Bayesian Model Selection and Averaging

Will Penny

SPM short course for M/EEG, London 2013
Ten Simple Rules

Stephan et al. Neuroimage, 2010
Model Structure
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

$$B_{ij} = \frac{1}{B_{ji}}$$
Posterior Model Probability

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

\[
p(m = i | y) = \frac{p(y | m = i)}{p(y | m = i) + p(y | m = j)}
\]

\[
= \frac{1}{1 + \frac{p(y | m = j)}{p(y | m = i)}}
\]

\[
= \frac{1}{1 + B_{ji}}
\]

\[
= \frac{1}{1 + \exp(\log B_{ji})}
\]

\[
= \frac{1}{1 + \exp(-\log B_{ij})}
\]
Posterior Model Probability

Hence

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

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

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>
<thead>
<tr>
<th>Table 1</th>
<th>Interpretation of Bayes factors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( B_{ij} )</td>
</tr>
<tr>
<td>( 1-3 )</td>
<td>( 50-75 )</td>
</tr>
<tr>
<td>( 3-20 )</td>
<td>( 75-95 )</td>
</tr>
<tr>
<td>( 20-150 )</td>
<td>( 95-99 )</td>
</tr>
<tr>
<td>( \geq 150 )</td>
<td>( \geq 99 )</td>
</tr>
</tbody>
</table>

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

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

respectively, 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.
Model Evidence

The model evidence is not, in general, straightforward to compute since computing it involves integrating out the dependence on model parameters

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

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

Because we have marginalised over \( \theta \) the evidence is also known as the marginal likelihood.

But for linear, Gaussian models there is an analytic solution.
Linear Models

For Linear Models

\[ y = Xw + e \]

where \( X \) is a design matrix and \( w \) are now regression coefficients. For prior mean \( \mu_w \), prior covariance \( C_w \), observation noise covariance \( C_y \) the posterior distribution is given by

\[
S_w^{-1} = X^T C_y^{-1} X + C_w^{-1}
\]

\[
m_w = S_w \left( X^T C_y^{-1} y + C_w^{-1} \mu_w \right)
\]
Bayesian Model Selection and Averaging

Will Penny

Model structure
- Bayes factors
- Linear Models
- Complexity
- Nonlinear Models

Families

Model Averaging
- FFX Model Inference
- RFX Model Inference

Parameter Inference
- FFX Parameter Inference

References
Structure Inference

- Definition of model space
- Inference on model structure or inference on model parameters?
  - Inference on individual models or model space partition?
    - Optimal model structure assumed to be identical across subjects?
      - Yes: FFX BMS
      - No: RFX BMS
    - Comparison of model families using FFX or RFX BMS
  - Inference on parameters of an optimal model or parameters of all models?
    - Optimal model structure assumed to be identical across subjects?
      - Yes: FFX analysis of parameter estimates (e.g. BPA)
      - No: RFX BMS
    - BMA
- FFX Model Inference
- RFX Model Inference
- FFX Parameter Inference
- RFX Parameter Inference
- References
The log model evidence comprises sum squared precision weighted prediction errors and Occam factors

\[
\log p(y|m) = -\frac{1}{2} e_y^T C_y^{-1} e_y - \frac{1}{2} \log |C_y| - \frac{N_y}{2} \log 2\pi \\
- \frac{1}{2} e_w^T C_w^{-1} e_w - \frac{1}{2} \log \left| \frac{C_w}{S_w} \right|
\]

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

\[
e_y = y - X m_w \\
e_w = m_w - \mu_w
\]

_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

$$Complexity(m) = \frac{1}{2} e_w^T C_w^{-1} e_w + \frac{1}{2} \log \frac{|C_w|}{|S_w|} \approx KL(prior||posterior)$$
Bayesian Model Selection and Averaging
Will Penny

Model structure
- Bayes factors
- Linear Models
- Complexity
- Nonlinear Models

Families

Model Averaging
- FFX Model Inference
- RFX Model Inference

FFX Parameter Inference
RFX Parameter Inference

References

Small KL

Prior

Posterior

$\text{b}_{31}$

$\text{b}_{21}$
Medium KL
Big KL

Bayesian Model Selection and Averaging
Will Penny

Model structure
Bayes factors
Linear Models
Complexity
Nonlinear Models

Families

Model Averaging
FFX Model Inference
RFX Model Inference

FFX Parameter Inference
RFX Parameter Inference

References
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

\[
F = -\frac{1}{2} e_y^T C_y^{-1} e_y - \frac{1}{2} \log |C_y| - \frac{N_y}{2} \log 2\pi - \frac{1}{2} e_w^T C_w^{-1} e_w - \frac{1}{2} \log \frac{|C_w|}{|S_w|}
\]

where

\[
e_y = y - g(m_w)
\]

\[
e_w = m_w - \mu_w
\]

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

\[
p(m|y) = \frac{p(y|m)p(m)}{p(y)}
\]
Families
Posterior Model Probabilities

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

![Histogram showing posterior model probabilities](image)

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

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

Bayesian Model Selection and Averaging
Will Penny

Model structure
Bayes factors
Linear Models
Complexity
Nonlinear Models

Families

Model Averaging
FFX Model Inference
RFX Model Inference
FFX Parameter Inference
RFX Parameter Inference

References
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

$$p(\theta|y) = \sum_m p(\theta, m|y)$$

$$= \sum_m p(\theta|m, y)p(m|y)$$

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 $\theta$ from $p(\theta | m, y)$.

If a connection doesn’t exist for model $m$ the relevant samples are set to zero.
Group Parameter Inference

If $i$th 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 ($p < 0.05$).
Bayesian Model Selection and Averaging
Will Penny

Model structure
Bayes factors
Linear Models
Complexity
Nonlinear Models

Families

Model Averaging

FFX Model Inference

RFX Model Inference

FFX Parameter Inference

RFX Parameter Inference

References

Fixed Effects BMS
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

Bayesian Model Selection and Averaging
Will Penny

Model structure
Bayes factors
Linear Models
Complexity
Nonlinear Models

Families

Model Averaging
FFX Model Inference
RFX Model Inference

FFX Parameter Inference
RFX Parameter Inference

References
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

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

RFX Model Inference
RFX Model Inference

![Graph showing iteration 2 and subject inference](image-url)
RFX Model Inference

Model structure
Bayes factors
Linear Models
Complexity
Nonlinear Models
Families
Model Averaging
FFX Model Inference
RFX Model Inference
FFX Parameter Inference
RFX Parameter Inference
References
Random Effects

11/12 = 92% subjects favoured model 2.

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

where the latter is called the exceedance probability.
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 preferring model 1 and 10 ten subjects preferring model 2. The model frequencies are \( r_1 = \frac{7}{17} = 0.41 \) and \( r_2 = \frac{10}{17} = 0.59 \).

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 \( r_1 = \frac{7}{17} = 0.41 \), \( r_2 = \frac{6}{17} = 0.35 \) and \( r_3 = \frac{4}{17} = 0.24 \).

This is like voting in elections.

Bayesian Parameter Averaging

- Bayes factors
- Linear Models
- Complexity
- Nonlinear Models
- Model Averaging
- FFX Model Inference
- RFX Model Inference
- FFX Parameter Inference
- RFX Parameter Inference
- References
Bayesian Parameter Averaging

If for the $i$th subject the posterior mean and precision are $\mu_i$ and $\Lambda_i$

Three subjects shown.
Bayesian Parameter Averaging

If for the $i$th subject the posterior mean and precision are $\mu_i$ and $\Lambda_i$ then the posterior mean and precision for the group are

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

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

*Kasses et al, Neuroimage, 2010.*

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

\[ \Lambda = \sum_{i=1}^{N} \Lambda_i \]

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

References
Informative Priors

If for the $i$th subject the posterior mean and precision are $\mu_i$ and $\Lambda_i$ then the posterior mean and precision for the group are

\[ \Lambda = \sum_{i=1}^{N} \Lambda_i - (N - 1)\Lambda_0 \]

\[ \mu = \Lambda^{-1} \left( \sum_{i=1}^{N} \Lambda_i \mu_i - (N - 1)\Lambda_0 \mu_0 \right) \]

Formulae augmented to accommodate non-zero priors $\Lambda_0$ and $\mu_0$. 

Bayesian Model Selection and Averaging
Will Penny
RFX Parameter Inference
RFX Parameter Inference

If \( i \)th 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.
References


