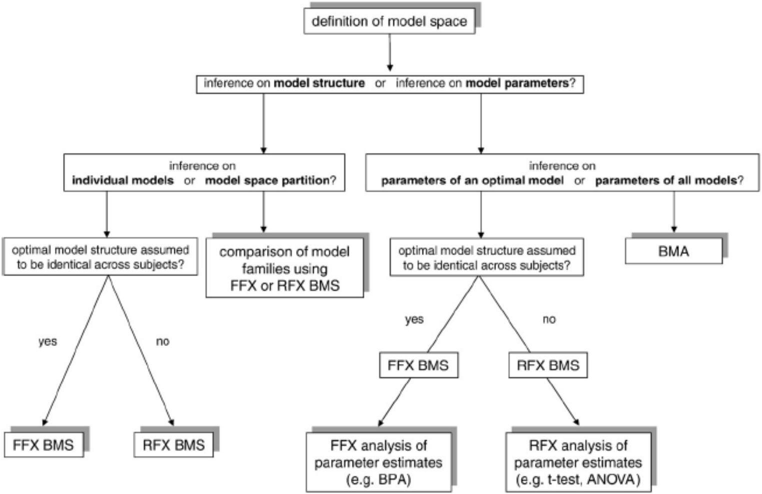
Bayesian Model Selection and Averaging

Will Penny

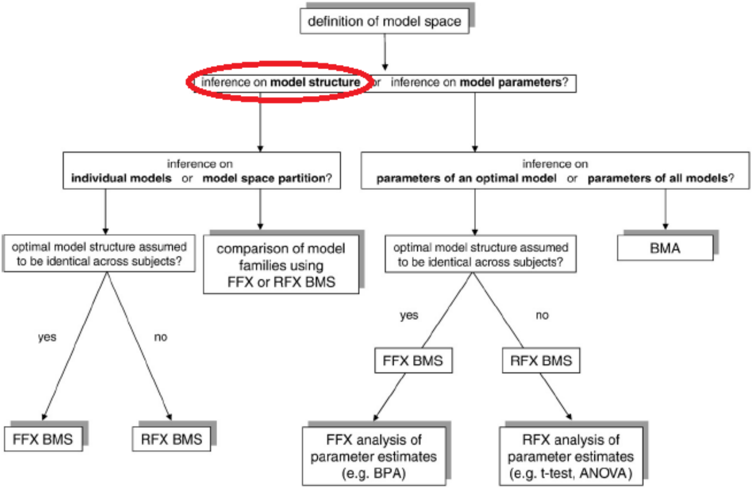
SPM short course for M/EEG,
London 2015

Ten Simple Rules


definition of model space
inference on **model structure** or inference on **model parameters**?
inference on **individual models** or **model space partition**?
inference on **parameters of an optimal model** or **parameters of all models**?
optimal model structure assumed to be identical across subjects?
comparison of model families using FFX or RFX BMS
optimal model structure assumed to be identical across subjects?
BMA
yes
no
yes
no
FFX BMS
RFX BMS
FFX BMS
RFX BMS
FFX analysis of parameter estimates (e.g. BPA)
RFX analysis of parameter estimates (e.g. t-test, ANOVA)

Stephan et al. Neuroimage, 2010

Model Structure



definition of model space

inference on **model structure** or inference on **model parameters**?

inference on **individual models** or **model space partition**?

inference on **parameters of an optimal model** or **parameters of all models**?

optimal model structure assumed to be identical across subjects?

comparison of model families using FFX or RFX BMS

optimal model structure assumed to be identical across subjects?

BMA

yes

no

yes

no

FFX BMS

RFX BMS

FFX BMS

RFX BMS

FFX analysis of parameter estimates (e.g. BPA)

RFX analysis of parameter estimates (e.g. t-test, ANOVA)

Model Evidence

The model evidence is given by integrating out the dependence on model parameters

$$
\begin{aligned}
p(y|m) &= \int p(y,\theta|m) d\theta \\
&= \int p(y|\theta,m) p(\theta|m) d\theta
\end{aligned}
$$

Because we have marginalised over θ the evidence is also known as the marginal likelihood.

For linear Gaussian models there is an analytic expression for the model evidence.

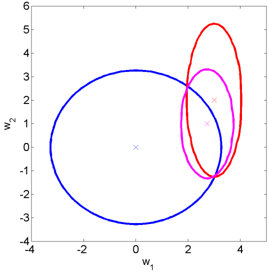
Linear Models

For Linear Models

$$y = Xw + e$$

where X is a design matrix and w are now regression coefficients. For prior mean μ_w, prior covariance C_w, observation noise covariance C_y the posterior distribution is 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

The log model evidence comprises sum squared precision weighted prediction errors and Occam factors

$$
\begin{aligned}
\log p(y|m) &= -\frac{1}{2} e_y^T C_y^{-1} e_y - \frac{1}{2} \log |C_y| - \frac{N_y}{2} \log 2\pi \\
&- \frac{1}{2} e_w^T C_w^{-1} e_w - \frac{1}{2} \log \frac{|C_w|}{|S_w|}
\end{aligned}
$$

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

$$
\begin{aligned}
e_y &= y - X m_w \\
e_w &= m_w - \mu_w
\end{aligned}
$$

Bishop, Pattern Recognition and Machine Learning, 2006

Accuracy and Complexity

The log evidence for model *m* can be split into an accuracy and a complexity term

$$\log p(y|m) = Accuracy(m) - Complexity(m)$$

where

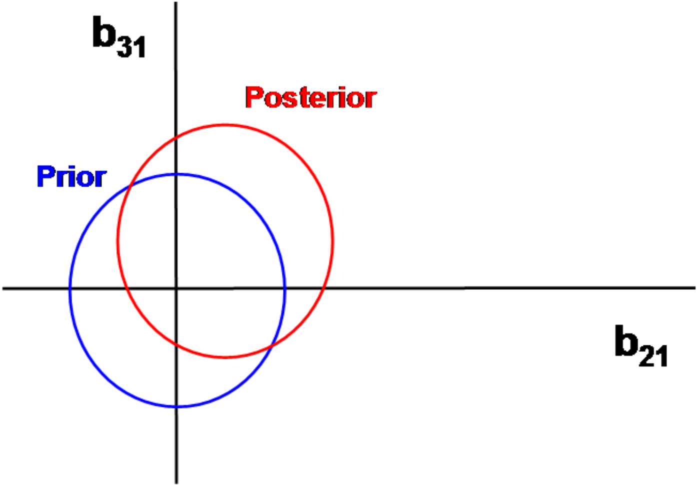
$$Accuracy(m) = -\frac{1}{2} e_y^T C_y^{-1} e_y - \frac{1}{2} \log |C_y| - \frac{N_y}{2} \log 2\pi$$

and

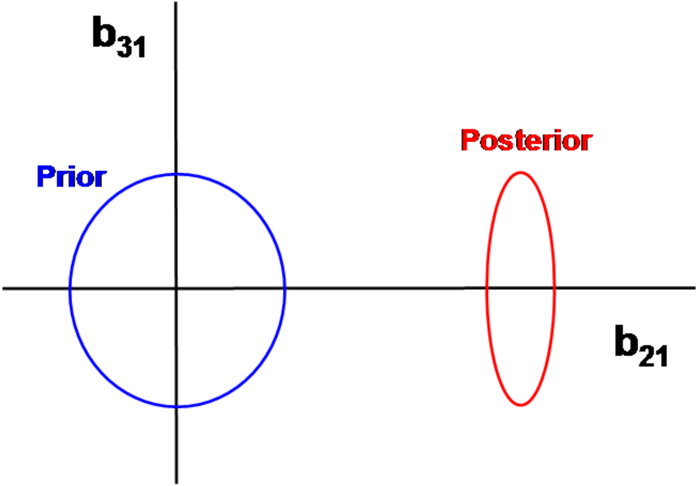
$$\begin{aligned} Complexity(m) &= \frac{1}{2} e_w^T C_w^{-1} e_w + \frac{1}{2} \log \frac{|C_w|}{|S_w|} \\ &\approx KL(prior||posterior) \end{aligned}$$

The Kullback-Leibler divergence measures the distance between probability distributions.

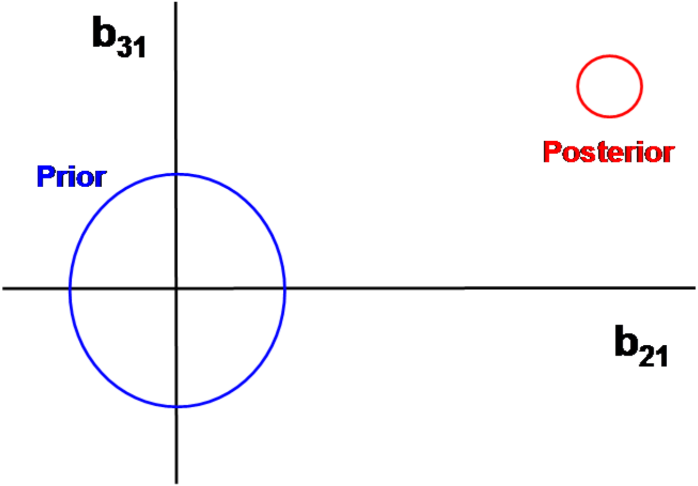
Small KL



Medium KL



Big KL



Nonlinear Models

For nonlinear models, we replace the true posterior with the approximate posterior (m_w, S_w), and the previous expression becomes an approximation to the log model evidence called the (negative) Free Energy

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

where

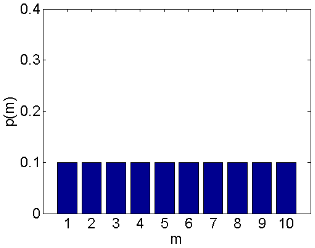
$$\begin{aligned} e_y &= y - g(m_w) \\ e_w &= m_w - \mu_w \end{aligned}$$

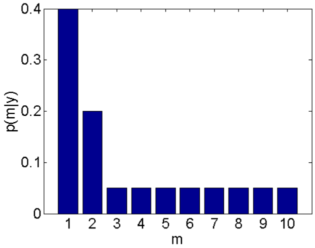
and $g(m_w)$ is the DCM prediction. This is used to approximate the model evidence for DCMs.

W Penny, Neuroimage, 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

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

Bayes Factors

The Bayes factor for model *j* versus *i* is the ratio of model evidences

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

We have

$$B_{ij} = \frac{1}{B_{ji}}$$

Hence

$$\begin{aligned} logB_{ji} &= \log p(y|m=j) - \log p(y|m=i) \\ &= F_j - F_i \end{aligned}$$

Posterior Model Probability

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

$$\begin{aligned}
p(m=i|y) &= \frac{p(y|m=i)}{p(y|m=i)+p(y|m=j)} \\
&= \frac{1}{1+\frac{p(y|m=j)}{p(y|m=i)}} \\
&= \frac{1}{1+B_{ji}} \\
&= \frac{1}{1+\exp(\log B_{ji})} \\
&= \frac{1}{1+\exp(-\log B_{ij})}
\end{aligned}$$

Posterior Model Probability

Hence

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

where is the Bayes factor for model i versus model j and

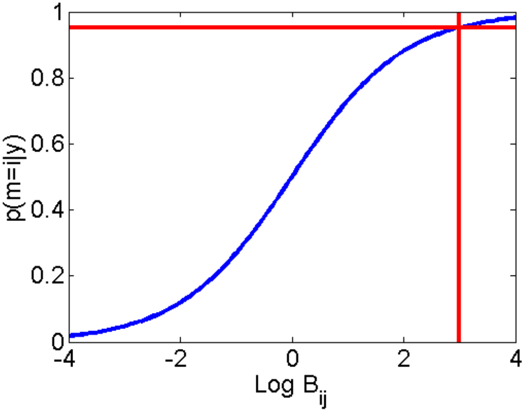
$$\sigma(x) = \frac{1}{1 + \exp(-x)}$$

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

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

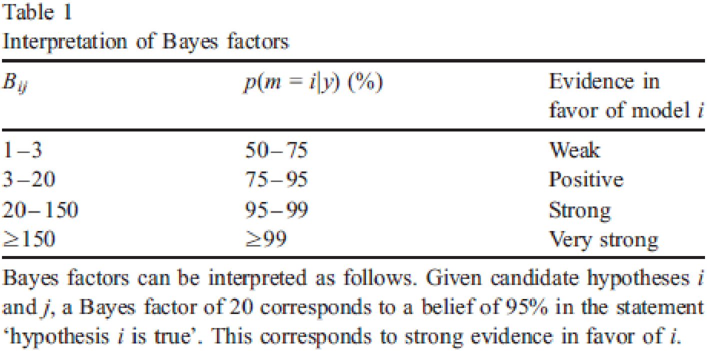


Table 1
Interpretation of Bayes factors

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.

Kass and Raftery, JASA, 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

$$\begin{aligned}\pi^0_{ij} &= \frac{p(m=i)}{p(m=j)} \\ \pi_{ij} &= \frac{p(m=i|y)}{p(m=j|y)}\end{aligned}$$

resepectively, and are related by the Bayes Factor

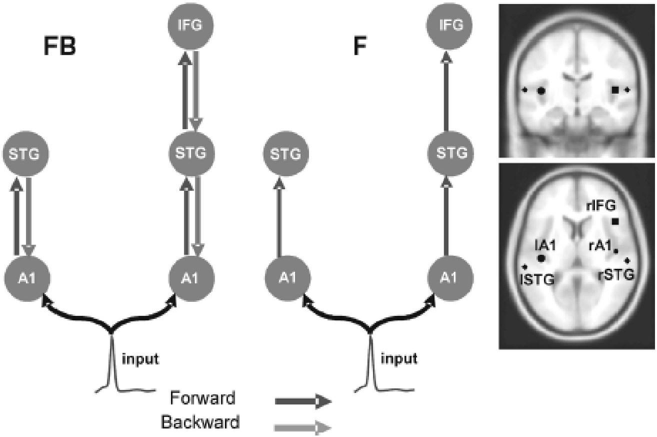
$$\pi_{ij} = B_{ij} \times \pi^0_{ij}$$

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.

Example

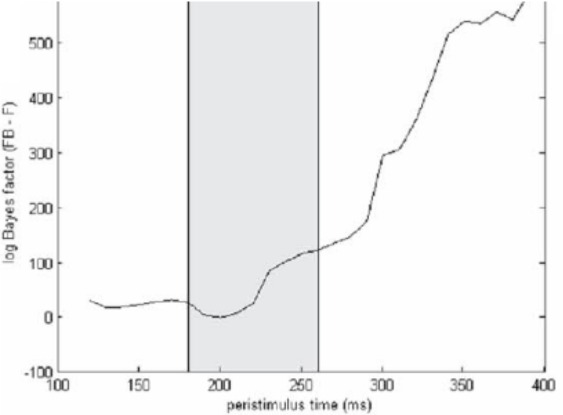
Modelling auditory responses with DCM for ERP



Garrido et al, PNAS, 2007

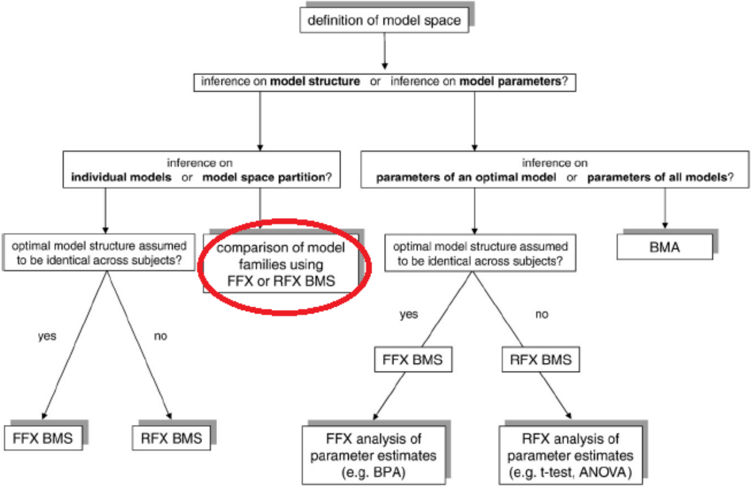
Example

Train DCMs from stimulus onset up to peristimulus time *T*. FB model favoured more heavily as *T* increases.



Evoked responses are generated by feedback loops.

Families



definition of model space

inference on **model structure** or inference on **model parameters**?

inference on **individual models** or **model space partition**?

inference on **parameters of an optimal model** or **parameters of all models**?

optimal model structure assumed to be identical across subjects?

comparison of model families using FFX or RFX BMS

optimal model structure assumed to be identical across subjects?

BMA

yes

no

yes

no

FFX BMS

RFX BMS

FFX BMS

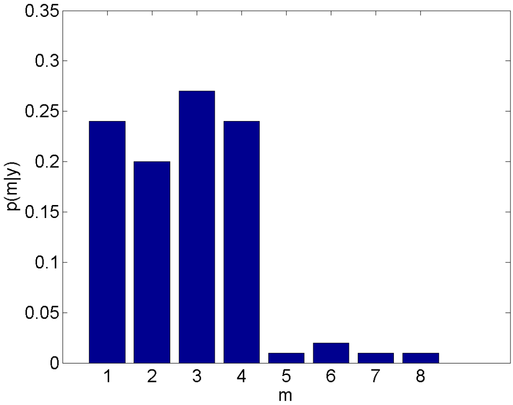
RFX BMS

FFX analysis of parameter estimates (e.g. BPA)

RFX analysis of parameter estimates (e.g. t-test, ANOVA)

Posterior Model Probabilities

Say we've fitted 8 DCMs and get the following distribution over models

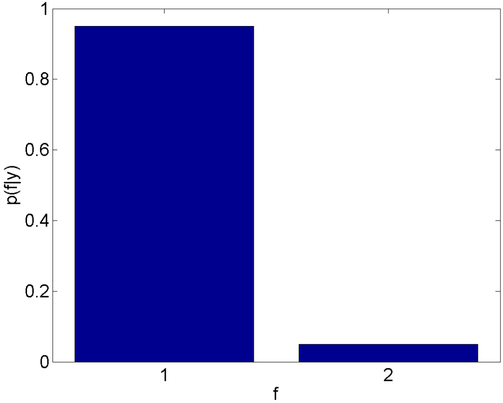


Similar models share probability mass (dilution). The probability for any single model can become very small esp. for large model spaces.

Model Families

Assign model m to family f eg. first four to family one, second four to family two. The posterior family probability is then

$$p(f|y) = \sum_{m \in S_f} p(m|y)$$



Different Sized Families

If we have K families, then to avoid bias in family inference we wish to have a uniform prior at the family level

$$p(f) = \frac{1}{K}$$

The prior family probability is related to the prior model probability

$$p(f) = \sum_{m \in S_f} p(m)$$

where the sum is over all N_f models in family f. So we set

$$p(m) = \frac{1}{KN_f}$$

for all models in family f before computing $p(m|y)$. This allows us to have families with unequal numbers of models.

Penny et al. PLOS-CB, 2010.

Different Sized Families

So say we have two families. We want a prior for each family of $p(f) = 0.5$.

If family one has $N_1 = 2$ models and family two has $N_2 = 8$ models, then we set

$$p(m) = \frac{1}{2} \times \frac{1}{2} = 0.25$$

for all models in family one and

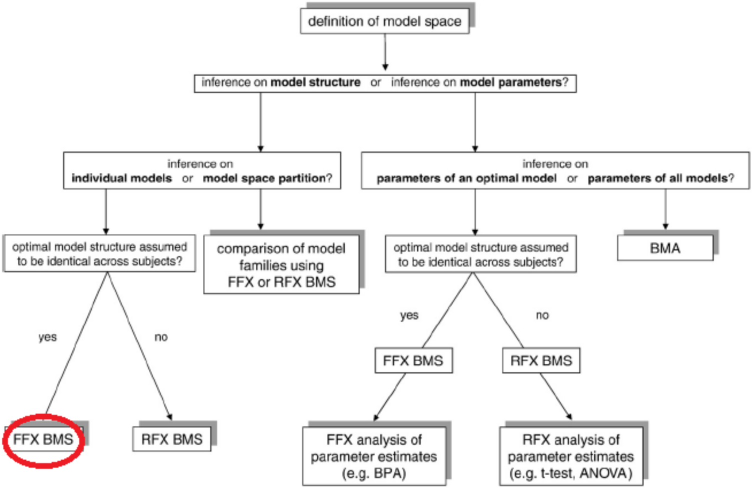
$$p(m) = \frac{1}{2} \times \frac{1}{8} = 0.0625$$

for all models in family two.

These are then used in Bayes rule for models

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

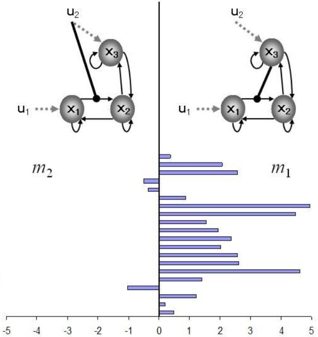
Fixed Effects BMS


definition of model space
inference on **model structure** or inference on **model parameters**?
inference on **individual models** or **model space partition**?
inference on **parameters of an optimal model** or **parameters of all models**?
optimal model structure assumed to be identical across subjects?
comparison of model families using FFX or RFX BMS
optimal model structure assumed to be identical across subjects?
BMA
yes
no
yes
no
FFX BMS
RFX BMS
FFX BMS
RFX BMS
FFX analysis of parameter estimates (e.g. BPA)
RFX analysis of parameter estimates (e.g. t-test, ANOVA)

Fixed Effects BMS

Two models, twenty subjects.

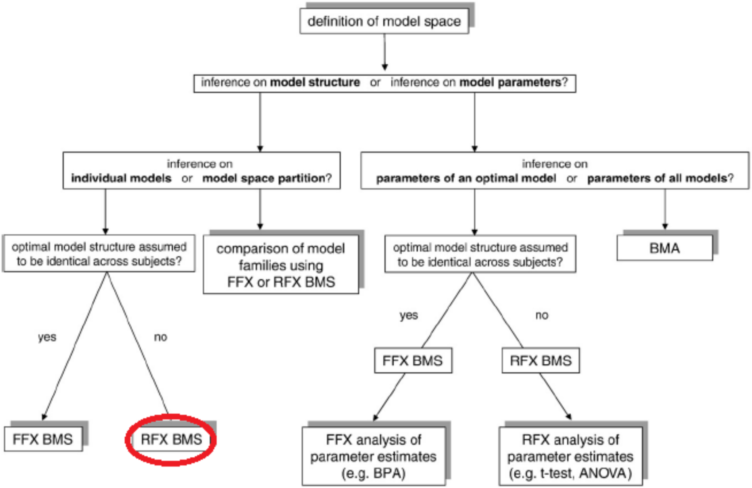
$$\log p(Y|m) = \sum_{n=1}^{N} \log p(y_n|m)$$



The Group Bayes Factor (GBF) is

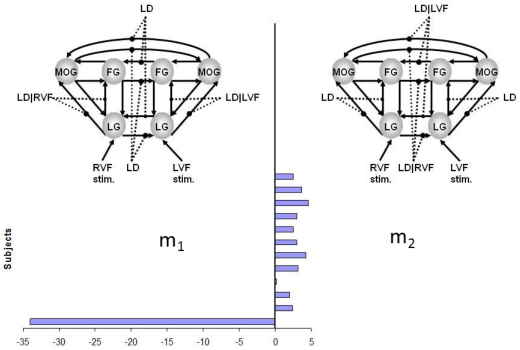
$$B_{ij} = \prod_{n=1}^{N} B_{ij}(n)$$

Random Effects BMS


definition of model space
inference on **model structure** or inference on **model parameters**?
inference on
individual models or **model space partition**?
inference on
parameters of an optimal model or **parameters of all models**?
optimal model structure assumed to be identical across subjects?
comparison of model families using FFX or RFX BMS
optimal model structure assumed to be identical across subjects?
BMA
yes
no
yes
no
FFX BMS
RFX BMS
FFX BMS
RFX BMS
FFX analysis of parameter estimates (e.g. BPA)
RFX analysis of parameter estimates (e.g. t-test, ANOVA)

Random Effects BMS

Stephan et al. J. Neurosci, 2007



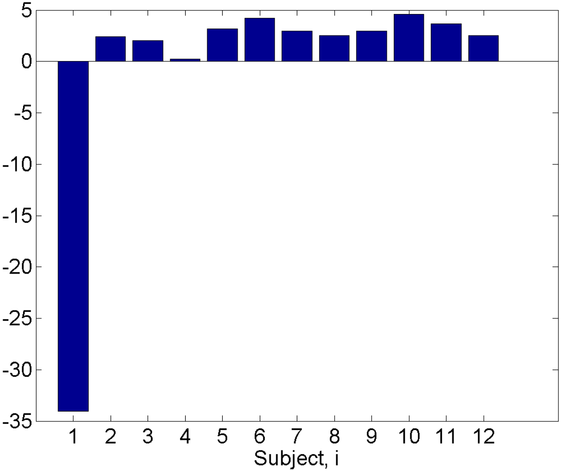
11/12=92% subjects favour model 2.

$GBF = 15$ in favour of model 1. FFX inference does not agree with the majority of subjects.

RFX Model Inference

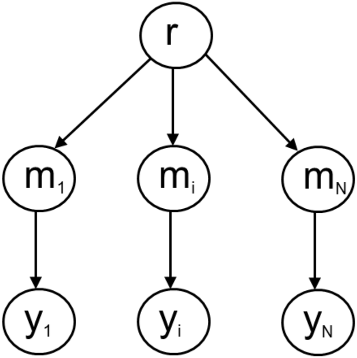
Log Bayes Factor in favour of model 2

$$\log \frac{p(y_i|m_i = 2)}{p(y_i|m_i = 1)}$$



RFX Model Inference

Model frequencies r_k, model assignments m_i, subject data y_i.

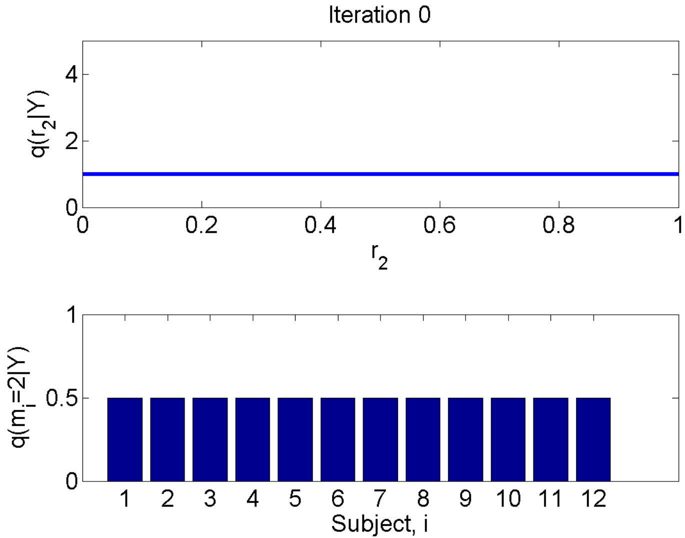


Approximate posterior

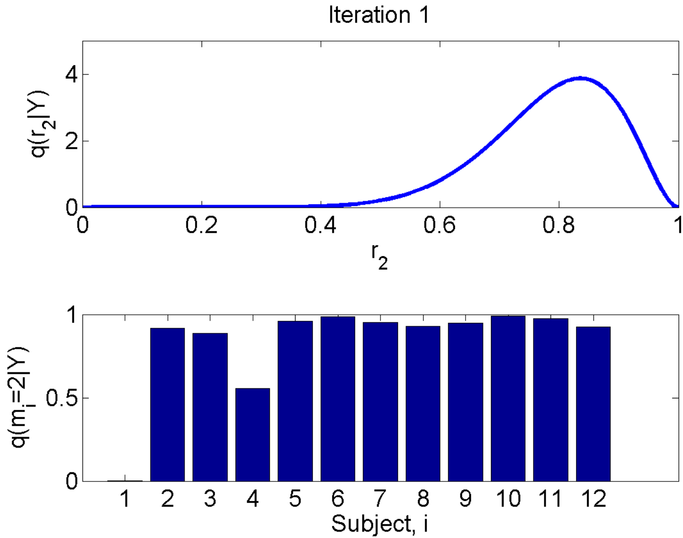
$$q(r, m|Y) = q(r|Y)q(m|Y)$$

Stephan et al, Neuroimage, 2009.

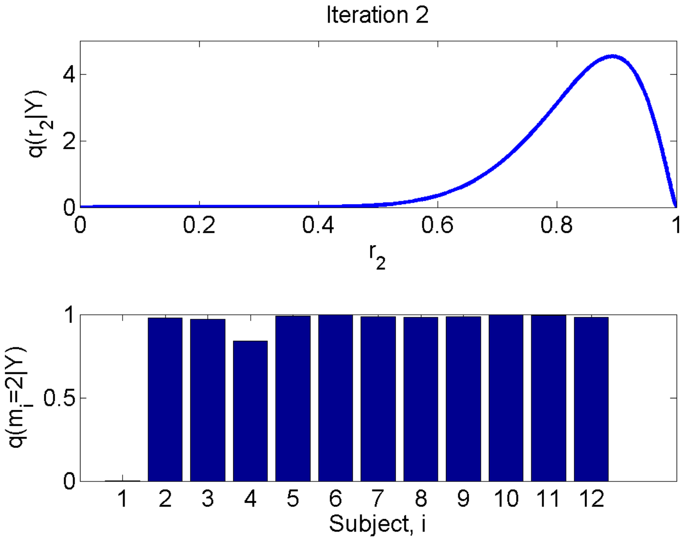
RFX Model Inference



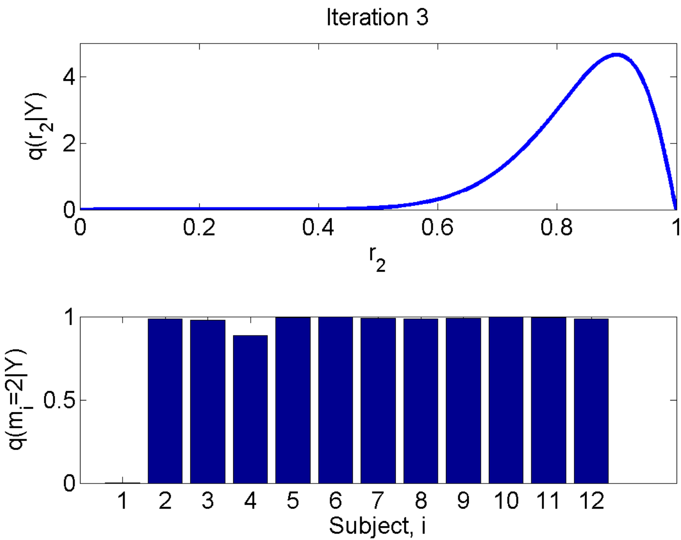
RFX Model Inference



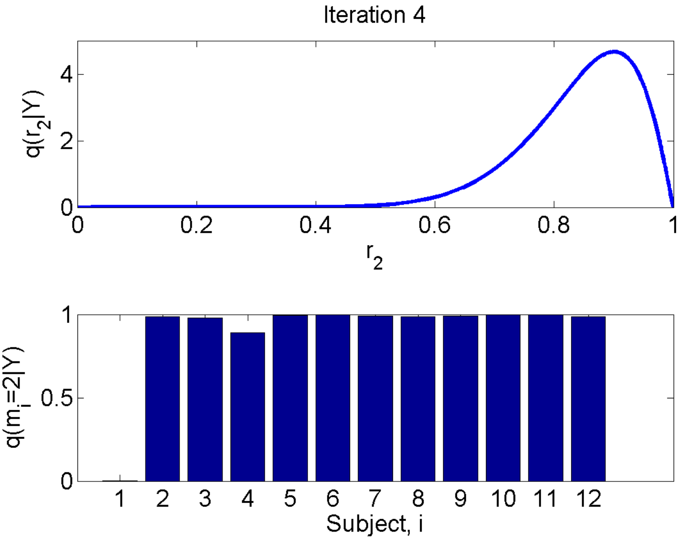
RFX Model Inference



RFX Model Inference

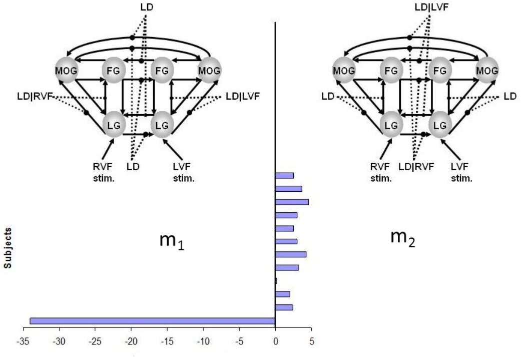


RFX Model Inference



Random Effects

11/12=92% subjects favoured model 2.

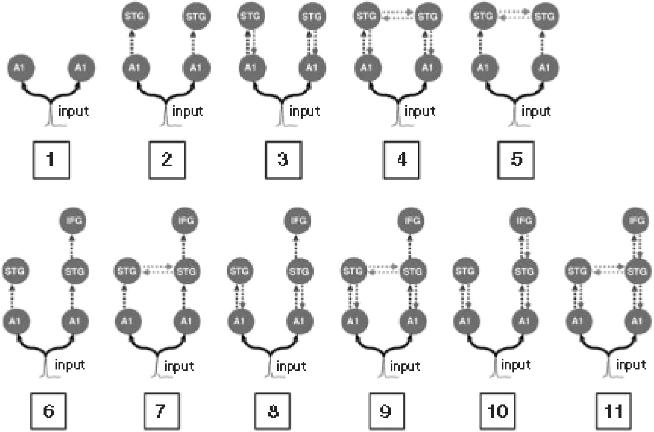


$$
\begin{aligned}
E[r_2|Y] &= 0.84 \\
p(r_2 > r_1|Y) &= 0.99
\end{aligned}
$$

where the latter is called the exceedance probability.

Example

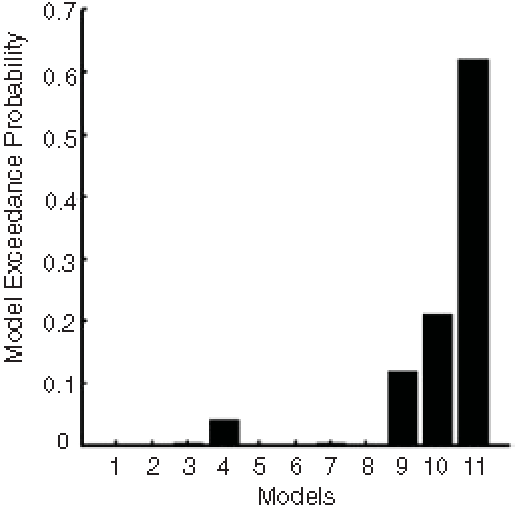
Auditory responses to stimuli with ‘roving’ frequencies modelled with DCM for ERP.



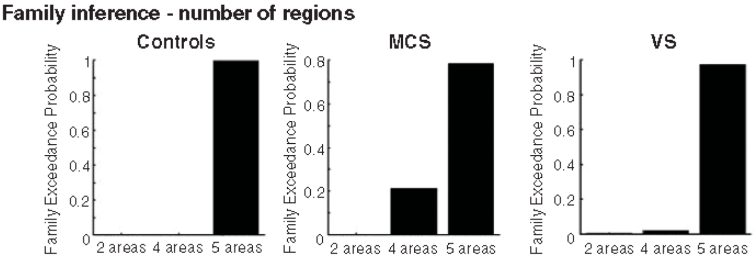
Boly et al, Science, 2011.

Example

Model Exceedance Probabilities

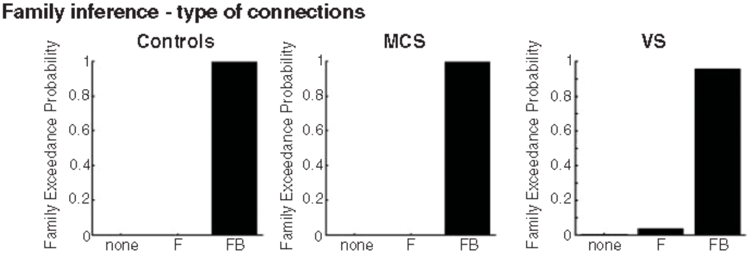


Example



This study used people in a Minimally Conscious State (MCS), in a Vegetative State (VS) or in a normal level of consciousness (Controls).

Example



This study used people in a Minimally Conscious State (MCS), in a Vegetative State (VS) or in a normal level of consciousness (Controls).

Protected Exceedance Probabilities

The use of Exceedance Probabilities (xp's) assumes the frequencies are different for each model.

But what if the model frequencies are all the same ? (H_0: omnibus hypothesis)

Let $p_0 = p(H_0|Y)$. Then the (posterior) probability that frequencies are different is $1 - p_0$.

Rigoux et al. (*Neuroimage, 2014*) show how to compute p_o and then define Protected Exceedance Probabilities as

$$pxp = xp(1 - p_o) + \frac{1}{K} p_o$$

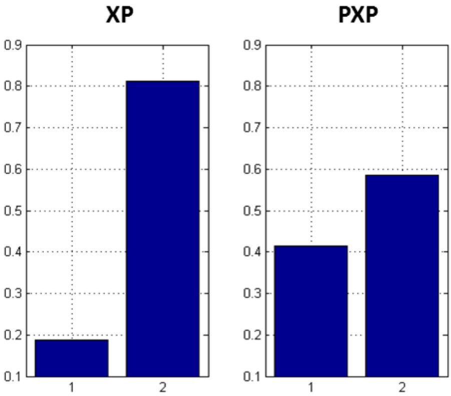
where K is the number of models.

p_o also referred to as 'Bayes Omnibus Risk (BOR)'.

Protected Exceedance Probabilities

The function `spm_BMS.m` reports *pxp*'s and p_0.

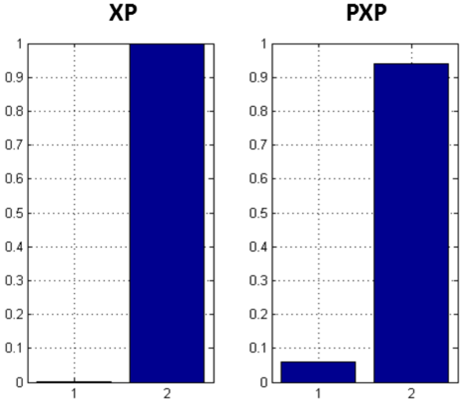
Synthetic data ($K = 2$ models, $N = 12$ subjects, mean log evidence difference=0) .



We have $p_0 = 0.72$.

Protected Exceedance Probabilities

Synthetic data ($K = 2$ models, $N = 12$ subjects, mean log evidence difference=1).


XP
PXP

We have $p_0 = 0.11$.

PXPs also very useful for large K.

Dependence on Comparison Set

The ranking of models from RFX inference can depend on the comparison set.

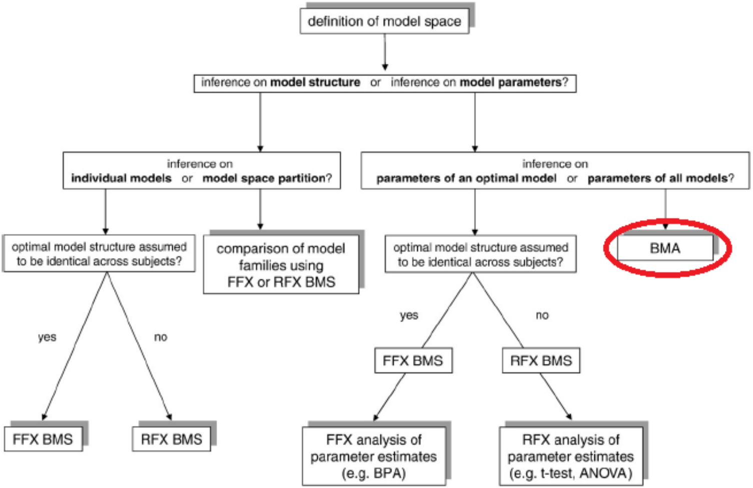
Say we have two models with 7 subjects prefering model 1 and 10 ten subjects preferring model 2. The model frequencies are $r_1 = 7/17 = 0.41$ and $r_2 = 10/17 = 0.59$.

Now say we add a third model which is similar to the second, and that 4 of the subjects that used to prefer model 2 now prefer model 3. The model frequencies are now $r_1 = 7/17 = 0.41$, $r_2 = 6/17 = 0.35$ and $r_3 = 4/17 = 0.24$.

This is like voting in elections.

Penny et al. PLOS-CB, 2010.

Model Averaging



definition of model space

inference on **model structure** or inference on **model parameters**?

inference on
individual models or **model space partition**?

inference on
parameters of an optimal model or **parameters of all models**?

optimal model structure assumed to be identical across subjects?

yes

no

FFX BMS

RFX BMS

comparison of model families using FFX or RFX BMS

optimal model structure assumed to be identical across subjects?

yes

no

FFX BMS

RFX BMS

FFX analysis of parameter estimates (e.g. BPA)

RFX analysis of parameter estimates (e.g. t-test, ANOVA)

BMA

Model Averaging

Each DCM.mat file stores the posterior mean (DCM.Ep) and covariance (DCM.Cp) for each fitted model. This defines the posterior mean over parameters for that model, $p(\theta|m,y)$.

This can then be combined with the posterior model probabilities $p(m|y)$ to compute a posterior over parameters

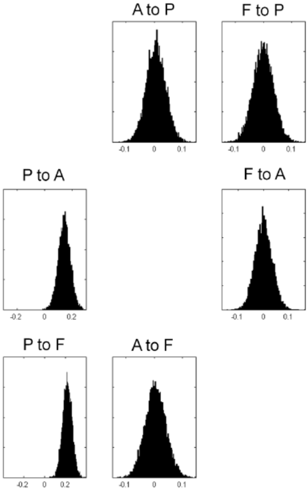
$$\begin{aligned} p(\theta|y) &= \sum_m p(\theta, m|y) \\ &= \sum_m p(\theta|m,y)p(m|y) \end{aligned}$$

which is independent of model assumptions (within the chosen set). Here, we marginalise over m.

The sum over m could be restricted to eg. models within the winning family.

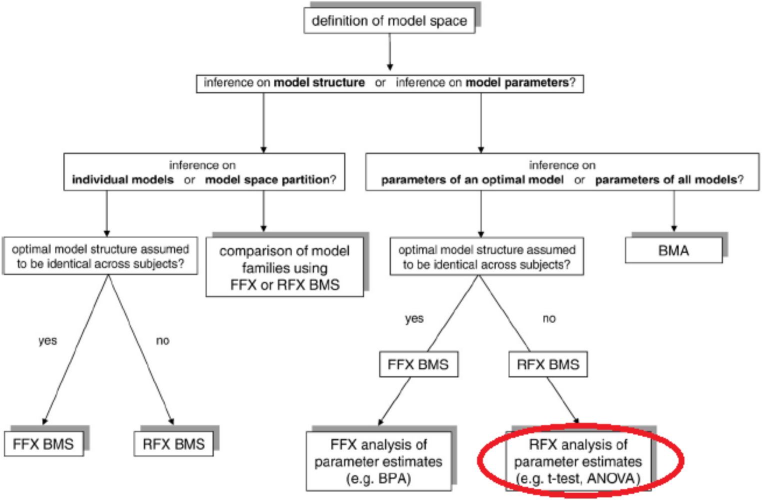
Model Averaging

The distribution $p(\theta|y)$ can be gotten by sampling; sample m from $p(m|y)$, then sample θ from $p(\theta|m, y)$.



If a connection doesn't exist for model m the relevant samples are set to zero.

RFX Parameter Inference



definition of model space

inference on **model structure** or inference on **model parameters**?

inference on
individual models or **model space partition**?

inference on
parameters of an optimal model or **parameters of all models**?

optimal model structure assumed to be identical across subjects?

comparison of model families using FFX or RFX BMS

optimal model structure assumed to be identical across subjects?

BMA

yes

no

yes

no

FFX BMS

RFX BMS

FFX BMS

RFX BMS

FFX analysis of parameter estimates (e.g. BPA)

RFX analysis of parameter estimates (e.g. t-test, ANOVA)

RFX Parameter Inference

If ith subject has posterior mean value m_i we can use these in Summary Statistic approach for group parameter inference (eg two-sample t-tests for control versus patient inferences).

eg P to A connection in controls: 0.20, 0.12, 0.32, 0.11, 0.01, ...

eg P to A connection in patients: 0.50, 0.42, 0.22, 0.71, 0.31, ...

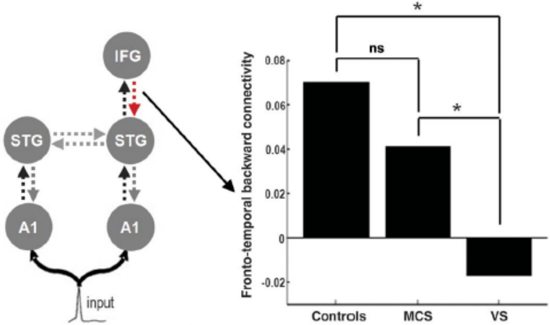
Two sample t-test shows the P to A connection is stronger in patients than controls ($p < 0.05$). Or one sample t-tests if we have a single group.

RFX is more conservative than BPA.

Example

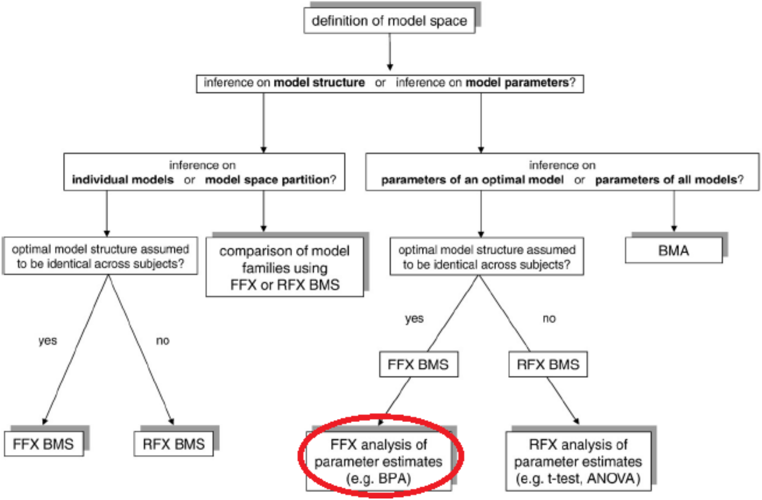
T-tests on backward connection from IFG to STG

Fig. 4. Quantitative effective connectivity analysis revealed that the only significant difference between VS patients and controls was an impairment of backward connectivity from frontal to temporal cortex. MCS subjects showed significantly preserved connectivity compared with VS subjects and were not significantly different from controls.



Boly et al. Science, 2011

FFX Parameter Inference


definition of model space
inference on model structure or inference on model parameters?
inference on individual models or model space partition?
inference on parameters of an optimal model or parameters of all models?
optimal model structure assumed to be identical across subjects?
comparison of model families using FFX or RFX BMS
optimal model structure assumed to be identical across subjects?
BMA
yes
no
yes
no
FFX BMS
RFX BMS
FFX BMS
RFX BMS
FFX analysis of parameter estimates (e.g. BPA)
RFX analysis of parameter estimates (e.g. t-test, ANOVA)

FFX Parameter Inference

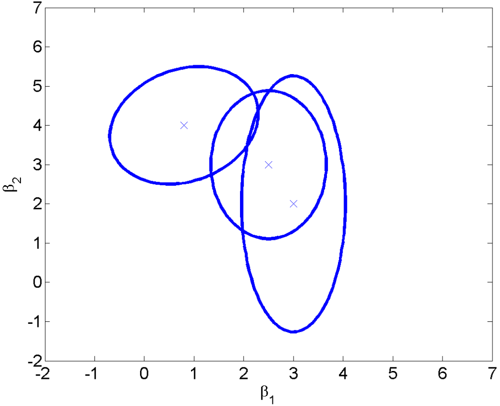
RFX parameter inference (eg. t-tests, F-tests) - allow for variability over eg. subjects.

FFX parameter inference - assumes no variability over eg. subjects/sessions.

FFX parameter inference - implemented using ‘Bayesian Parameter Averaging’ (BPA)

Bayesian Parameter Averaging

If for the *i*th subject the posterior mean and precision are μ_i and Λ_i



Three subjects shown.

Bayesian Parameter Averaging

If for the *i*th subject the posterior mean and precision are μ_i and Λ_i then the posterior mean and precision for the group are

$$\begin{aligned}\Lambda &= \sum_{i=1}^{N} \Lambda_i \\ \mu &= \Lambda^{-1} \sum_{i=1}^{N} \Lambda_i \mu_i\end{aligned}$$

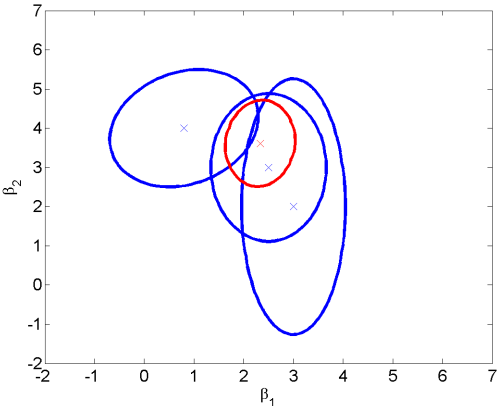
Kasses et al, Neuroimage, 2010.

This is a FFX analysis where each subject adds to the posterior precision.

Bayesian Parameter Averaging

$$\Lambda = \sum_{i=1}^{N} \Lambda_i$$

$$\mu = \Lambda^{-1} \sum_{i=1}^{N} \Lambda_i \mu_i$$



Informative Priors

If for the *i*th subject the posterior mean and precision are μ_i and Λ_i then the posterior mean and precision for the group are

$$\begin{aligned}
\Lambda &= \sum_{i=1}^{N} \Lambda_i - (N-1)\Lambda_0 \\
\mu &= \Lambda^{-1} \left(\sum_{i=1}^{N} \Lambda_i \mu_i - (N-1)\Lambda_0 \mu_0 \right)
\end{aligned}$$

Formulae augmented to accomodate non-zero priors Λ_0 and μ_0.

References

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

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

W. Penny (2011) Comparing Dynamic Causal Models using AIC, BIC and Free Energy. Neuroimage Available online 27 July 2011.

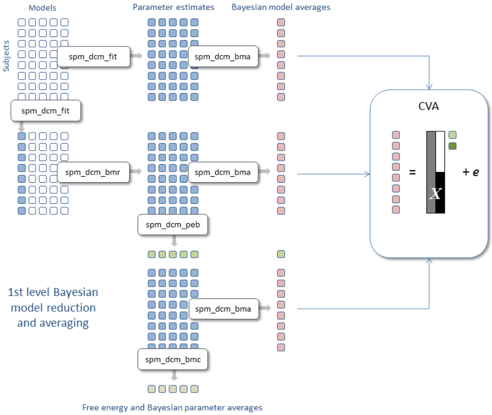
W. Penny et al (2010) Comparing Families of Dynamic Causal Models. PLoS CB, 6(3).

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

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

Forthcoming

A new method for taking fitted DCMs from a group of subjects, and ‘refitting’ them according to a mixed effects model.



Models

Parameter estimates

Bayesian model averages

Subjects

spm_dcm_fit

spm_dcm_bma

spm_dcm_fit

spm_dcm_bmr

spm_dcm_bma

spm_dcm_peb

CVA

$= X + e$

1st level Bayesian model reduction and averaging

spm_dcm_bma

spm_dcm_bmc

Free energy and Bayesian parameter averages

The method is highly computationally efficient and is very flexible, allowing e.g. for parametric random effects, and comparison of models at the group level.

K. Friston et al. Bayesian model reduction and empirical Bayes for group (DCM) studies, Submitted, 2015.