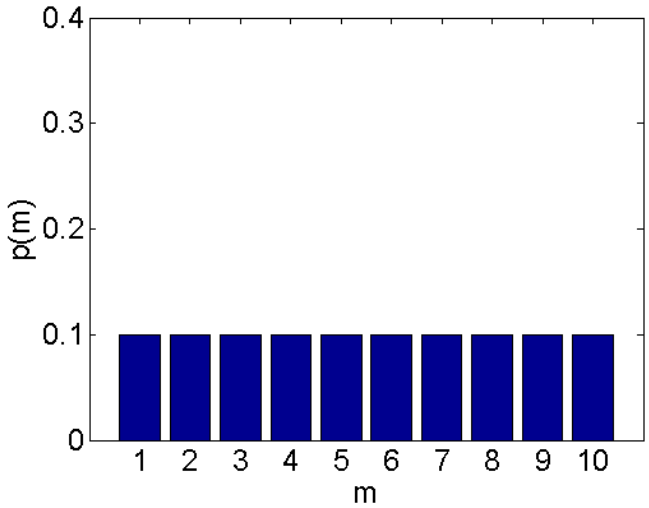
Bayesian Model Comparison

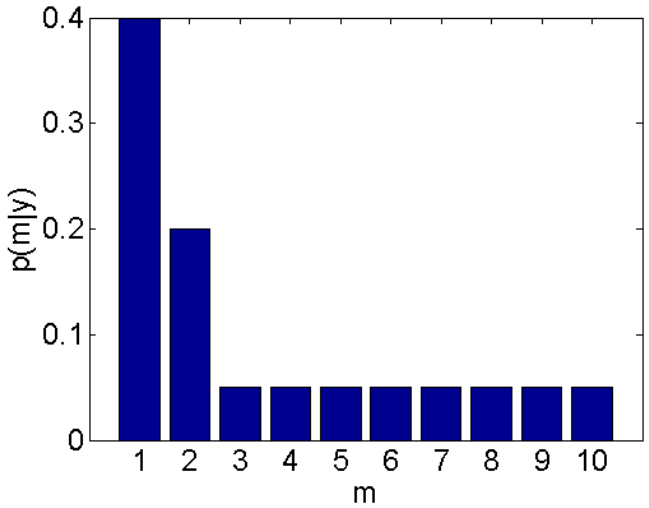
Will Penny

June 2nd 2011

Bayes rule for models

A prior distribution over model space $p(m)$ (or 'hypothesis space') can be updated to a posterior distribution after observing data y.





This is implemented using Bayes rule

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

where $p(y|m)$ is referred to as the evidence for model m and the denominator is given by

$$p(y) = \sum_{m'} p(y|m')p(m')$$

Model Evidence

The evidence is the denominator from the first (parameter) level of Bayesian inference

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

The model evidence is not straightforward to compute, since this computation involves integrating out the dependence on model parameters

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

Posterior Model Probability

Given equal priors, $p(m=i) = p(m=j)$ the posterior model probability is

$$\begin{aligned} p(m=i|y) &= \frac{p(y|m=i)}{p(y|m=i)+p(y|m=j)} \\ &= \frac{1}{1+\frac{p(y|m=j)}{p(y|m=i)}} \end{aligned}$$

Hence

$$p(m=i|y) = \sigma(\log B_{ij})$$

where

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

is the Bayes factor for model 1 versus model 2 and

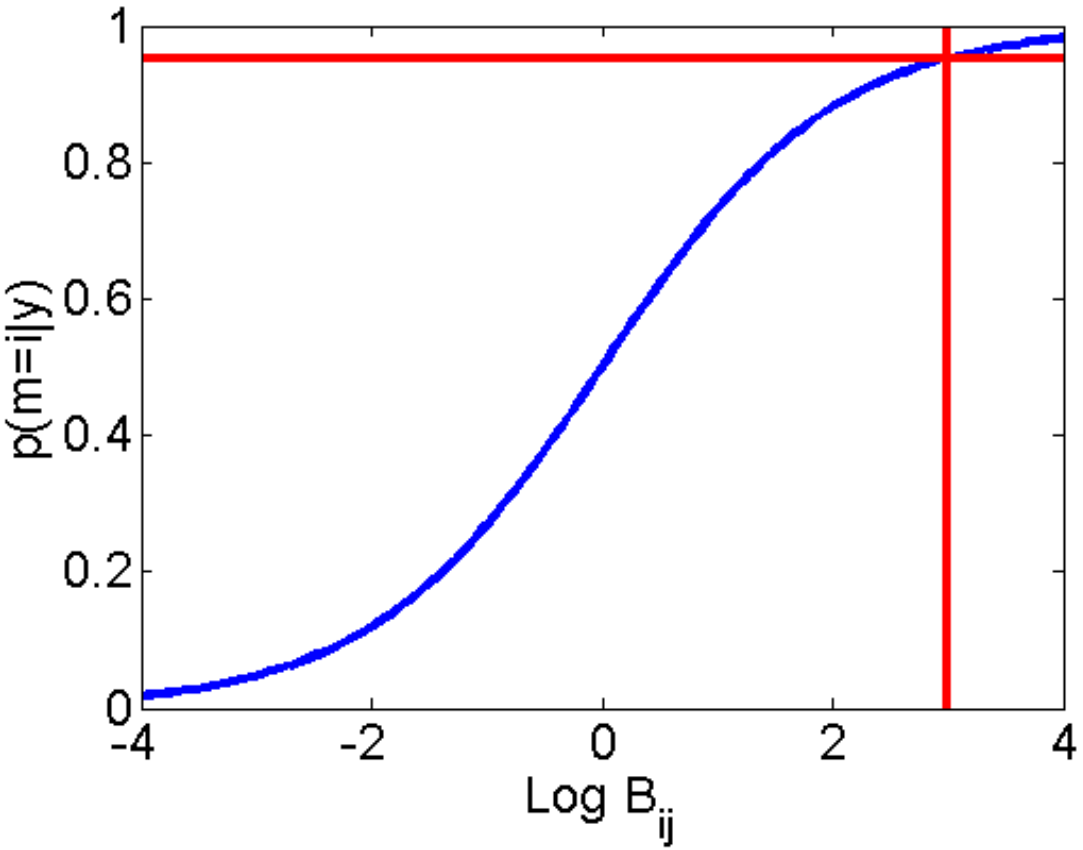
$$\sigma(x) = \frac{1}{1+\exp(-x)}$$

is the sigmoid function.

Bayes factors

The posterior model probability is a sigmoidal function of the log Bayes factor

$$p(m = i|y) = \sigma(\log B_{ij})$$



Bayes factors

The posterior model probability is a sigmoidal function of the log Bayes factor

$$p(m = i|y) = \sigma(\log B_{ij})$$

Table 1
Interpretation of Bayes factors

B_{ij}	$p(m = i\|y)$ (%)	Evidence in favor of model i
1–3	50–75	Weak
3–20	75–95	Positive
20–150	95–99	Strong
≥150	≥99	Very strong

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 favor of i.

From Raftery (1995).

Odds Ratios

If we don't have uniform priors one can work with odds ratios.

The prior and posterior odds ratios are defined as

$$\pi_{ij}^0 = \frac{p(m=i)}{p(m=j)}$$
$$\pi_{ij} = \frac{p(m=i|y)}{p(m=j|y)}$$

resepectively, and are related by the Bayes Factor

$$\pi_{ij} = B_{ij} \times \pi_{ij}^0$$

eg. priors odds of 2 and Bayes factor of 10 leads posterior odds of 20.

An odds ratio of 20 is 20-1 ON in bookmakers parlance.

Likelihood

We consider the same frameworks as in lecture 4, ie Bayesian estimation of nonlinear models of the form

$$y = g(w) + e$$

where $g(w)$ is some nonlinear function, and e is zero mean additive Gaussian noise with covariance C_y. The likelihood of the data is therefore

$$p(y|w, \lambda) = \mathrm{N}(y; g(w), C_y)$$

The error covariances are assumed to decompose into terms of the form

$$C_y^{-1} = \sum_i \exp(\lambda_i) Q_i$$

where Q_i are known precision basis functions and λ are hyperparameters.

Priors

Typically $Q = I_{N_y}$ and the observation noise precision $s = \exp(\lambda)$ where λ is a latent variable, also known as a hyperparameter.

The hyperparameters are constrained by the prior

$$p(\lambda) = \mathrm{N}(\lambda; \mu_\lambda, C_\lambda)$$

We allow Gaussian priors over model parameters

$$p(w) = \mathrm{N}(w; \mu_w, C_w)$$

where the prior mean and covariance are assumed known.

This is not Empirical Bayes.

Joint Log Likelihood

Writing all unknown random variables as $\theta = \{w, \lambda\}$, the above distributions allow one to write down an expression for the joint log likelihood of the data, parameters and hyperparameters

$$\log p(y, \theta) = \log[p(y|w, \lambda)p(w)p(\lambda)]$$

Here it splits into three terms

$$\begin{aligned} \log p(y, \theta) &= \log p(y|w, \lambda) \\ &+ \log p(w) \\ &+ \log p(\lambda) \end{aligned}$$

Joint Log Likelihood

The joint log likelihood is composed of sum squared precision weighted prediction errors and entropy terms

$$\begin{aligned} L &= -\frac{1}{2} e_y^T C_y^{-1} e_y - \frac{1}{2} \log |C_y| - \frac{N_y}{2} \log 2\pi \\ &- \frac{1}{2} e_w^T C_w^{-1} e_w - \frac{1}{2} \log |C_w| - \frac{N_w}{2} \log 2\pi \\ &- \frac{1}{2} e_\lambda^T C_\lambda^{-1} e_\lambda - \frac{1}{2} \log |C_\lambda| - \frac{N_\lambda}{2} \log 2\pi \end{aligned}$$

where N_y, N_w and N_λ are the numbers of data points, parameters and hyperparameters. The prediction errors are the difference between what is expected and what is observed

$$\begin{aligned} e_y &= y - g(m_w) \\ e_w &= m_w - \mu_w \\ e_\lambda &= m_\lambda - \mu_\lambda \end{aligned}$$

Variational Laplace

The Variational Laplace (VL) algorithm assumes an approximate posterior density of the following factorised form

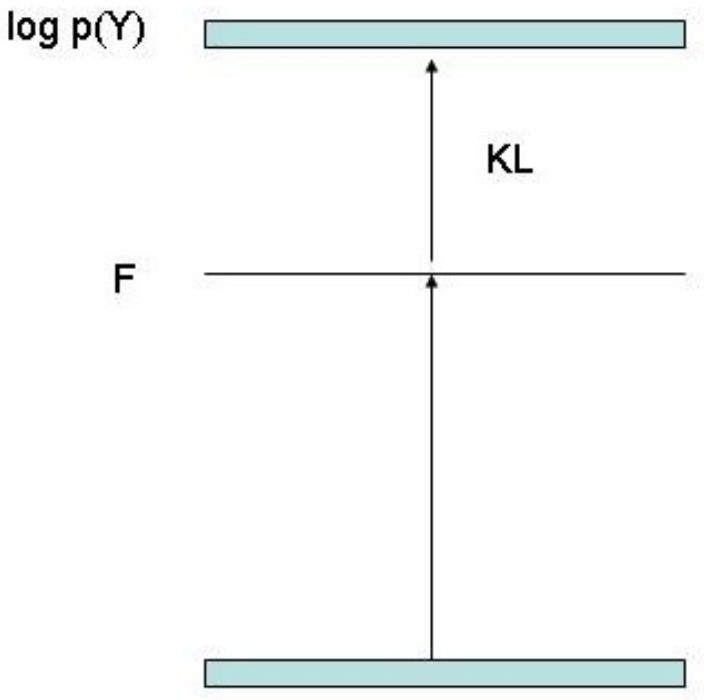
$$\begin{aligned} q(\theta|m) &= q(w|m)q(\lambda|m) \\ q(w|m) &= \mathrm{N}(w; m_w, S_w) \\ q(\lambda|m) &= \mathrm{N}(\lambda; m_\lambda, S_\lambda) \end{aligned}$$

See lecture 4 and Friston et al. (2007).

Free Energy

In the lecture on approximate inference we showed that

$$\log p(y|m) = F(m) + KL[q(\theta|m)||p(\theta|y,m)]$$



Because KL is always positive F provides a lower bound on the model evidence.

F is known as the negative variational free energy. Henceforth 'free energy'.

Free Energy

In lecture 4 we showed that

$$F = \int q(\theta) \log \frac{p(Y,\theta)}{q(\theta)} d\theta$$

We can write is as two terms

$$\begin{aligned} F &= \int q(\theta) \log p(Y|\theta) d\theta + \int q(\theta) \log \frac{p(\theta)}{q(\theta)} d\theta \\ &= \int q(\theta) \log p(Y|\theta) d\theta + KL[q(\theta)||p(\theta)] \end{aligned}$$

where this KL is between the approximate posterior and the prior - not to be confused with the earlier. The first term is referred to as the averaged likelihood.

Free Energy

For our nonlinear model with Gaussian priors and approximate Gaussian posteriors F is composed of sum squared precision weighted prediction errors and Occam factors

$$\begin{aligned} F &= -\frac{1}{2} e_y^T C_y^{-1} e_y - \frac{1}{2} \log |C_y| - \frac{N_y}{2} \log 2\pi \\ &- \frac{1}{2} e_w^T C_w^{-1} e_w - \frac{1}{2} \log \frac{|C_w|}{|S_w|} \\ &- \frac{1}{2} e_\lambda^T C_\lambda^{-1} e_\lambda - \frac{1}{2} \log \frac{|C_\lambda|}{|S_\lambda|} \end{aligned}$$

where prediction errors are the difference between what is expected and what is observed

$$\begin{aligned} e_y &= y - g(m_w) \\ e_w &= m_w - \mu_w \\ e_\lambda &= m_\lambda - \mu_\lambda \end{aligned}$$

Accuracy and Complexity

The free energy for model m can be split into an accuracy and a complexity term

$$F(m) = Accuracy(m) - Complexity(m)$$

where

$$Accuracy(m) = -\frac{1}{2} e_y^T C_y^{-1} e_y - \frac{1}{2} \log |C_y| - \frac{N_y}{2} \log 2\pi$$

and

$$\begin{aligned} Complexity(m) &= \frac{1}{2} e_w^T C_w^{-1} e_w + \frac{1}{2} \log \frac{|C_w|}{|S_w|} \\ &+ \frac{1}{2} e_\lambda^T C_\lambda^{-1} e_\lambda + \frac{1}{2} \log \frac{|C_\lambda|}{|S_\lambda|} \end{aligned}$$

Complexity

Model complexity will tend to increase with the number of parameters N_w because distances tend to be larger in higher dimensional spaces.

For the parameters we have

$$Complexity(m) = \frac{1}{2} e_w^T C_w^{-1} e_w + \frac{1}{2} \log \frac{|C_w|}{|S_w|}$$

where

$$e_w = m_w - \mu_w$$

But this will only be the case if these extra parameters diverge from their prior values and have smaller posterior (co)variance than prior (co)variance.

Complexity

In the limit that the posterior equals the prior ($e_w = 0, C_w = S_w$), the complexity term equals zero.

$$Complexity(m) = \frac{1}{2} e_w^T C_w^{-1} e_w + \frac{1}{2} \log \frac{|C_w|}{|S_w|}$$

Because the determinant of a matrix corresponds to the volume spanned by its eigenvectors, the last term gets larger and the model evidence smaller as the posterior volume, $|S_w|$, reduces in proportion to the prior volume, $|C_w|$.

Models for which parameters have to specified precisely (small posterior volume) are brittle. They are not good models (complexity is high).

Correlated Parameters

Other factors being equal, models with strong correlation in the posterior are not good models.

For example, given a model with just two parameters the determinant of the posterior covariance is given by

$$|S_w| = (1 - r^2)\sigma_{w_1}^2 \sigma_{w_2}^2$$

where r is the posterior correlation, σ_{w_1} and σ_{w_2} are the posterior standard deviations of the two parameters.

For the case of two parameters having a similar effect on model predictions the posterior correlation will be high, therefore implying a large complexity penalty.

Decompositions

It is also instructive to decompose approximations to the model evidence into contributions from specific sets of parameters or predictions. In the context of DCM, one can decompose the accuracy terms into contributions from different brain regions (Penny et al 2004).

Table 12
Visual object category data

Source	Model 1 vs. model 2 relative cost (bits)	Bayes factor B_{12}
V3 error	−10.59	1545
MO error	−6.01	64.6
SPC error	3.65	0.08

If the relative cost is E then the contribution to the Bayes factor is 2^{-E}. This enables insight to be gained into why one model is better than another.

Decompositions

Similarly, it is possible to decompose the complexity term into contributions from different sets of parameters. If we ignore correlation among different parameter sets then the complexity is approximately

$$Complexity(m) \approx \frac{1}{2} \sum_j \left(e_{w_j}^T C_{w_j}^{-1} e_{w_j} + \log \frac{|C_{w_j}|}{|S_{w_j}|} \right)$$

where j indexes the jth parameter set. In the context of DCM these could index input connections ($j = 1$), intrinsic connections ($j = 2$), modulatory connections ($j = 3$) etc.

Bayesian Information Criterion

A simple approximation to the log model evidence is given by the Bayesian Information Criterion (Schwarz, 1978)

$$BIC = \log p(y|\hat{w}, m) - \frac{N_w}{2} \log N_y$$

where $\hat{w}$ are the estimated parameters, N_w is the number of parameters, and N_y is the number of data points.

BIC is a special case of the Free Energy approximation that drops all terms that do not scale with the number of data points (see Penny et al, 2004).

There is a complexity penalty of $\frac{1}{2} \log N_y$ for each parameter.

An Information Criterion

An alternative approximation is Akaike's Information Criterion or 'An Information Criterion' (AIC) - Akaike (1973)

$$AIC = \log p(y|\hat{w}, m) - N_w$$

There is a complexity penalty of 1 for each parameter.

AIC and BIC are attractive because they are so easy to implement.

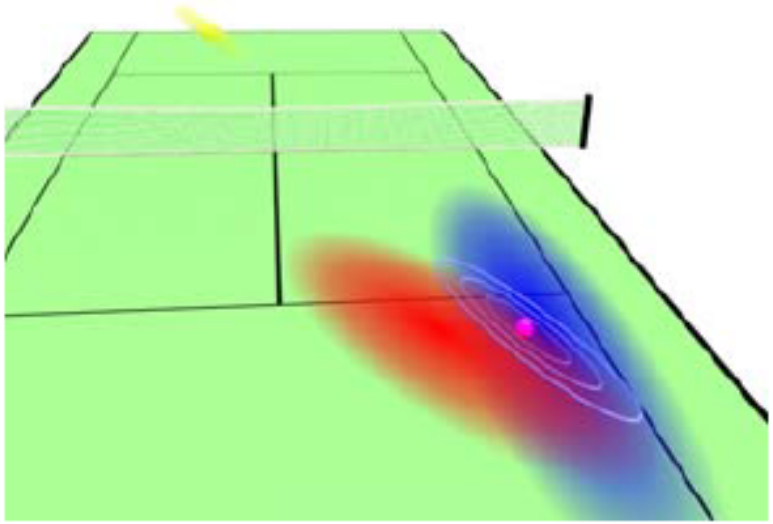
Linear Models

For Linear Models

$$y = Xw + e$$

where X is a design matrix and w are now regression coefficients. The posterior distribution is analytic and given by

$$\begin{aligned} S_w^{-1} &= X^T C_y^{-1} X + C_w^{-1} \\ m_w &= S_w \left(X^T C_y^{-1} y + C_w^{-1} \mu_w \right) \end{aligned}$$



Model Evidence

If we assume the error precision is known then for linear models the free energy is exactly equal to the log model evidence.

We have sum squared precision weighted prediction errors and Occam factors as before

$$\begin{aligned} L &= -\frac{1}{2} e_y^T C_y^{-1} e_y - \frac{1}{2} \log |C_y| - \frac{N_y}{2} \log 2\pi \\ &+ -\frac{1}{2} e_w^T C_w^{-1} e_w - \frac{1}{2} \log \frac{|C_w|}{|S_w|} \end{aligned}$$

where prediction errors are the difference between what is expected and what is observed

$$\begin{aligned} e_y &= y - g(m_w) \\ e_w &= m_w - \mu_w \end{aligned}$$

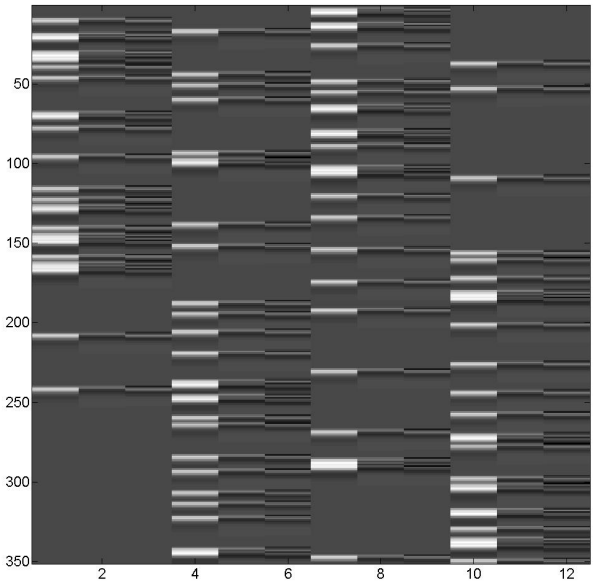
See Bishop (2006) for derivation.

fMRI example

We use a linear model

$$y = Xw + e$$

with design matrix from Henson et al (2002).



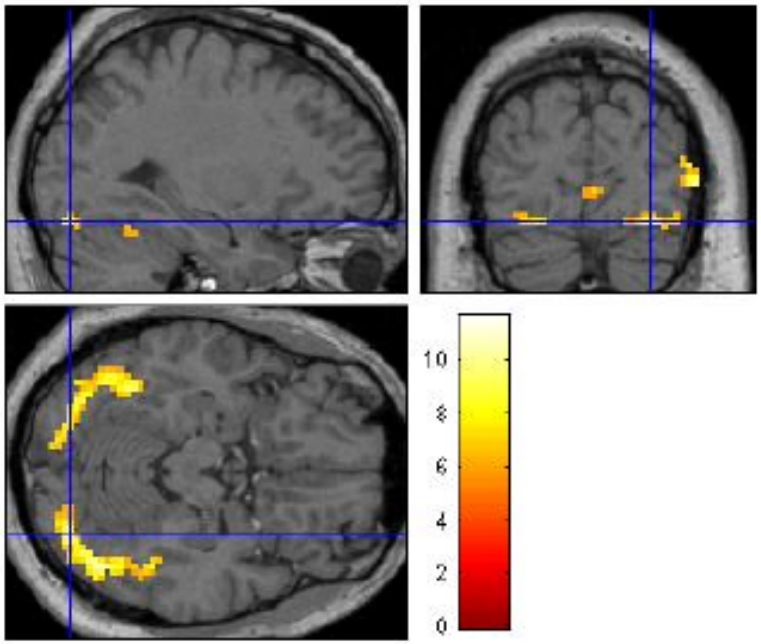
See also SPM Manual. We considered ‘complex’ models (with 12 regressors) and ‘simple’ models (with last 9 only).

fMRI example

Likelihood

$$p(y|w) = \mathrm{N}(y; Xw, C_y)$$

where $C_y = \sigma_e^2 I_{N_y}$.



Parameters were drawn from the prior

$$p(w) = \mathrm{N}(w; \mu_w, C_w)$$

with $\mu_w = 0$ and $C_w = \sigma_p^2 I_p$ with set σ_p to correspond to the magnitude of coefficients in a face responsive area.

fMRI example

The parameter σ_e (observation noise SD) was set to give a range of SNRs where

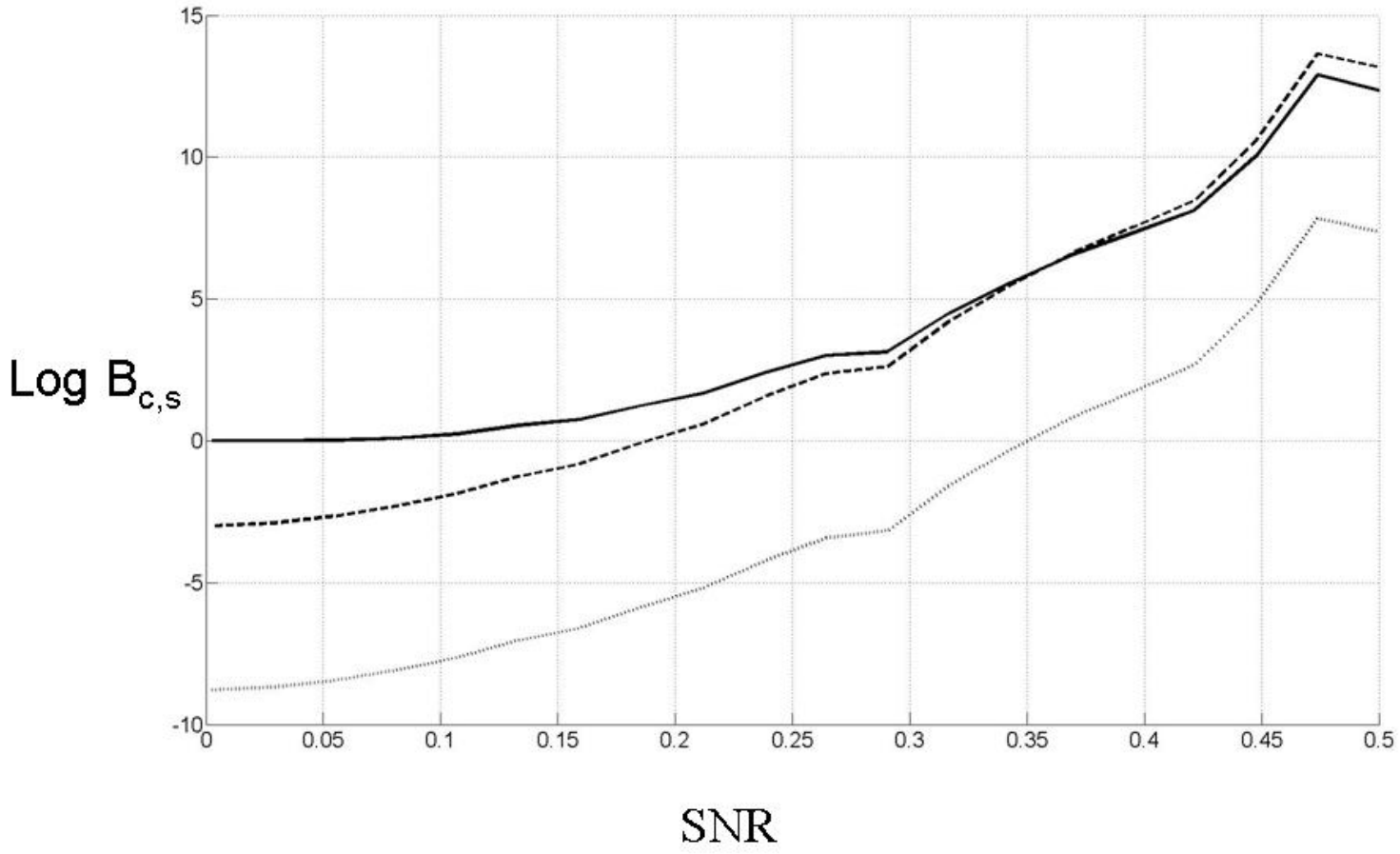
$$SNR = \frac{std(g)}{\sigma_e}$$

and $g = Xw$ is the signal. For each SNR we generated 100 data sets.

We first look at model comparison behaviours when the true model is complex. For each generated data set we fitted both simple and complex models and computed the log Bayes factor. We then averaged this over runs.

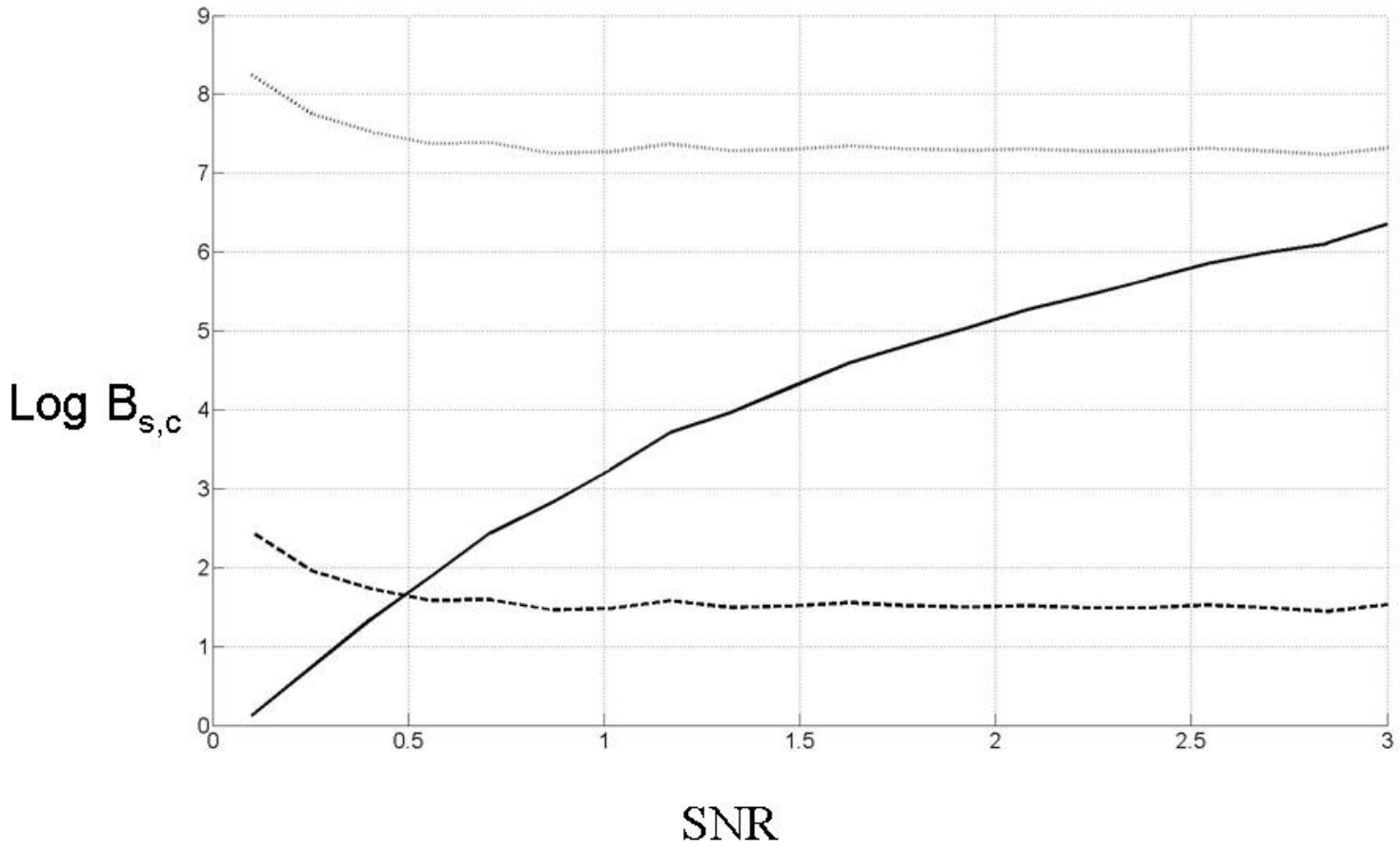
True Model: Complex GLM

Log Bayes factor of complex versus simple model versus the signal to noise ratio, SNR, when true model is the complex GLM for F (solid), AIC (dashed) and BIC (dotted).



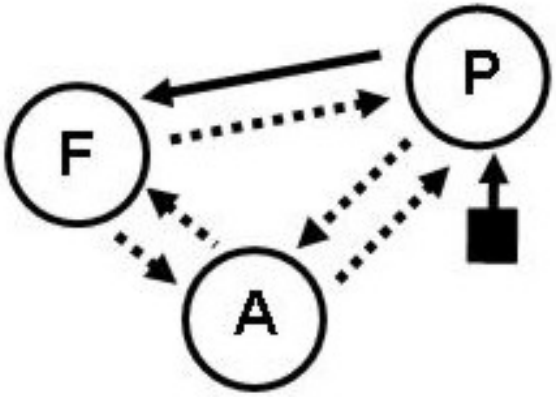
True Model: Simple GLM

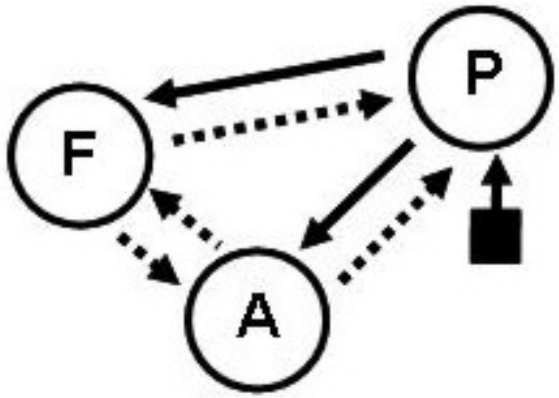
Log Bayes factor of simple versus complex model versus the signal to noise ratio, SNR, when true model is the simple GLM for F (solid), AIC (dashed) and BIC (dotted).



DCM for fMRI

We consider a 'simple' and 'complex' DCM for fMRI (Friston et al 2003)





Neurodynamics evolve according to

$$\begin{bmatrix} \dot{z}_P \\ \dot{z}_F \\ \dot{z}_A \end{bmatrix} = \left(\begin{bmatrix} a_{PP} & a_{PF} & a_{PA} \\ a_{FP} & a_{FF} & a_{FA} \\ a_{AP} & a_{AF} & a_{AA} \end{bmatrix} + u_{int} \begin{bmatrix} 0 & 0 & 0 \\ b_{FP} & 0 & 0 \\ b_{AP} & 0 & 0 \end{bmatrix} \right) \begin{bmatrix} z_P \\ z_F \\ z_A \end{bmatrix} + u_{aud} \begin{bmatrix} c_P \\ 0 \\ 0 \end{bmatrix}$$

where u_{aud} is a train of auditory input spikes and u_{int} indicates whether the input is intelligible (Leff et al 2008). For the simple DCM we have $b_{AP} = 0$.

Priors

Neurodynamic priors are

$$
\begin{aligned}
p(a_{ii}) &= \mathrm{N}(a_{ii}; -1, \sigma^2_{self}) \\
p(a_{ij}) &= \mathrm{N}(a_{ij}; 1/64, \sigma^2_{cross}) \\
p(b) &= \mathrm{N}(b; 0, \sigma^2_b) \\
p(c) &= \mathrm{N}(c; 0, \sigma^2_c)
\end{aligned}
$$

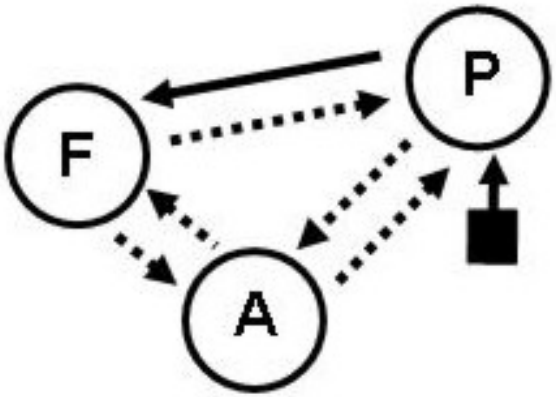
where

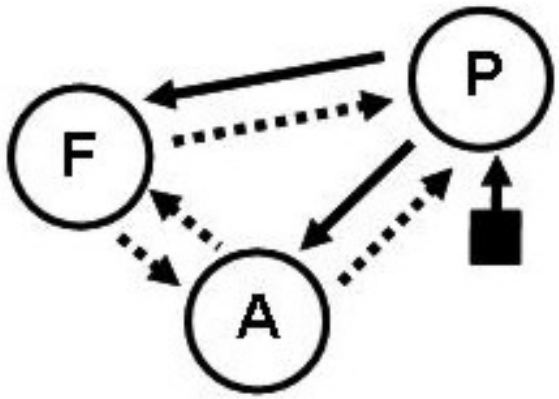
$$
\begin{aligned}
\sigma_{self} &= 0.25 \\
\sigma_{cross} &= 0.50 \\
\sigma_b &= 2 \\
\sigma_c &= 2
\end{aligned}
$$

Simulations

To best reflect the empirical situation, data were generated using *a* and *c* parameters as estimated from original fMRI data (Leff et al 2008).

For each simulation the modulatory parameters b_{FP} (and b_{AP} for the complex model) were drawn from the prior. Neurodynamics and hemodynamics were then integrated to generate signals *g*.

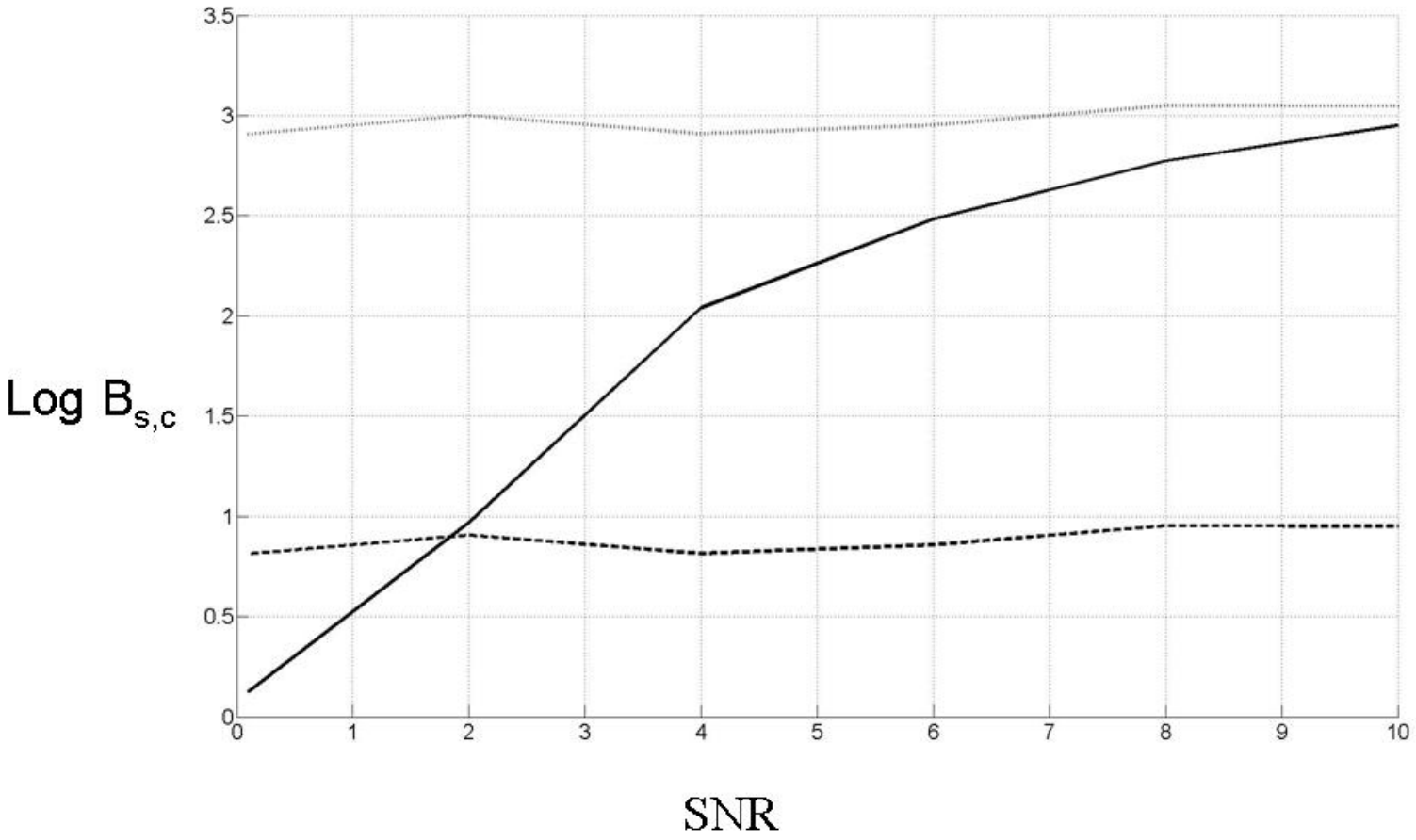




Observation noise SD σ_e was set to a achieve a range of SNRs. Models were fitted using Variational Laplace (lecture 4).

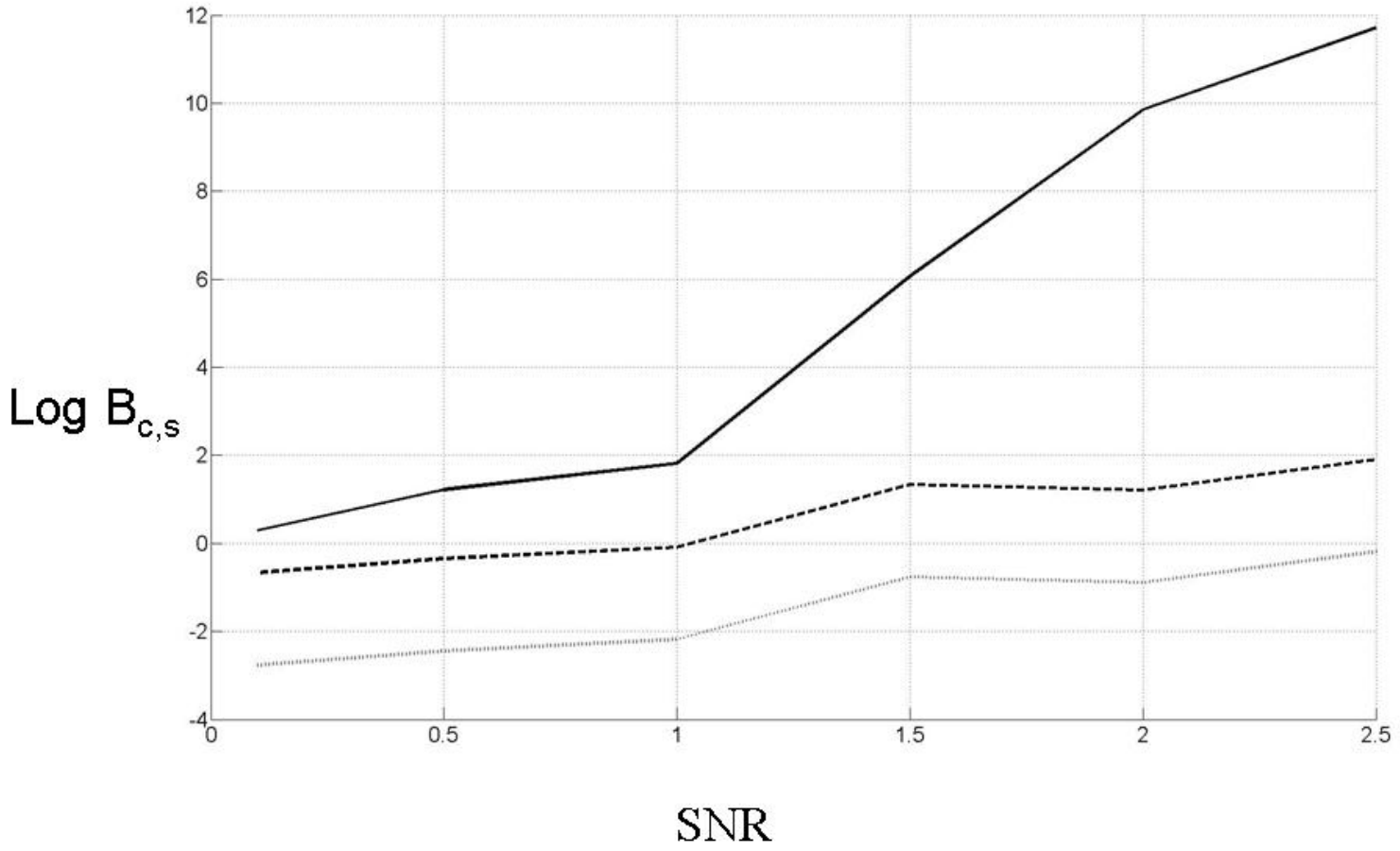
True Model: Simple DCM

Log Bayes factor of simple versus complex model versus the signal to noise ratio, SNR, when true model is the simple DCM for F (solid), AIC (dashed) and BIC (dotted).



True Model: Complex DCM

Log Bayes factor of complex versus simple model versus the signal to noise ratio, SNR, when true model is the complex DCM for F (solid), AIC (dashed) and BIC (dotted).



Decomposition

The complex model was correctly detected by F at high SNR but not by AIC or BIC.

Decomposition of F into accuracy and complexity terms

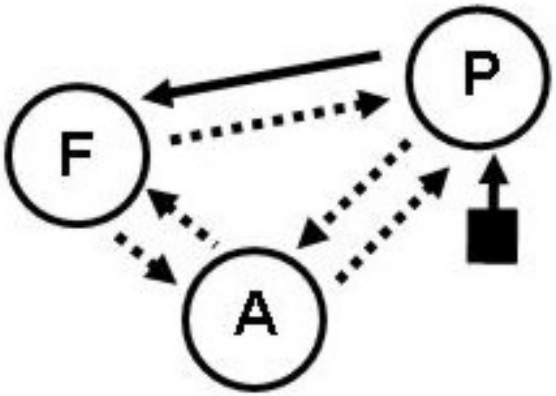
$$F(m) = Accuracy(m) - Complexity(m)$$

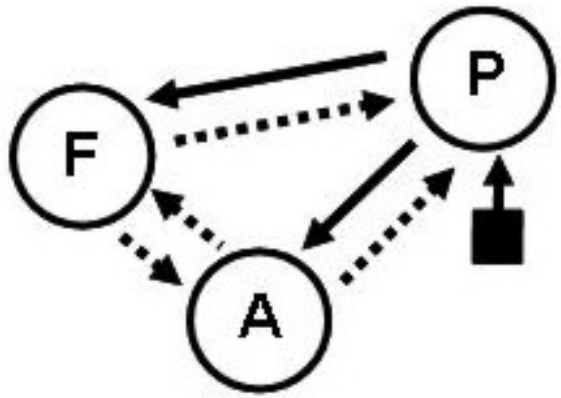
showed that the accuracy of the simple and complex models was very similar. So differences in accuracy did not drive differences in F.

Decomposition

The simple model was able to fit the data from the complex model quite well by having a very strong intrinsic connection from F to A.

Typically, for the simple model $\hat{a}_{AF} = 1.5$. Whereas for the complex model $\hat{a}_{AF} = 0.3$.





This leads to a large complexity penalty (in F) for the simple model

$$Complexity \approx \frac{1}{2} \sum_j e_{w_j}^T C_{w_j}^{-1} e_{w_j} + \ldots$$

Decomposition

Typically, for the simple model $\hat{a}_{AF} = 1.5$. Whereas for the complex model $\hat{a}_{AF} = 0.3$.

$$\begin{aligned} \textit{Complexity} &\approx \frac{1}{2}\sum_j e_{w_j}^T C_{w_j}^{-1} e_{w_j} + \ldots \\ &\approx \frac{1}{2\sigma_{cross}^2}(\hat{a}_{AF} - 1/64)^2 \end{aligned}$$

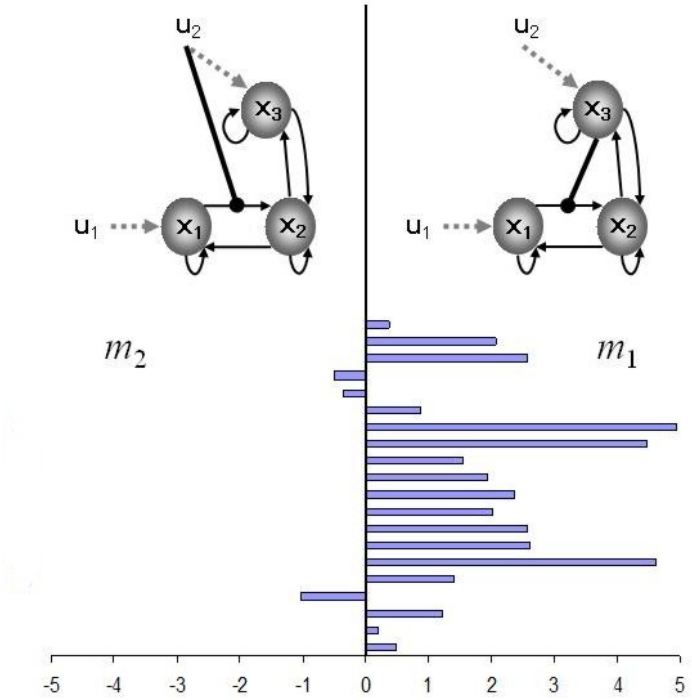
So the simple model pays a bigger complexity penalty.

Hence it is detected by F as the worst model. But BIC and AIC do not detect this as they pay the same penalty for each parameter (regardless of its estimated magnitude).

Fixed Effects

Two models, twenty subjects.

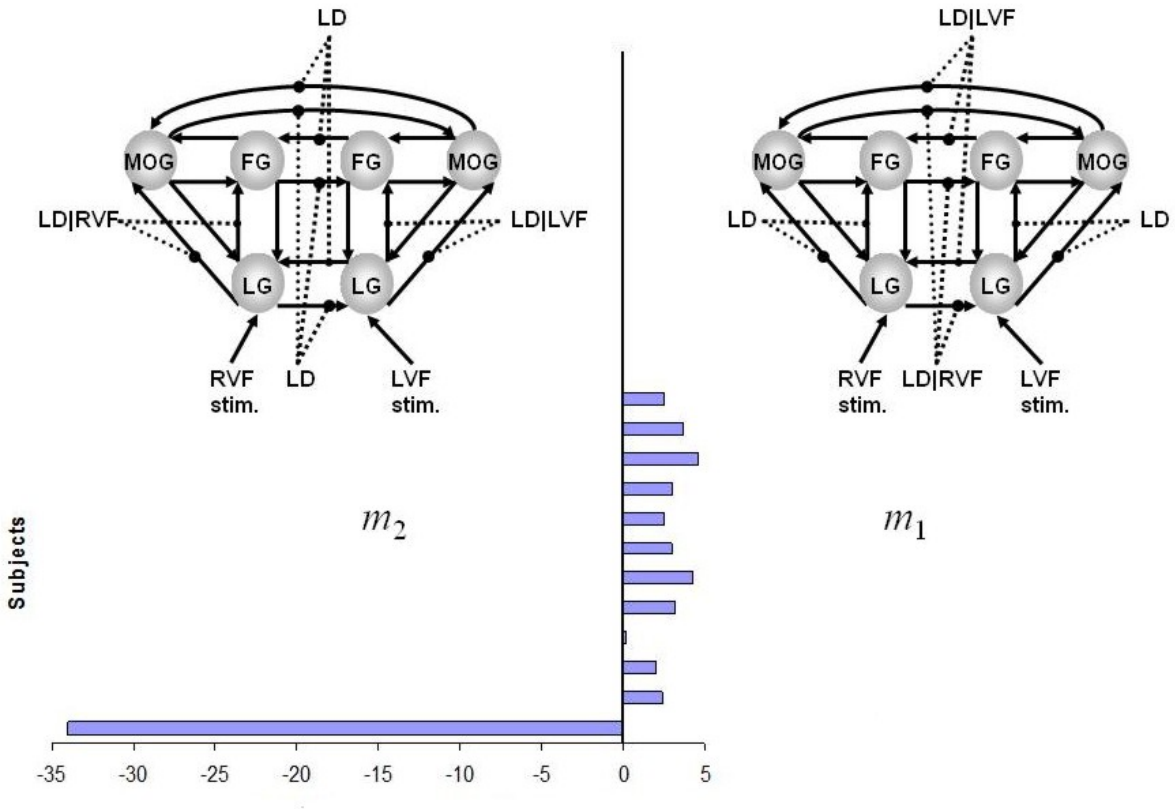
$$\log p(Y|m) = \sum_{n=1}^{N} \log p(y_n|m)$$



The Group Bayes Factor (GBF) is

$$B_{ij} = \prod_{n=1}^{N} B_{ij}(n)$$

Random Effects



11/12=92% subjects favour model 1.

$GBF = 15$ in favour of model 2. FFX inference does not agree with the majority of subjects.

Random Effects

For RFX analysis it is possible that different subjects use different models. If we knew exactly which subjects used which models then this information could be represented in a $[N \times M]$ assignment matrix, A, with entries $a_{nm} = 1$ if subject m used model n, and $a_{nm} = 0$ otherwise.

For example, the following assignment matrix

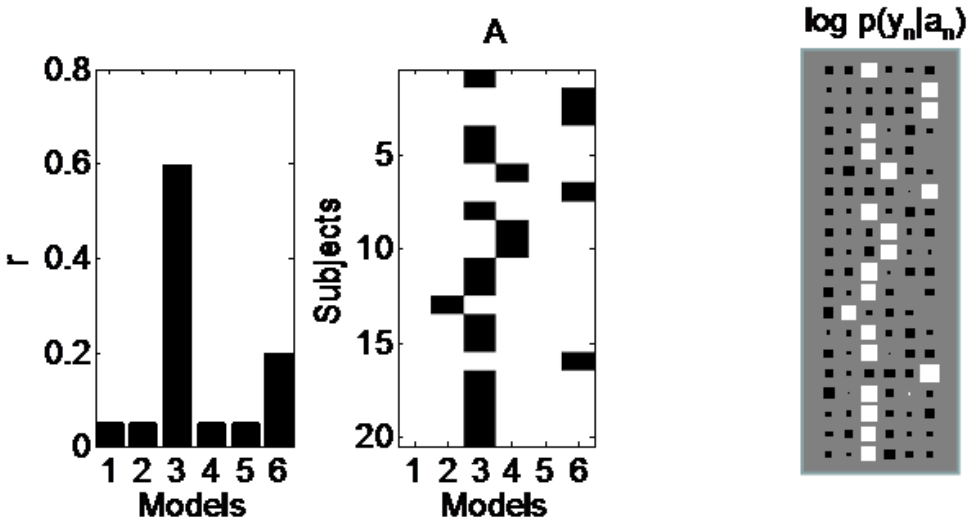
$$A = \begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$$

indicates that subjects 1 and 2 used model 2 and subject 3 used model 1.

We denote r_m as the frequency with which model m is used in the population. We also refer to r_m as the model probability.

Generative Model

In our generative model we have a prior $p(r|\alpha)$. A vector of probabilities is then drawn from this.



An assigment for each subject a_n is then drawn from $p(a_n|r)$. Finally a_n specifies which log evidence value to use for each subject. This specifies $p(y_n|a_n)$.

The joint likelihood for the RFX model is

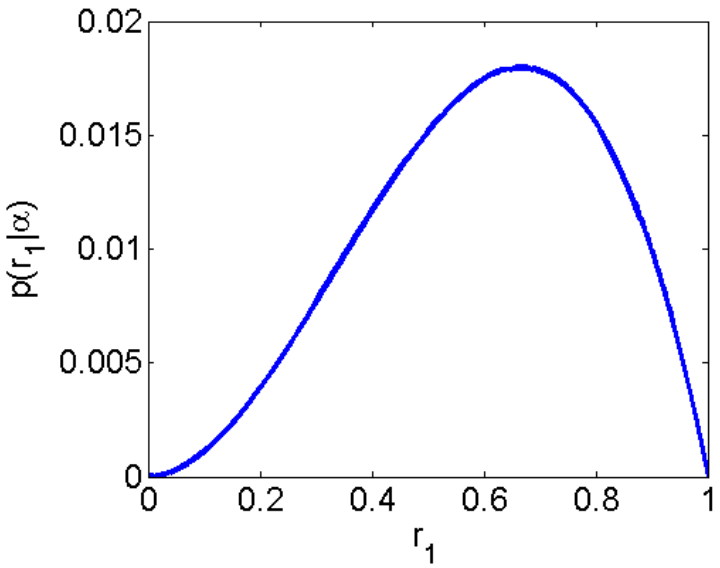
$$p(y, a, r|\alpha) = \prod_{n=1}^{N} [p(y_n|a_n)p(a_n|r)]p(r|\alpha)$$

Prior Model Frequencies

We define a prior distribution over r which is a Dirichlet

$$p(r|\alpha_0) = \text{Dir}(\alpha_0) = \frac{1}{Z}\prod_{m=1}^{M} r_m^{\alpha_0(m)-1}$$

where Z is a normalisation term and the parameters, α_0, are strictly positively valued and the mth entry can be interpreted as the number of times model m has been selected.



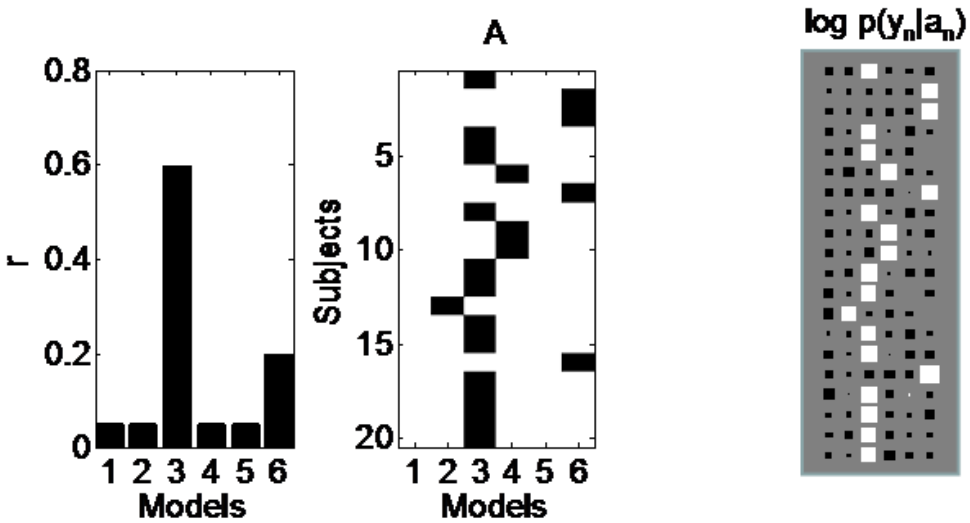
Example with $\alpha_0 = [3, 2]$ and $r = [r_1, 1 - r_1]$.

In the RFX generative model we use a uniform prior $\alpha_0 = [1, 1]$ or more generally $\alpha_0 = ones(1, M)$.

Model Assignment

The probability of the 'assignation vector', a_n, is then given by the multinomial density

$$p(a_n|r) = \text{Mult}(r) = \prod_{m=1}^{M} r_m^{a_{nm}}$$



The assignments then indiciate which entry in the model evidence table to use for each subject, $p(y_n|a_n)$.

Gibbs Sampling

Samples from the posterior densities $p(r|y)$ and $p(a|y)$ can be drawn using Gibbs sampling (Gelman et al 1995).

This can be implemented by alternately sampling from

$$\begin{aligned} r &\sim p(r|a,y) \\ a &\sim p(a|r,y) \end{aligned}$$

and discarding samples before convergence.

This is like a sample-based EM algorithm.

Gibbs Sampling

STEP 1: model probabilites are drawn from the prior distribution

$$r \sim \text{Dir}(\alpha_{prior})$$

where by default we set $\alpha_{prior}(m) = \alpha_0$ for all m (but see later).

STEP 2: For each subject $n = 1..N$ and model $m = 1..M$ we use the model evidences from model inversion to compute

$$\begin{aligned} u_{nm} &= \exp\left(\log p(y_n|m) + \log r_m\right) \\ g_{nm} &= \frac{u_{nm}}{\sum_{m=1}^{M} u_{nm}} \end{aligned}$$

Here, g_{nm} is our posterior belief that model m generated the data from subject n.

Gibbs Sampling

STEP 3: For each subject, model assignation vectors are then drawn from the multinomial distribution

$$a_n \sim \text{Mult}(g_n)$$

We then compute new model counts

$$\begin{aligned} \beta_m &= \sum_{n=1}^{N} a_{nm} \\ \alpha_m &= \alpha_{prior}(m) + \beta_m \end{aligned}$$

and draw new model probabilities

$$r \sim \text{Dir}(\alpha)$$

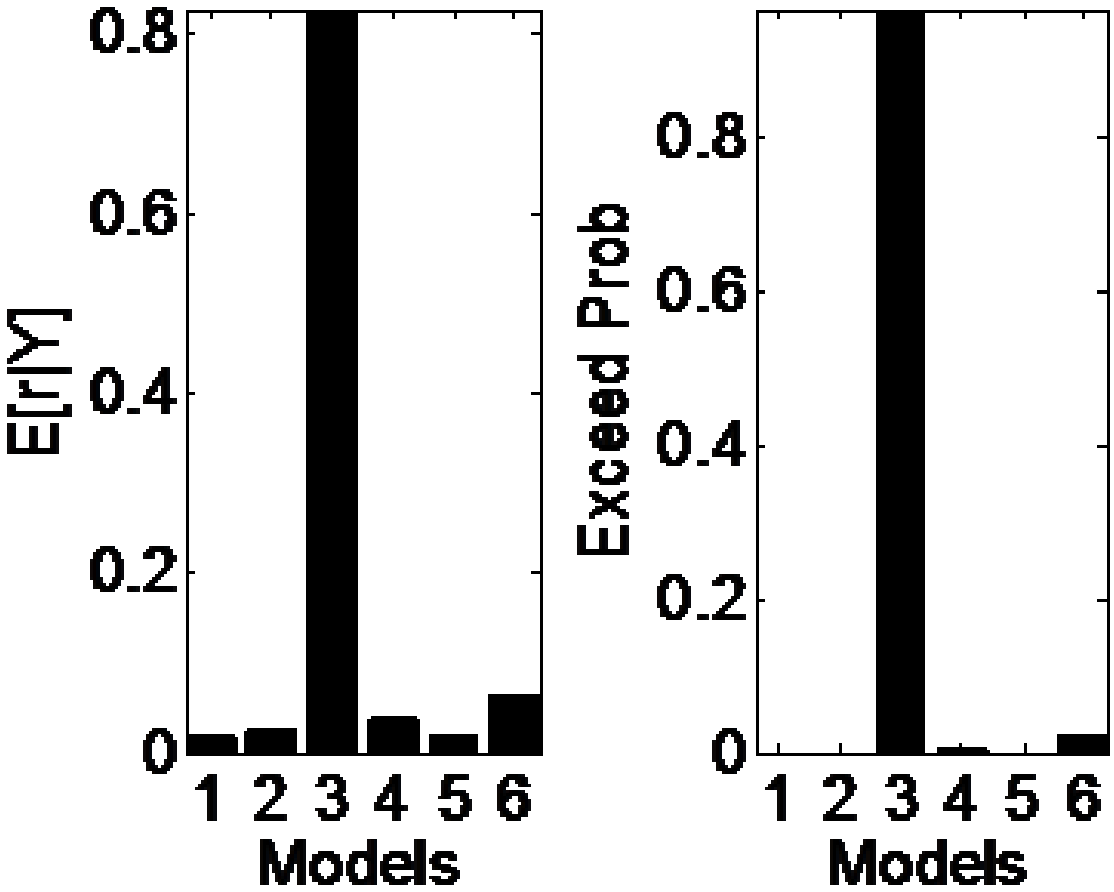
Go back to STEP 2 !

Gibbs Sampling

These steps are repeated N_d times. For the following results we used a total of $N_d = 20,000$ samples and discarded the first 10,000.

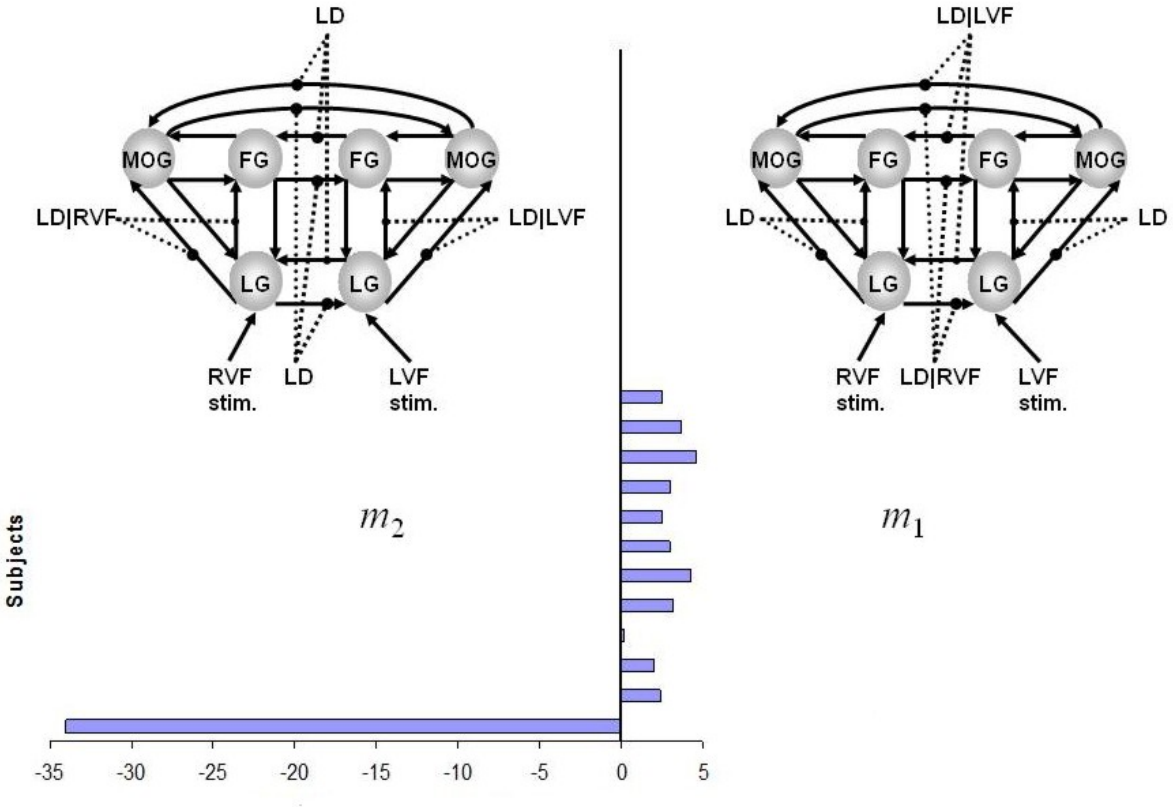
Gibbs Sampling

These remaining samples then constitute our approximation to the posterior distribution $p(r|Y)$. From this density we can compute usual quantities such as the posterior expectation, $E[r|Y]$.



Random Effects

11/12=92% subjects favoured model 1.



$$E[r_1|Y] = 0.84$$
$$p(r_1 > r_2|Y) = 0.99$$

where the latter is called the exceedance probability.

References

H Akaike (1973) Information measures and model selection. Bull. Inst. Int. Stat 50, 277-290.

C. Bishop (2006) Pattern Recognition and Machine Learning. Springer.

K. Friston et al. (2003) Dynamic Causal Models. Neuroimage, 19(4) 1273-1302.

K. Friston et al. (2007) Variational Free Energy and the Laplace Approximation. Neuroimage 34(1), 220-234.

A. Gelman et al. (1995) Bayesian Data Analysis. Chapman and Hall.

R. Henson et al (2002) Face repetition effects in implicit and explicit memory tests as measured by fMRI. Cerebral Cortex, 12, 178-186.

A. Leff et al (2008) The cortical dynamics of intelligible speech. J. Neurosci 28(49):132009-15.

W. Penny et al (2004) Comparing Dynamic Causal Models, Neuroimage 22, 1157-1172.

W. Penny et al (2010) Comparing Families of Dynamic Causal Models, PLoS Computational Biology 6(3), e1000709.

W. Penny et al (2011) Comparing Dynamic Causal Models using AIC, BIC and Free Energy. In revision.

A Raftery (1995) Bayesian model selection in social research. In Marsden, P (Ed) Sociological Methodology, 111-196, Cambridge.

G. Schwarz (1978) Estimating the dimension of a model. Ann. Stat. 6, 461-464.

K Stephan et al (2009). Bayesian model selection for group studies. Neuroimage, 46(4):1004-17