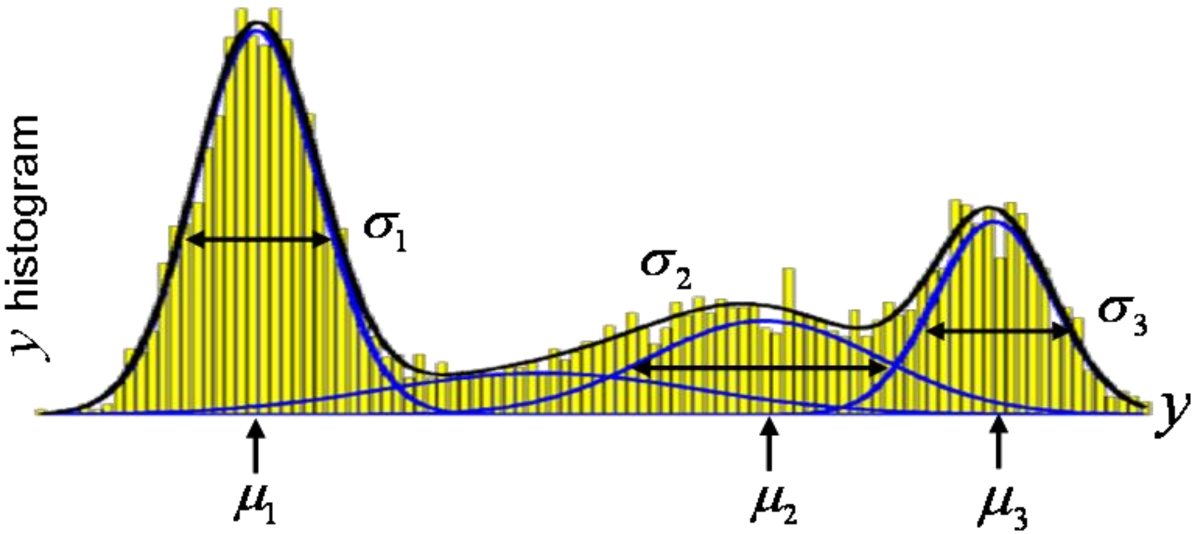
Bayesian Inference

Jérémie Mattout, PhD

Lyon Neuroscience Research Center, France
Brain Dynamics & Cognition Team (DYCOG)

With many thanks to Jean Daunizeau, Guillaume Flandin, Karl Friston & Will Penny

Some prior belief to start with



The ubiquity of Bayesian inference

Bayesian inference to test biophysical models of neuronal activity (neuroimaging data)

NeuroImage **16**, 465–483 (2002)
doi:10.1006/nimg.2002.1090, available online at http://www.idealibrary.com on IDEAL®

Classical and Bayesian Inference in Neuroimaging: Theory

K. J. Friston, W. Penny, C. Phillips, S. Kiebel, G. Hinton, and J. Ashburner

The Wellcome Department of Imaging Neuroscience, and The Gatsby Computational Neuroscience Unit, University College London, Queen Square, London WC1N 3BG, United Kingdom

Bayesian inference to test computational models of the mind (behavioural data)

A comparison of fixed-step-size and Bayesian staircases for sensory threshold estimation
Alcalá-Quintana, Rocío; García-Pérez, Miguel A.
Spatial Vision, 2007

Bayesian inference as a model of cognitive processes (sensory data)

How to Grow a Mind: Statistics, Structure, and Abstraction

Joshua B. Tenenbaum,[1*] Charles Kemp,[2] Thomas L. Griffiths,[3] Noah D. Goodman[4]

Science 2011

ELSEVIER journal homepage: www.elsevier.com/locate/ynimg

Review

The history of the future of the Bayesian brain

Karl Friston *

The Wellcome Trust Centre for Neuroimaging, UCL, 12 Queen Square, London WC1N 3BG, UK

What does « inference » mean ?

Statistics: concerned with the collection, analysis and interpretation of data to make decisions

Applied statistics

Descriptive statistics
summary statistics, graphics…

Probability

Inferential statistics
Data interpretations, decision making
(Modeling, accounting for randomness and unvertainty, hypothesis testing, infering hidden parameters)

The logic of probability

"The true logic for this world is the calculus of Probabilities, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind."

***James Clerk Maxwell** (1850)*

- ***Logic***

You know that : if A is true, then B is true

Then if B is false, you **deduce** for sure that A is false

→ Binary values

- ***Plausibility***

You know that : if A is true, then B is true

What if A is false ? Isn't B a little less likely to be true ?
What if you observe B ? What could be **induced** about A ?

→ Real numbers

The logic of probability

"The true logic for this world is the calculus of Probabilities, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind."

***James Clerk Maxwell** (1850)*

B. Pascal (1623-1662)

P. de Fermat (1601-1665)

A.N. Kolmogorov (1903-1987)

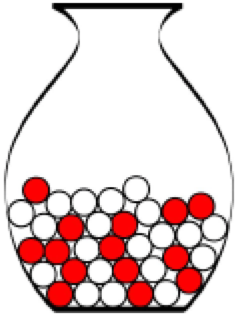
<u>Cox-Jaynes theorem:</u>

1. Divisibility and comparability
2. Common sense
3. Consistency

$\longrightarrow$

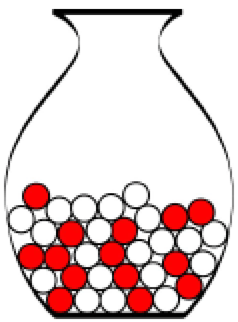
$$p(w) = 0$$
$$p(w) = 1$$

$$p(w) = Nr/N$$



From consequences to causes

Given a bag with twice more white balls than red balls, what is the probability to draw 2 red balls ?

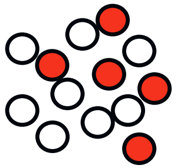


Deductive reasoning

$$P(R=2)?$$

Given that we drawn 5 red and 15 white balls, what is the proportion of red balls in the bag ?

Inductive reasoning



Bayes theorem

Révérend Thomas Bayes

~1748

Pierre Simon Laplace

~1774

prior belief + objective observations = up-dated belief

$$p(C)\; p(F|C) \propto p(C|F)$$

C: causes

F: facts

Joint probability : $p(F,C) = p(F|C)p(C)$

Conditional probability : $p(F|C)$

Marginal probability : $p(F) = \sum_i p(F|C_i)p(C_i)$

$$\boxed{P(C|F) = \frac{P(F|C)P(C)}{P(F)}}$$

Frequentist vs. Bayesian

Frequentist interpretation

- ***Probability*** *= frequency of the occurrence of an event, given an infinite number of trials*

- *Is only defined for random processes that can be observed many times*

- *Is meant to be* ***Objective***

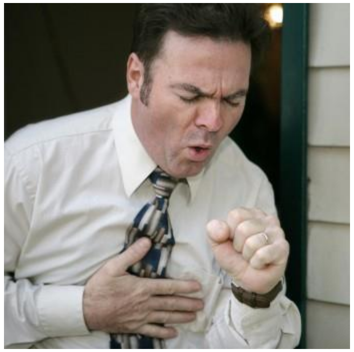
Bayesian interpretation

- ***Probability*** *= degree of belief, measure of uncertainty*

- *Can be arbitrarily defined for any type of event*

- *Is considered as* ***Subjective*** *in essence*

An example of Bayesian reasoning



we observe John coughing (d), and we consider three hypotheses as explanations: John has h_1, a cold; h_2, lung disease; or h_3, heartburn. Intuitively only h_1 seems compelling. Bayes's rule explains why...

MAP

ML

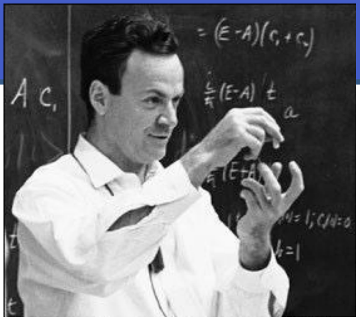
The likelihood favors h_1 and h_2 over h_3: only colds and lung disease cause coughing and thus elevate the probability of the data above baseline. The prior, in contrast, favors h_1 and h_3 over h_2: Colds and heartburn are much more common than lung disease.

Tenenbaum et al., Science, 2011

In the context of neuroimaging

What I cannot create, I do not understand.

Richard Feynman (1918 – 1988)



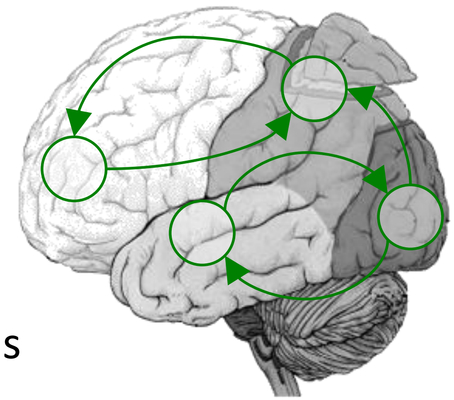
Deductive Reasoning / Predictions / Generative model / Forward problem

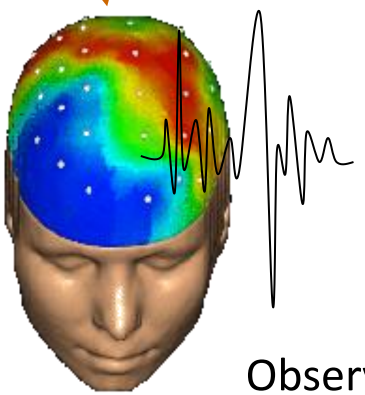
$$P(Y|\theta, M)$$

likelihood

θ

Causes





Y

Observations

posterior distribution

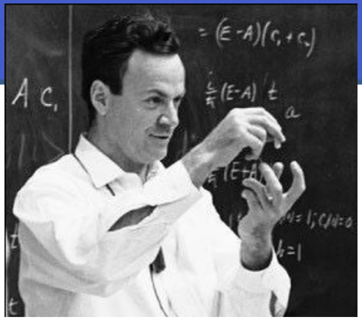
$$P(\theta|Y, M)$$

Inductive Reasoning / Estimations / Inference method / Inverse problem

Bayesian inference

What I cannot create, I do not understand.

Richard Feynman (1918 – 1988)



Model/Hypothesis

Likelihood

Prior

$$P(\theta|Y,M)=\frac{P(Y|\theta,M)P(\theta|M)}{P(Y|M)}$$

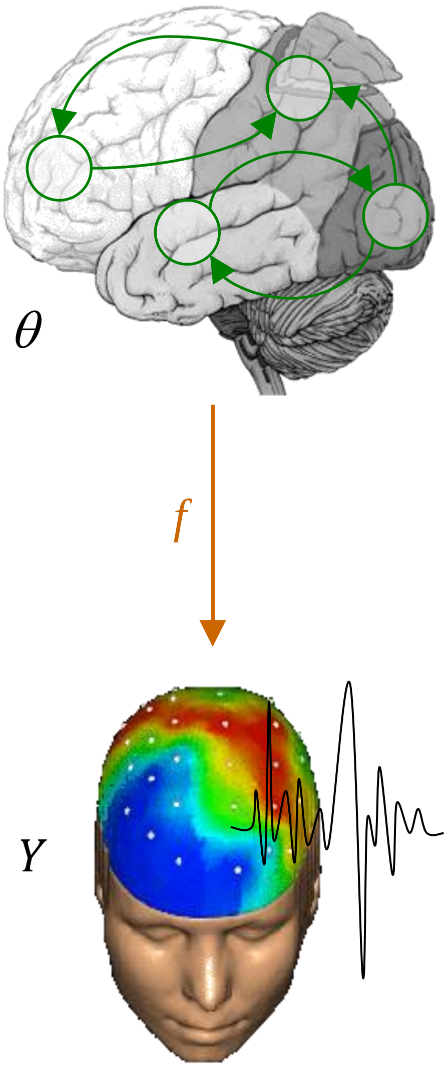
Posterior or conditional

Marginal likelihood or evidence

To be infered

Likelihood function

$$\boxed{P\left(Y \middle| \theta, M\right)}$$

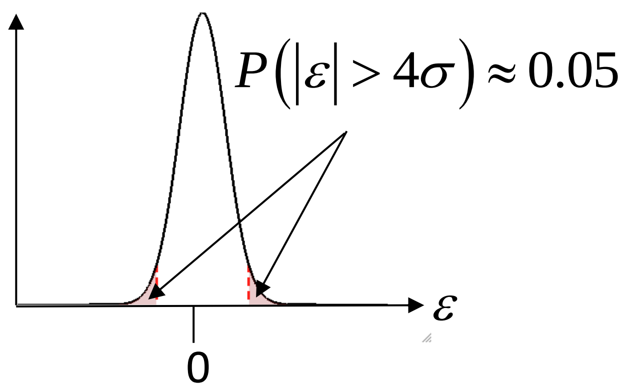


Assumption $Y = f(\theta)$

e.g. linear model $Y = X\theta$

But data are noisy $Y = X\theta + \varepsilon$

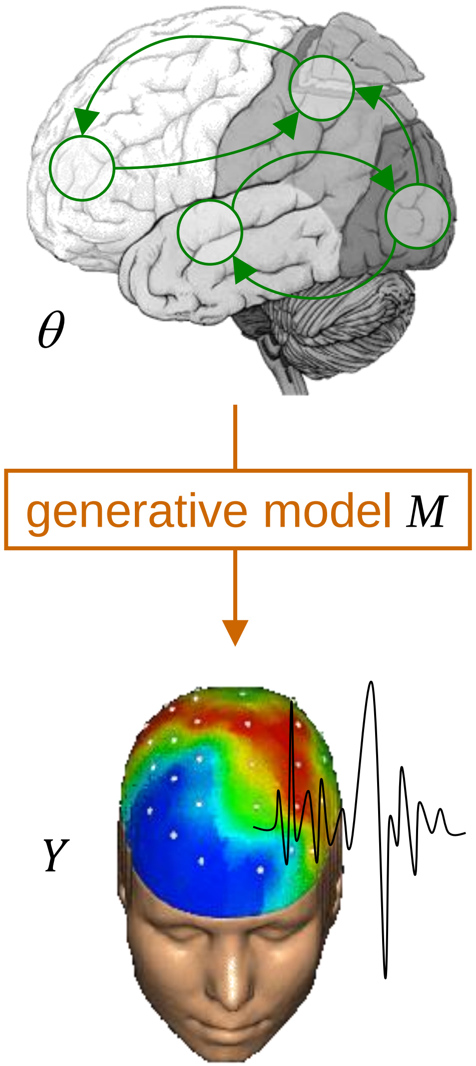
$$p(\varepsilon) \propto \exp\left(-\frac{1}{2\sigma^2}\varepsilon^2\right)$$



Data distribution, *given the parameter value*:

$$p\left(y \middle| \theta\right) \propto \exp\left(-\frac{1}{2\sigma^2}\left(y - f(\theta)\right)^2\right)$$

Incorporating priors



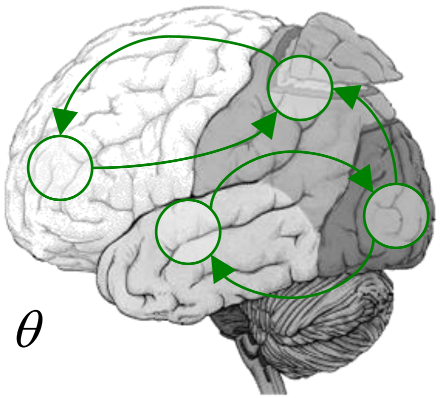
$$\boxed{P(\theta|M)}$$

Likelihood $Y = X\theta + \varepsilon \qquad \varepsilon \sim N(0,\gamma)$

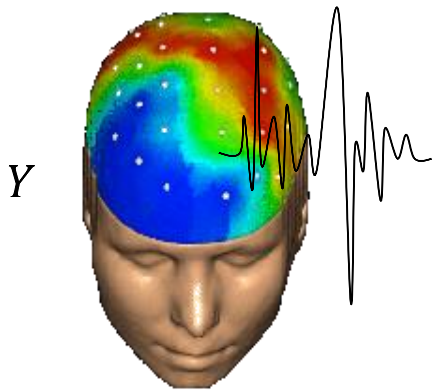
Prior $\theta \sim N(\mu,\sigma)$

Incorporating priors

$\boxed{P(\theta|M)}$

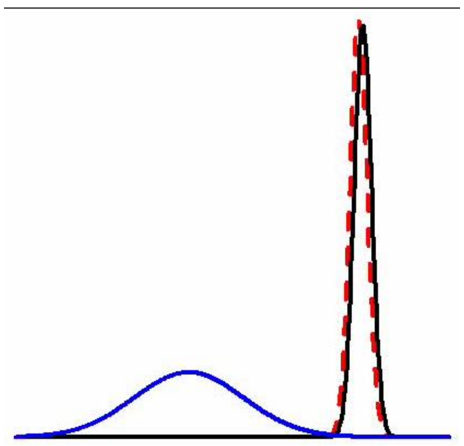


generative model M



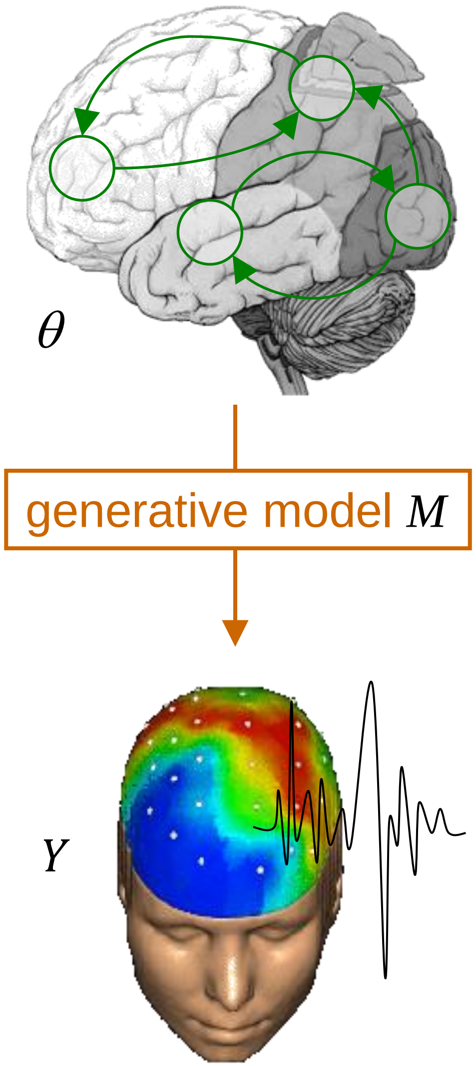
Likelihood $Y = X\theta + \varepsilon$ $\quad \varepsilon \sim N(0,\gamma)$

Prior $\theta \sim N(\mu,\sigma)$



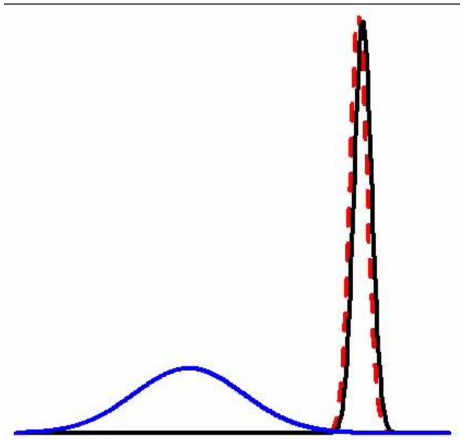
Incorporating priors

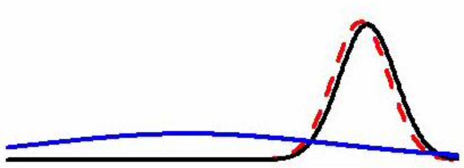
$$\boxed{P(\theta|M)}$$



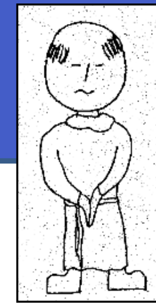
Likelihood $Y = X\theta + \varepsilon \qquad \varepsilon \sim N(0,\gamma)$

Prior $\theta \sim N(\mu,\sigma)$





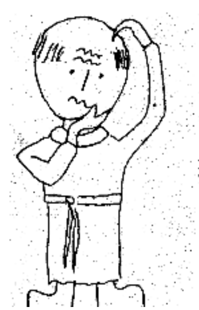
About priors

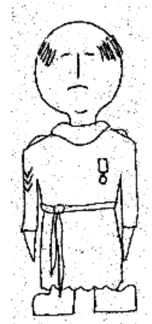


« Ir-reverend Bayes »

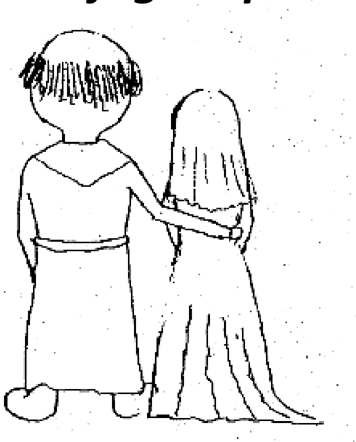
Shrinkage prior $\theta \sim N(0, \sigma)$

Uninformative (objective) prior $\theta \sim N(0, \sigma)$ with large σ



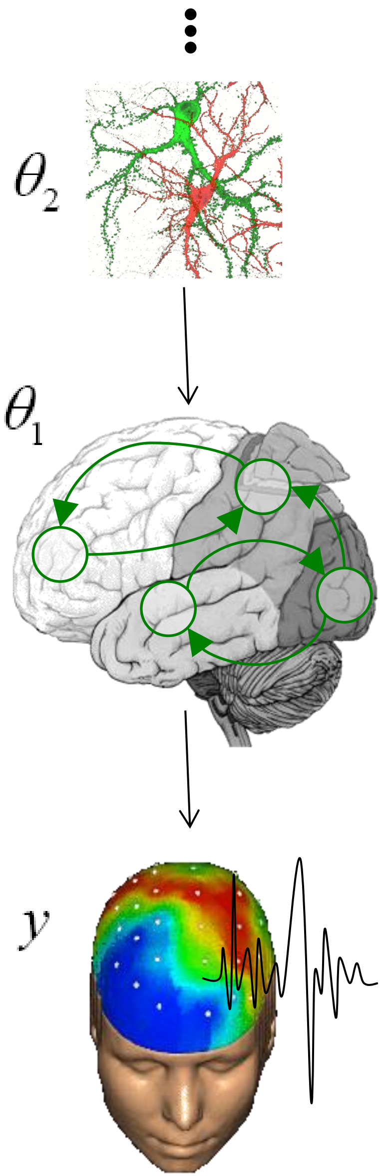


Conjugate prior when the prior and posterior distributions belong to the same family



Likelihood dist.	Conjugate prior dist.
Binomiale	Beta
Multinomiale	Dirichlet
Gaussian	Gaussian
Gamma	Gamma

Hierarchical models and empirical priors



Likelihood $Y = X\theta_1 + \varepsilon \qquad \varepsilon \sim N(0, \gamma)$

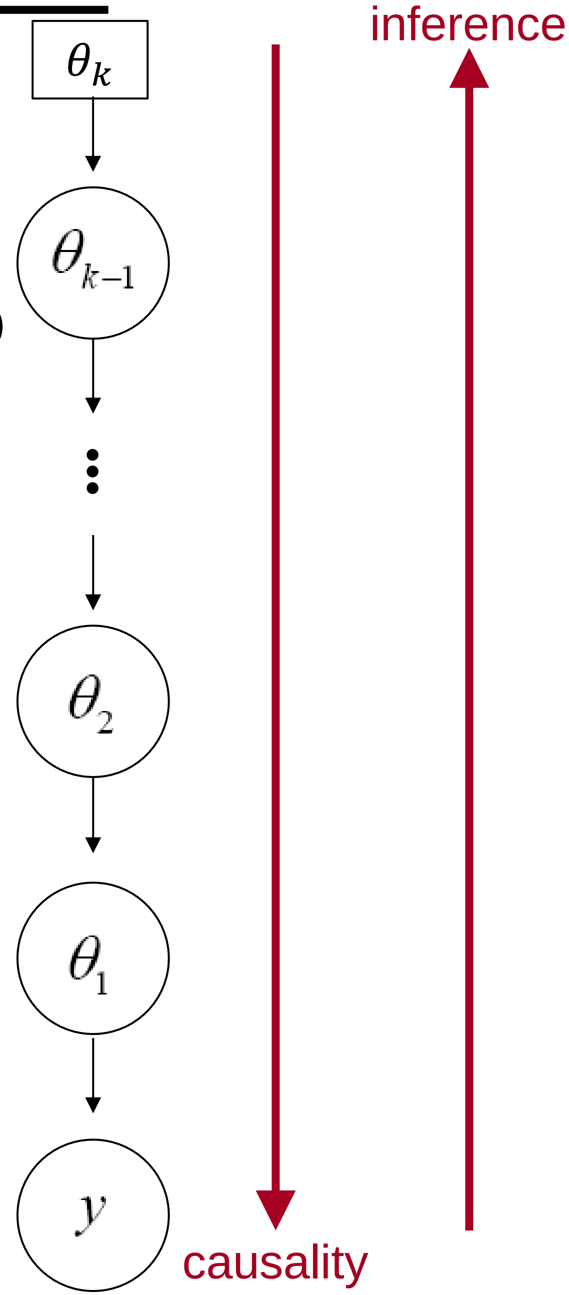
Prior $\theta = \{\theta_1, \theta_2, \ldots, \theta_{k-1}\}$

$$\theta_1 \sim N(\theta_2, \sigma_2)$$

$$\theta_2 \sim N(\theta_3, \sigma_3)$$

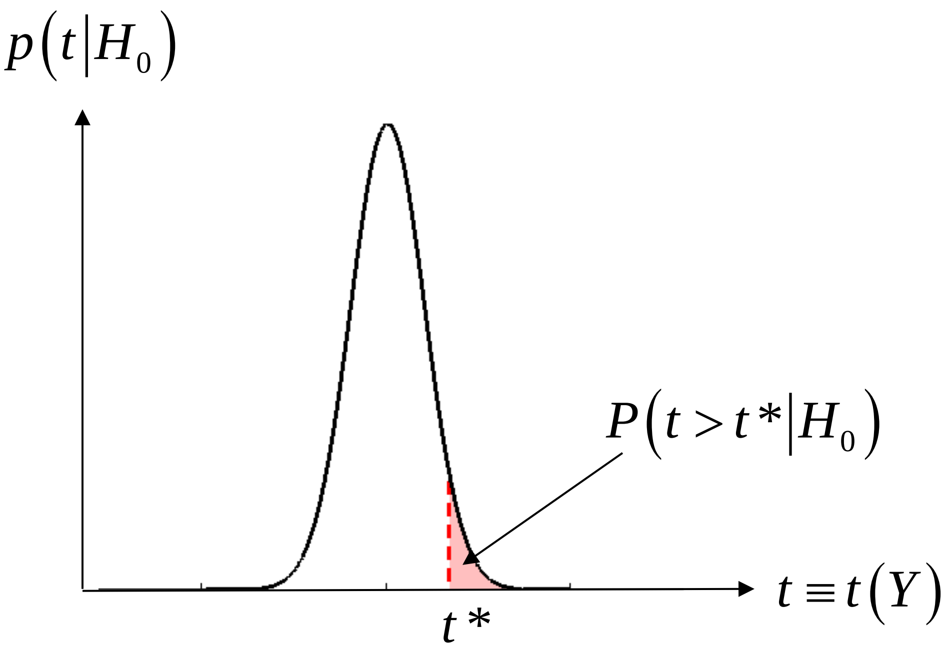
$$\vdots$$

$$\theta_{k-1} \sim N(\theta_k, \sigma_k)$$



Hypothesis testing : classical vs. bayesian

- define the null, e.g.: $H_0 : \theta = 0$

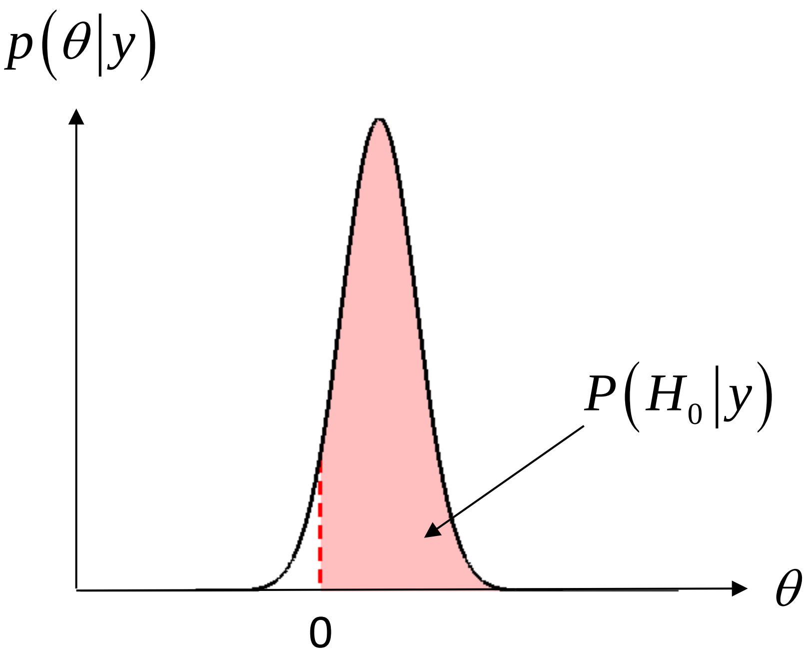


- estimate parameters (obtain test stat.)
- apply decision rule, i.e.:

if $P(t > t^* | H_0) \leq \alpha$ then reject H0

classical SPM

- invert model (obtain posterior pdf)



- define the null, e.g.: $H_0 : \theta > 0$
- apply decision rule, i.e.:

if $P(H_0|y) \geq \alpha$ then accept H0

Bayesian PPM

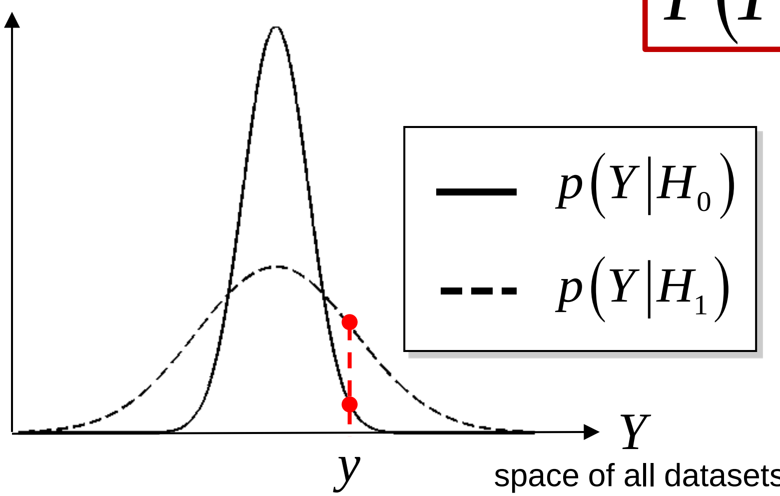
Hypothesis testing : Bayesian model comparison

- define the null and the alternative hypothesis *in terms of priors*, e.g.:

$$\boxed{P(Y|M)}$$

$$H_0 : p(\theta|H_0) = \begin{cases} 1 & \text{if } \theta = 0 \\ 0 & \text{otherwise} \end{cases}$$

$$H_1 : p(\theta|H_1) = N(0, \Sigma)$$



- apply decision rule, i.e.: if $\frac{P(y|H_0)}{P(y|H_1)} < u$ then reject H0

Inference on models

if $$P(Y|M_1) > P(Y|M_2)$$, select model M_1

In practice, compute the Bayes Factor…

$$BF_{12} = \frac{P(Y|M_1)}{P(Y|M_2)}$$

… and apply the decision rule

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

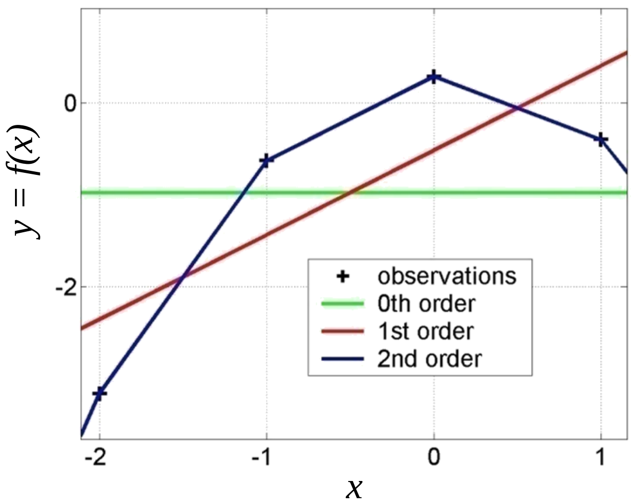
Bayes factors can be interpreted as follows. Given candidate hypotheses i and j, a Bayes factor of 20 corresponds to a belief of 95% in the statement ‘hypothesis i is true’. This corresponds to strong evidence in favor of i.

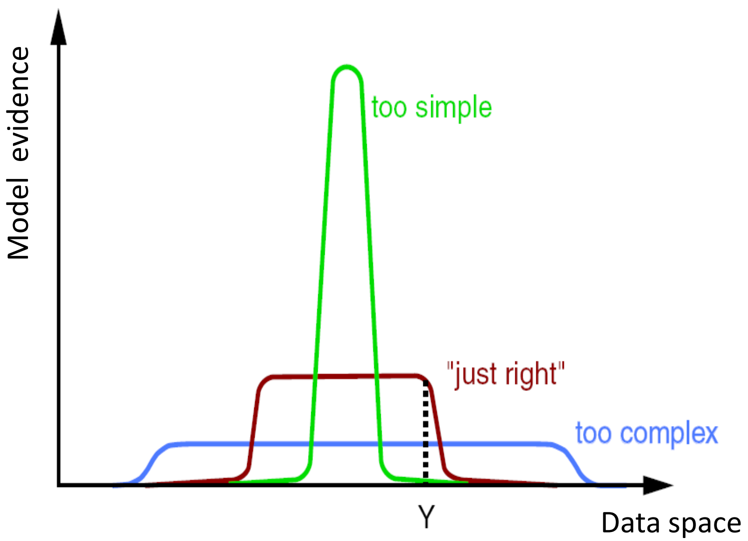
Kass and Raftery, JASA, 1995.

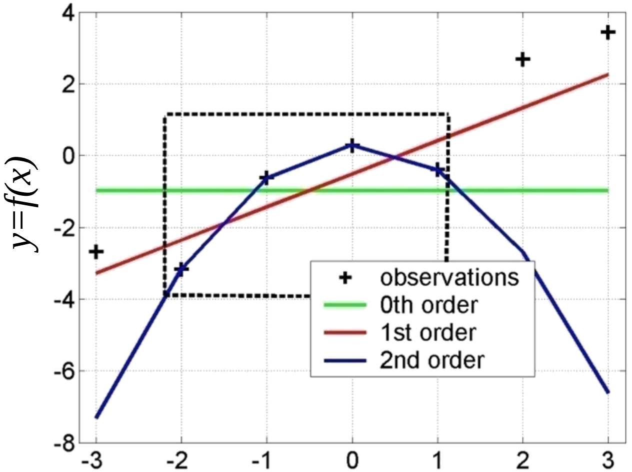
Hypothesis testing and principle of parsimony

Occam's razor

Complex models should not be considered without necessity







$$\boxed{p(Y \mid M) = \int p(Y \mid \theta, M)\, p(\theta \mid M)\, d\theta}$$

Usually no exact analytic solution !!

Approximations to the model evidence

$$\Delta BIC = -2\log\left[\frac{\sup P(Y|\theta, M_1)}{\sup P(Y|\theta, M_2)}\right] - (n2 - n1)\log N$$

$$\Delta AIC = -2\log\left[\frac{\sup P(Y|\theta, M_1)}{\sup P(Y|\theta, M_2)}\right] - 2(n2 - n1)$$

$\boxed{\text{Free energy } \mathbf{F}}$ ← Obtained from Variational Bayes inference

For non-linear models, F is used as a proxy for the model evidence

Variational Bayes Inference

Variational Bayes (VB) ≡ Expectation Maximization (EM) ≡ Restricted Maximum Likelihood (ReML)

Main features

- Iterative optimization procedure
- Yields a twofold inference on parameters θ and models M
- Uses a fixed-form approximate posterior $q(\theta)$
- Make use of approximations (e.g. mean field, Laplace)

to approach $P(\theta|Y,M)$ and $P(Y|M)$

The criterion to be maximized is the free-energy F

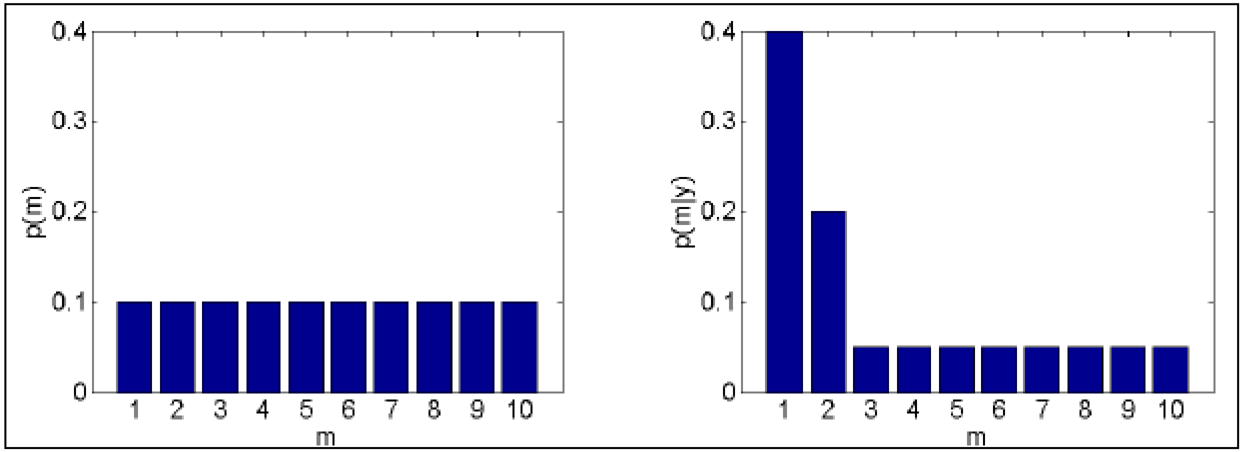
F is a lower bound to the log-evidence

$$\boldsymbol{F} = \ln P(Y|M) - D_{KL}\big(Q(\theta); P(\theta|Y,M)\big)$$
$$= \langle \ln P(Y|\theta,M)\rangle_Q - D_{KL}\big(Q(\theta); P(\theta|M)\big)$$

F = accuracy - complexity

Bayes rule for models

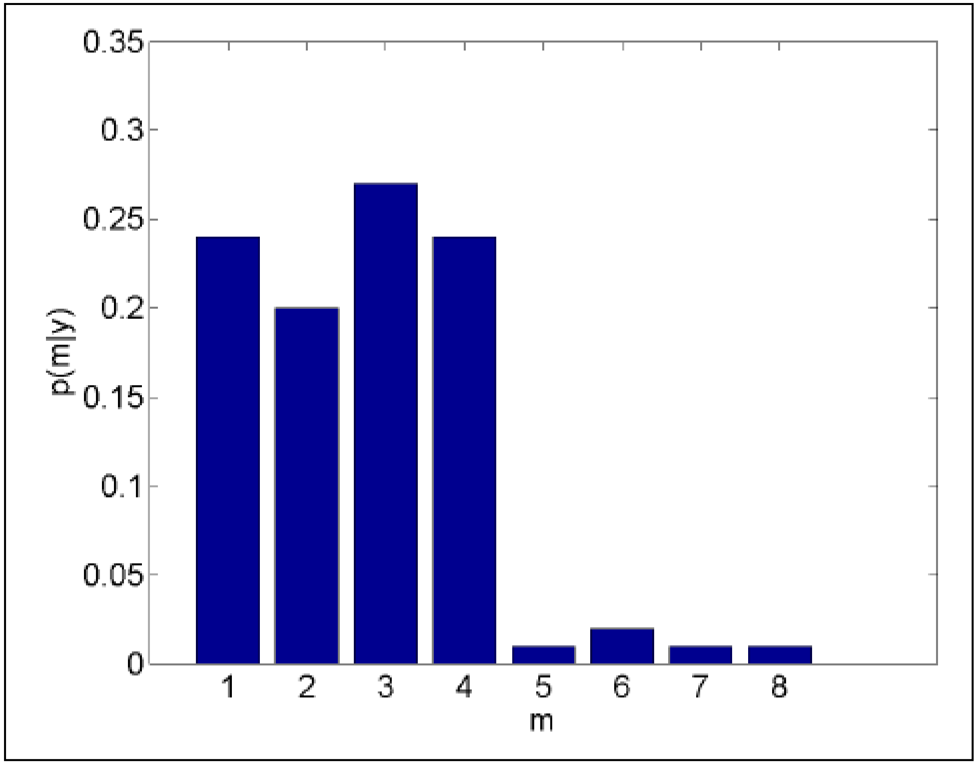
$$p(m|y) = \frac{p(y|m)p(m)}{p(y)}$$



For non-linear models, F is used as a proxy for the model evidence

Family level inference

Example of model posterior (8 models)

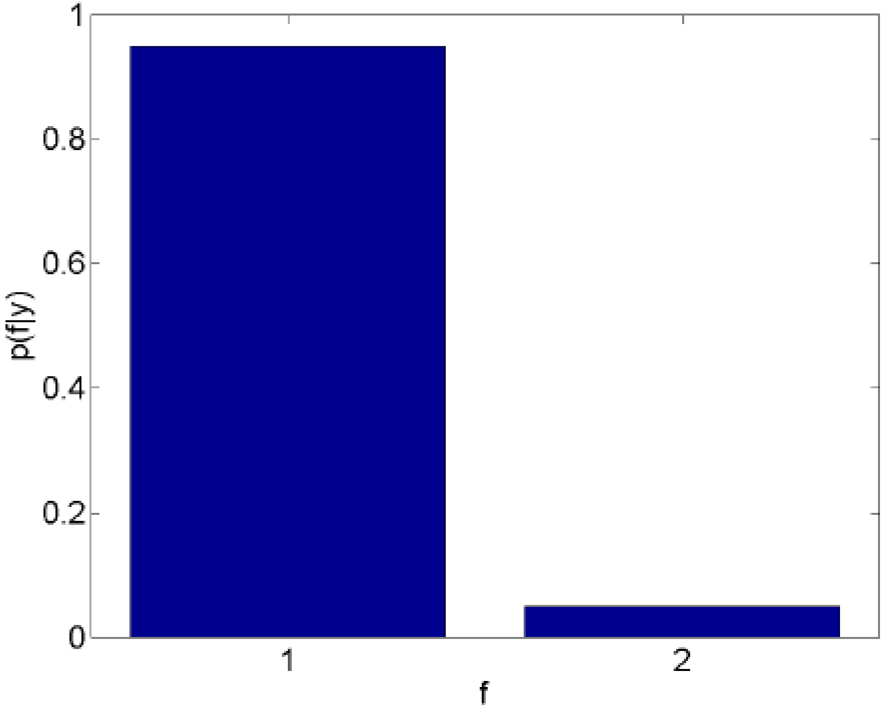


Similar models share probability mass (dilution).
The probability for any single model can become very small, especially for large model spaces.

Family level inference

If we can assign each model m to a family f, one can compare families based on their posterior probabilities which write simply

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



Within family parameter estimation : Bayesian model averaging

Each DCM.mat file stores the posterior mean (DCM.Ep) and posterior covariance (DCM.Cp) for that particular model M. They define the posterior distribution over parameters

$$P(\theta|Y,M)$$

The posterior can be combined with the posterior model probabilities to compute a posterior over parameters independent of model assumptions (within the chosen set)

$$P(M|Y)$$

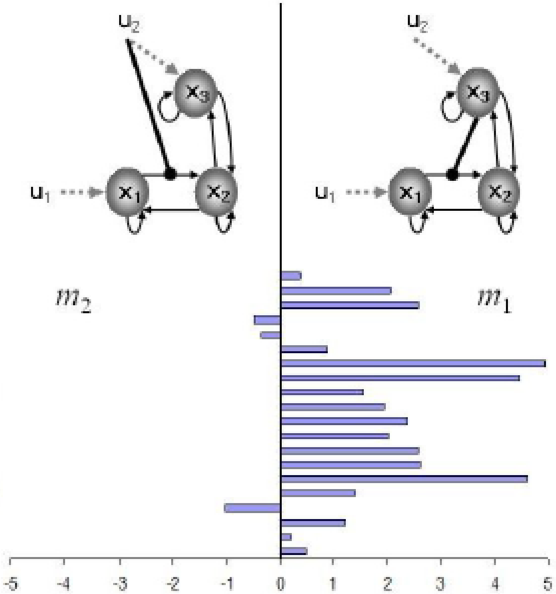
$$\begin{aligned} p(\theta|y) &= \sum_m p(\theta, m|y) \\ &= \sum_m p(\theta|m, y)p(m|y) \end{aligned}$$

We marginalized over model space (usually restricted to the winning family)

Group model comparison : fixed effect (FFX)

Two models, twenty subjects.

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

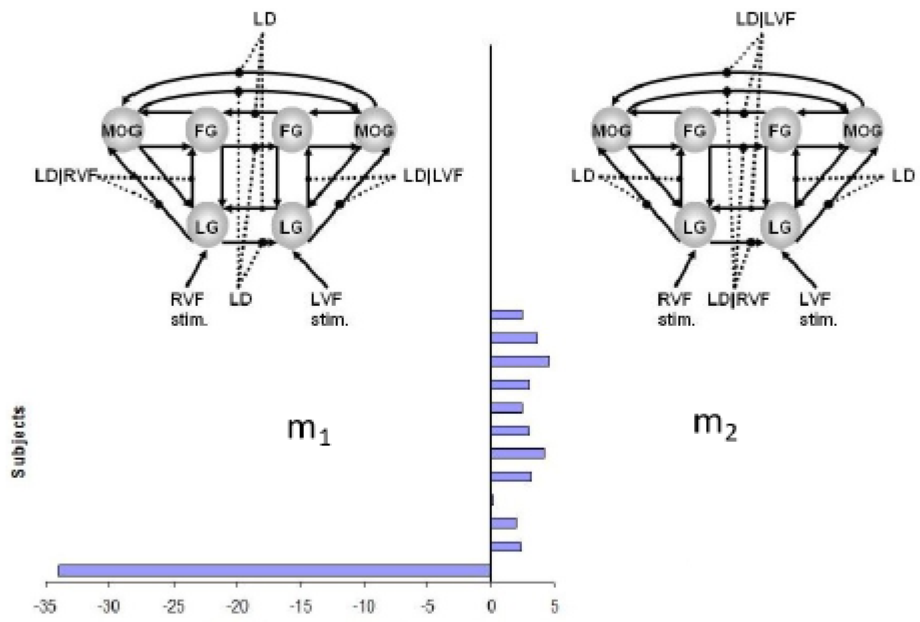


The Group Bayes Factor (GBF) is

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

Group model comparison : random effect (FFX)

Stephan et al. J. Neurosci, 2007

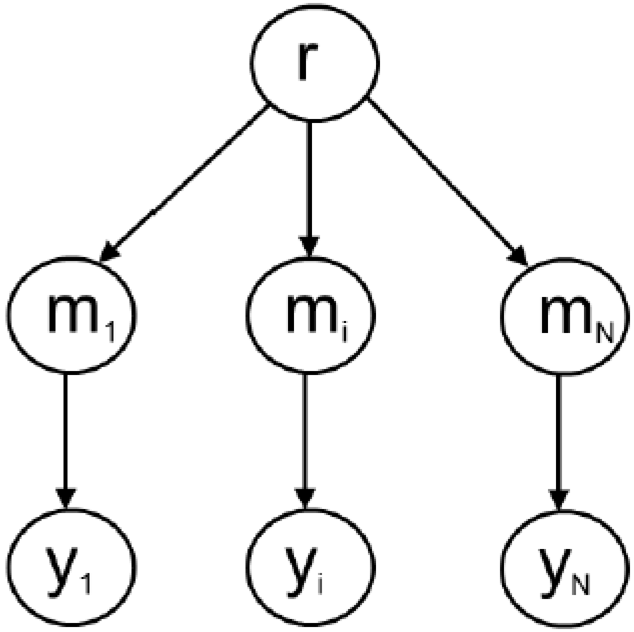


11/12=92% subjects favour model 2.

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

Group model comparison : random effect (FFX)

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

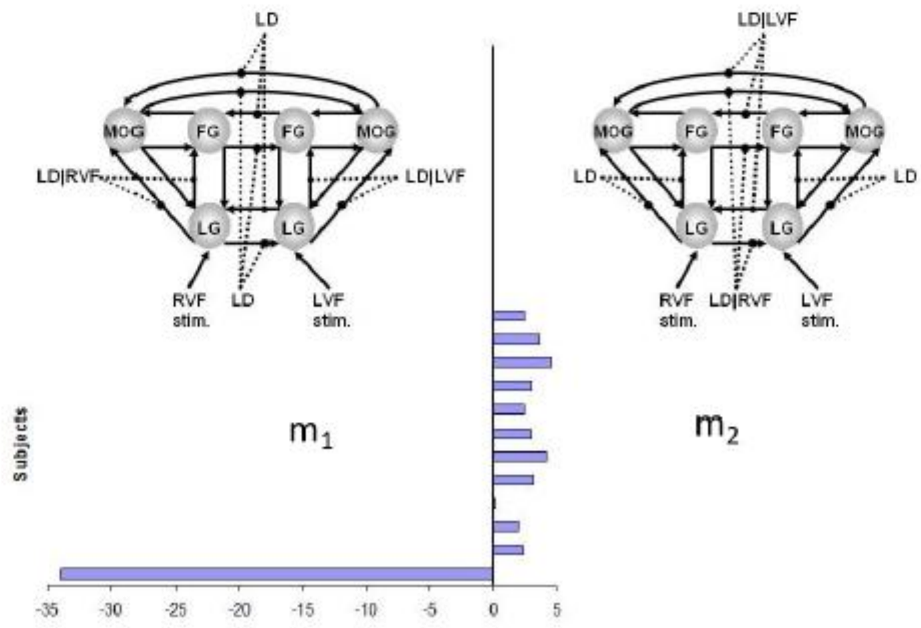


Approximate posterior

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

Group model comparison : random effect (FFX)

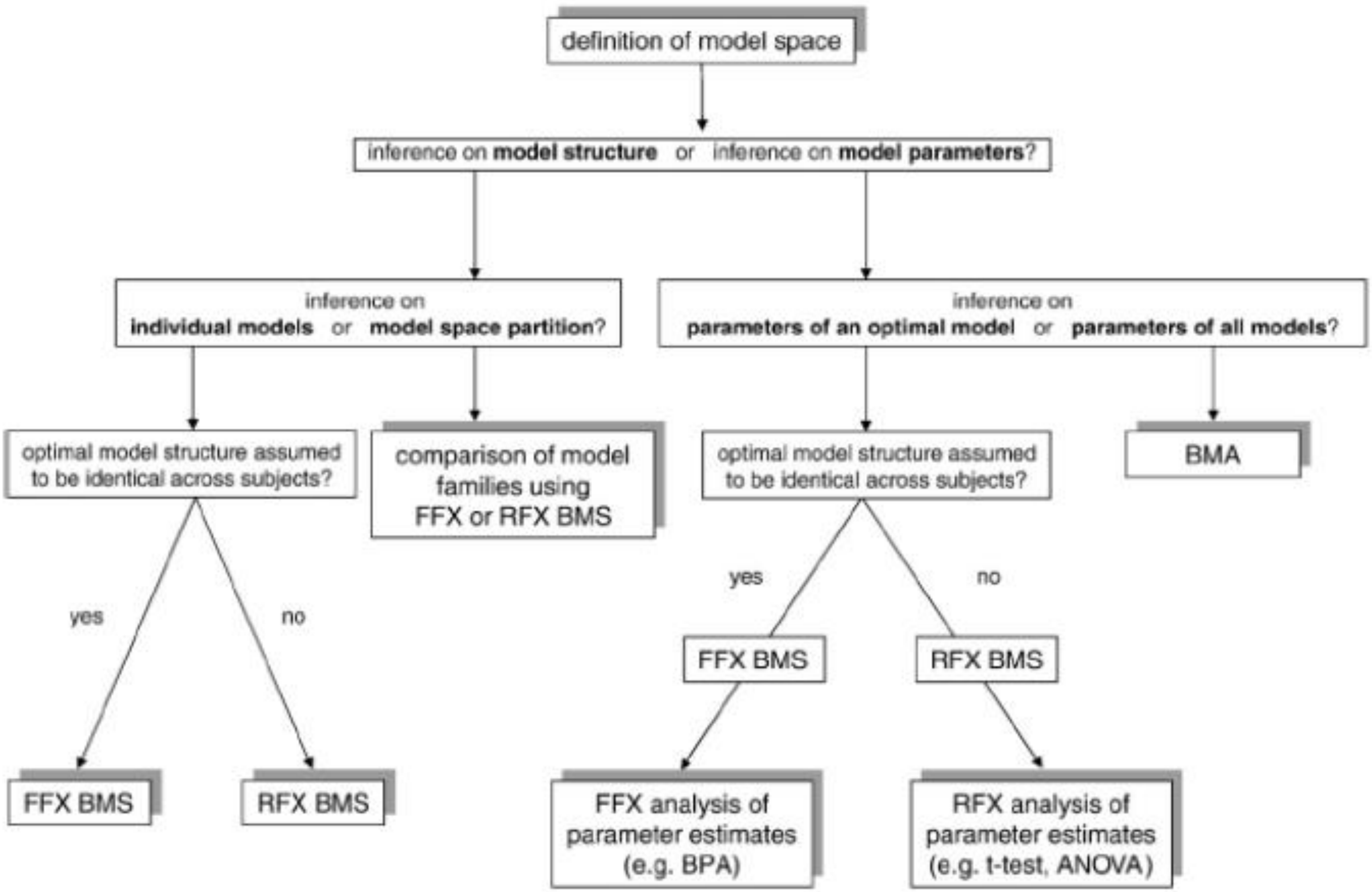
11/12=92% subjects favoured model 2.



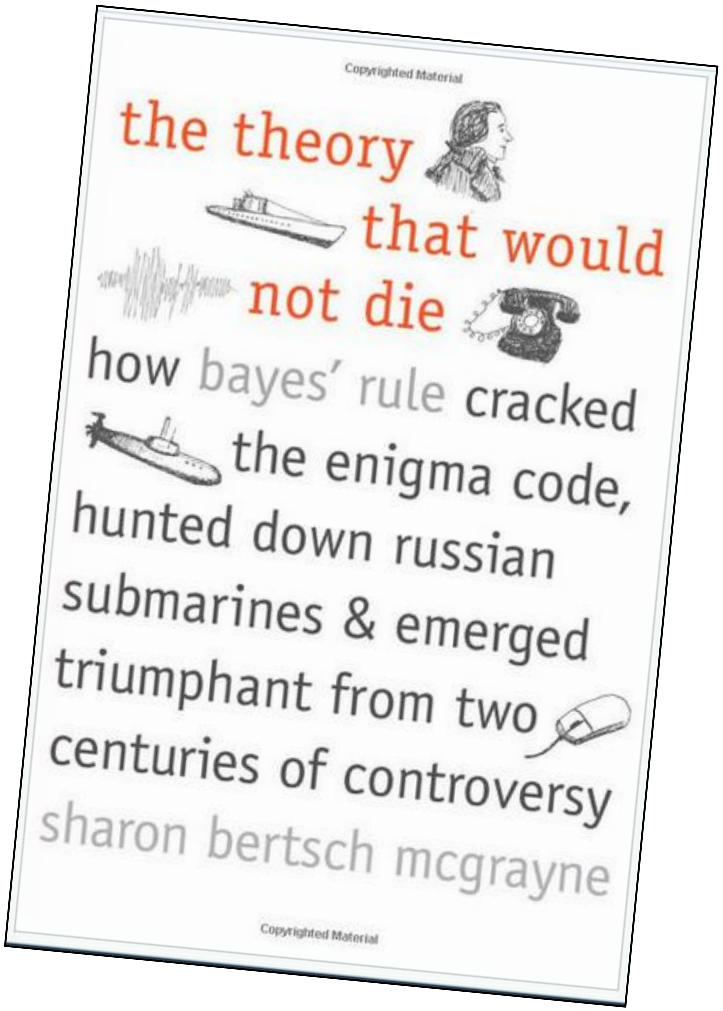
$$E[r_2|Y] = 0.84$$
$$p(r_2 > r_1|Y) = 0.99$$

where the latter is called the exceedance probability.

Overview


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)

Suggestion for further reading



"When the facts change, I change my mind, what do you do, sir ?"

John Maynard Keynes

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