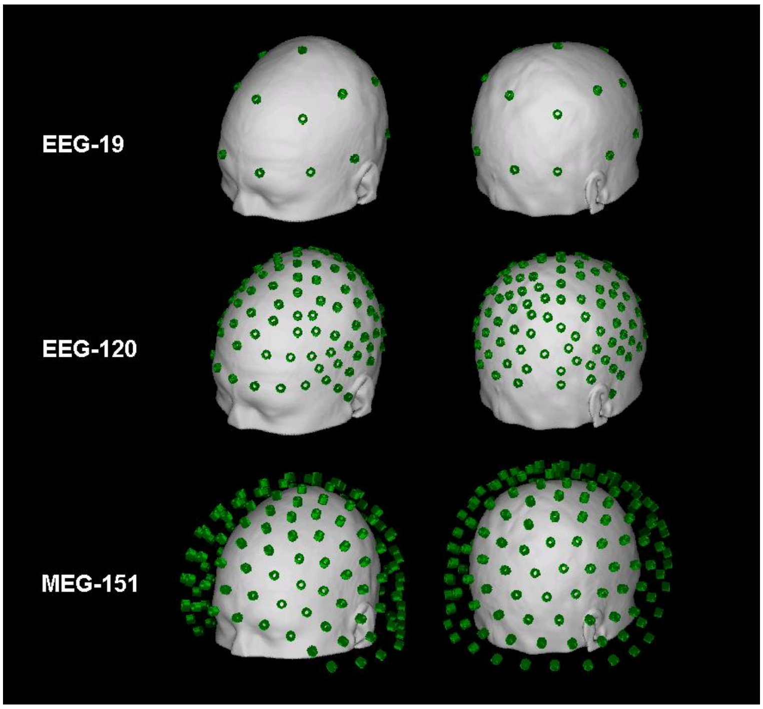

Figure 9. Different arrays of sensors used in the simulations. EEG-19 represents the 10/20 electrode system; EEG-120 is obtained by extending and refining the 10/20 system; and MEG-151 corresponds to the spatial configuration of MEG sensors in the helmet of the CTF System Inc.

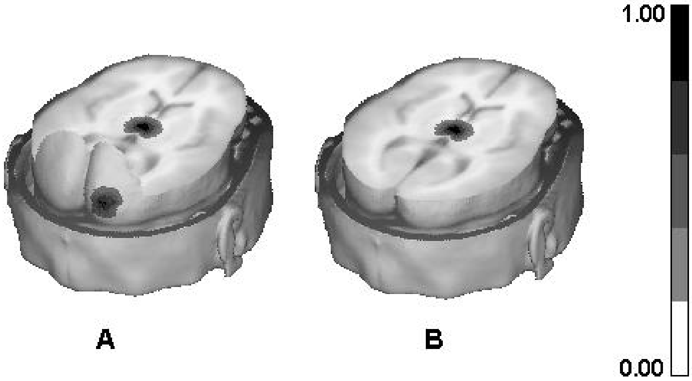

Figure 10. Spatial distributions of the simulated primary current densities. A) Simultaneous activation of two sources at different depths: one in the right Occipital Pole and the other in the Thalamus (OPR+TH). B) Simulation of a single source in the Thalamus (TH).

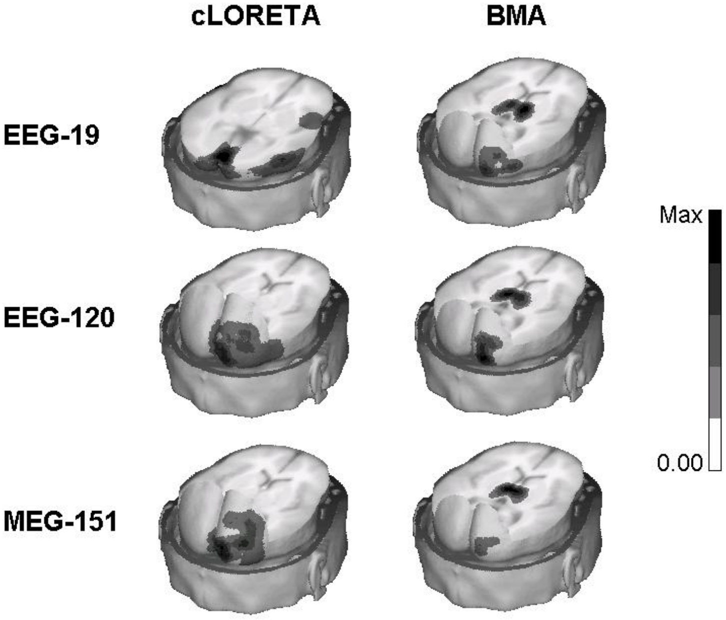

Figure 11. 3D reconstructions of the absolute values of BMA and cLORETA solutions for the OPR+TH source case. The first column indicates the array of sensors used in each simulated data set. The maximum of the scale is different for each case. For cLORETA (from top to bottom): Max = 0.21, 0.15 and 0.05; for BMA (from top to bottom): Max = 0.41, 0.42 and 0.27.

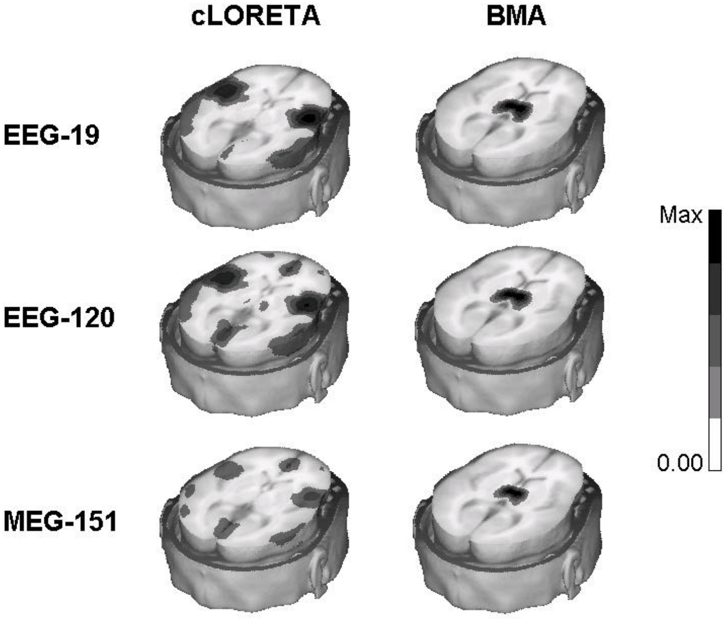

Figure 12. 3D reconstructions of the absolute values of BMA and cLORETA solutions for the TH source case. The first column indicates the array of sensors used in each simulated data set. The maximum of the scale is different for each case. For cLORETA (from top to bottom): Max = 0.06, 0.01 and 0.003 ; for BMA (from top to bottom): Max = 0.36, 0.37 and 0.33.