

Bayesian Inference

“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)

Jérémie Mattout

Lyon Neuroscience Research Center, France

With many thanks to

Jean Daunizeau

Guillaume Flandin

Karl Friston

Will Penny

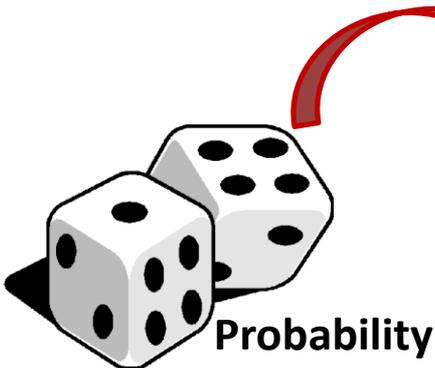
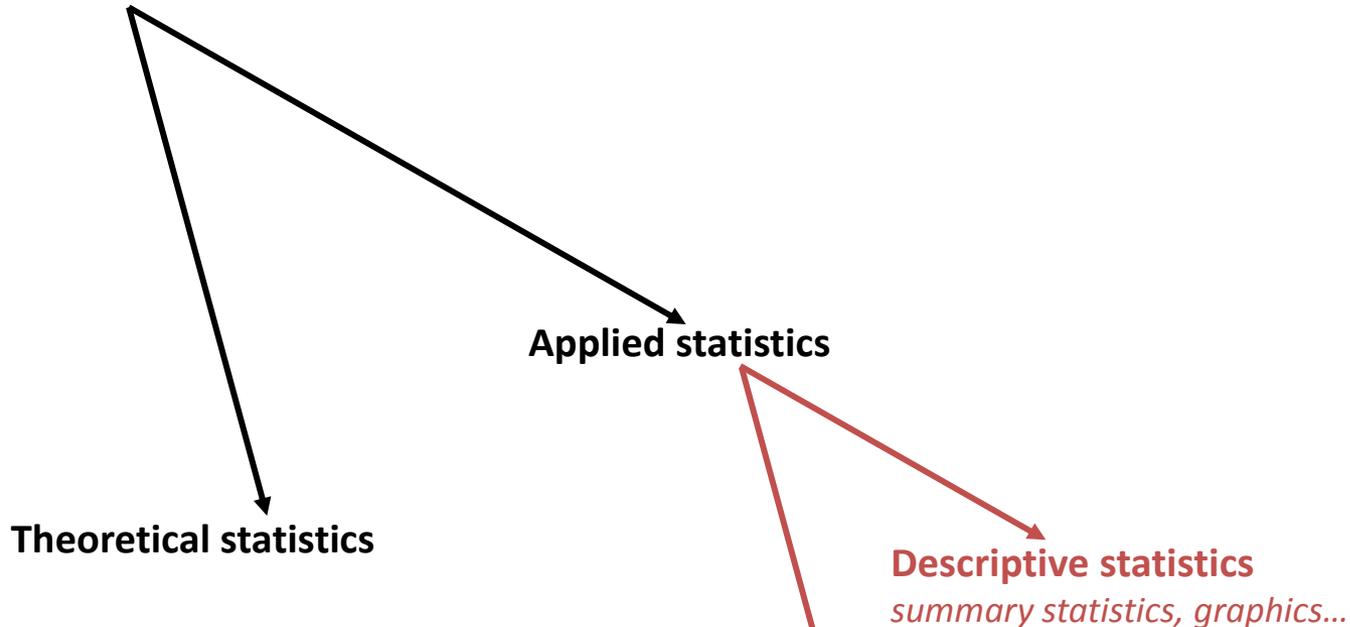
Outline

- General principles
- The Bayesian way
- SPM examples

- **General principles**
- The Bayesian way
- SPM examples

A starting point

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



Inferential statistics

*Data interpretations, decision making
(Modeling, accounting for randomness and uncertainty, hypothesis testing, inferring hidden parameters)*

The notion(s) of probability



To express belief that an event has or will occur

Ω : All possible events

A_i : one particular event



B. Pascal (1623-1662)



P. de Fermat (1601-1665)

Kolmogorov axioms

(1) $0 \leq P(A) \leq 1$

(2) $P(\Omega) = 1$

(3) $P(A_1 \cup A_2 \cdots \cup A_k) = \sum_{i=1}^k P(A_i)$
(for mutually exclusive events)



A.N. Kolmogorov (1903-1987)

A few consequences...

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

(joint probability)

$$P(A \cap B) = 0$$

(if mutually exclusive events)

$$P(A \cap B) = P(A) \cdot P(B)$$

(if independent events)

The notion(s) of probability

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



The notion(s) of probability

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



Joint and conditional probabilities

- *Joint probability of A and B* $P(A \cap B) = P(A, B)$
- *Conditional probability of A given B* $P(A|B)$

$$P(A, B) = P(A|B)P(B)$$

- *Note that if A and B are independent*

$$P(A|B) = P(A)$$

and

$$P(A, B) = P(A)P(B)$$

Joint and conditional probabilities

- Joint probability of A and B $P(A \cap B) = P(A, B)$
- Conditional probability of A given B $P(A|B)$

$$P(A, B) = P(A|B)P(B)$$

$$P(A, B) = P(B, A) = P(B|A)P(A)$$

$$P(A|B)P(B) = P(B|A)P(A)$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$



T. Bayes (1702-1761)

Extension to multiple variables

$$\begin{aligned} P(A, B, C) &= P(A, B|C)P(C) = P(A|B, C)P(B|C)P(C) \\ &= P(B|A, C)P(A|C)P(C) \end{aligned}$$

$$P(A|B, C) = \frac{P(B|A, C)P(A|C)}{P(B|C)}$$



T. Bayes (1702-1761)

Marginalisation

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- Discrete case

$$P(B) = \sum_A P(A, B) = \sum_A P(B|A)P(A)$$

- Continuous case

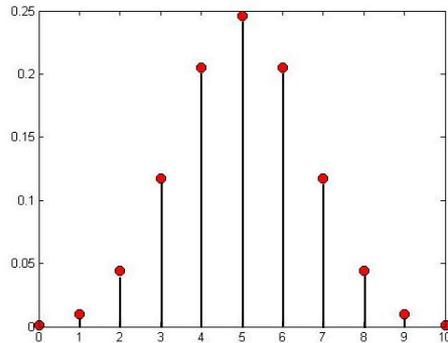
$$P(B) = \int P(A, B)dA = \int P(B|A)P(A)dA$$

Probability distributions (quick reminder)

Discrete variable
(e.g. Binomial distribution)



$$P(\text{Heads}) = 1 - P(\text{Tails})$$



Number of Heads in 10 trials

$$p(X = x) = C_x^n p^x (1-p)^{n-x}$$

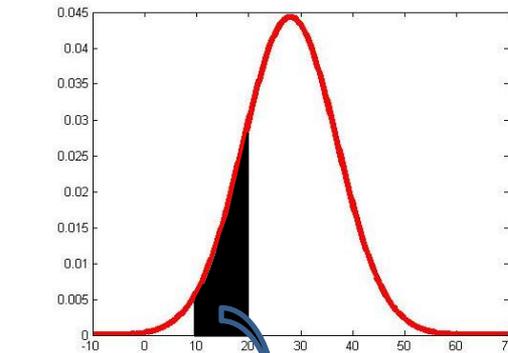
$$p(X \leq x) = \sum_0^x f(x)$$

Continuous variable
(e.g. Gaussian distribution)



$$p(X) \sim N(\mu, \sigma)$$

$$p(X = x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



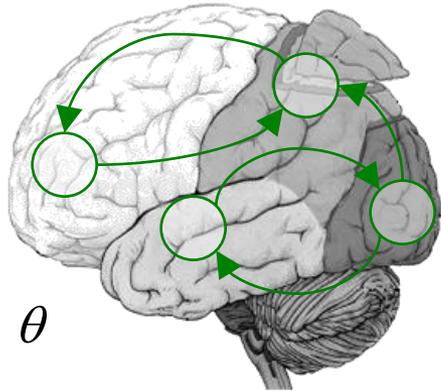
Temperature X

$$p(10 \leq X \leq 20) = \int_{x=10}^{20} f(x) dx$$

Likelihood function

Assumption $Y = f(\theta)$

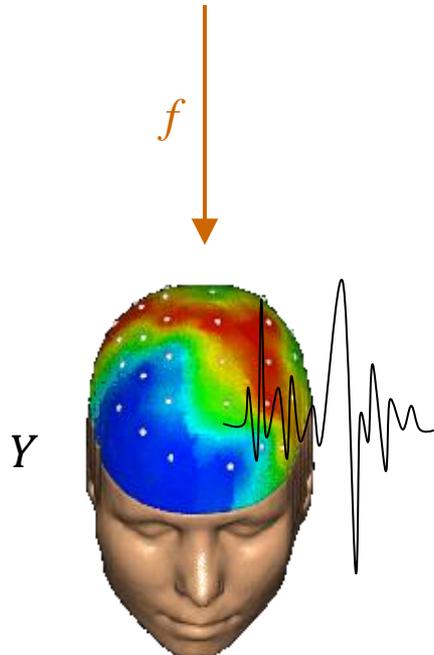
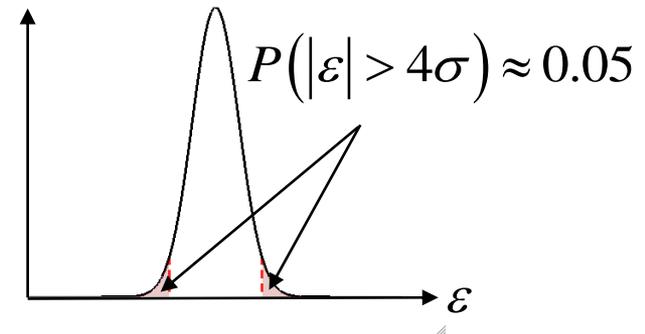
e.g. linear model $Y = X\theta$



But data are noisy

$Y = X\theta + \varepsilon$

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



Distribution of data, given fixed parameters:

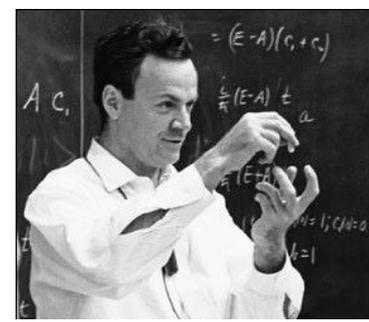
$$p(y|\theta) \propto \exp\left(-\frac{1}{2\sigma^2} (y - f(\theta))^2\right)$$

- General principles
- **The Bayesian way**
- SPM examples

A word on generative models

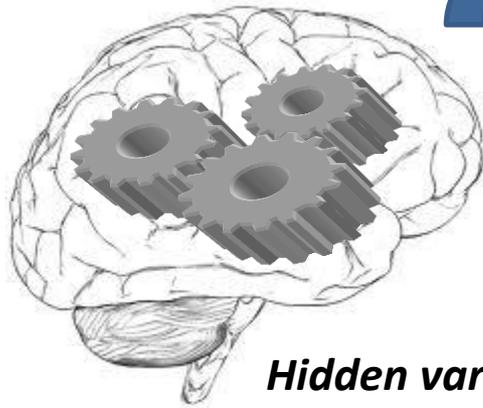
What I cannot create, I do not understand.

Richard Feynman (1918 – 1988)

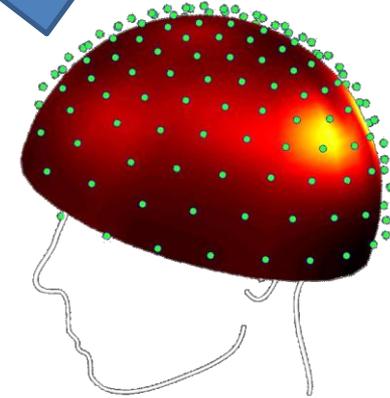


Model: mathematical formulation of a system or process (set of hypothesis and approximations)

Data generative process ?



Hidden variables (θ)



Observations (Y)

A Probabilistic Model enables to:

- **Account for prior knowledge and uncertainty**
(due to randomness, noise, incomplete observations)
- Simulate data
- Make predictions
- **Estimate hidden parameters**
- **Test Hypothesis**

Another look at Bayes rule

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

Model/Hypothesis

Likelihood $P(Y|\theta, M)$ **Prior** $P(\theta|M)$

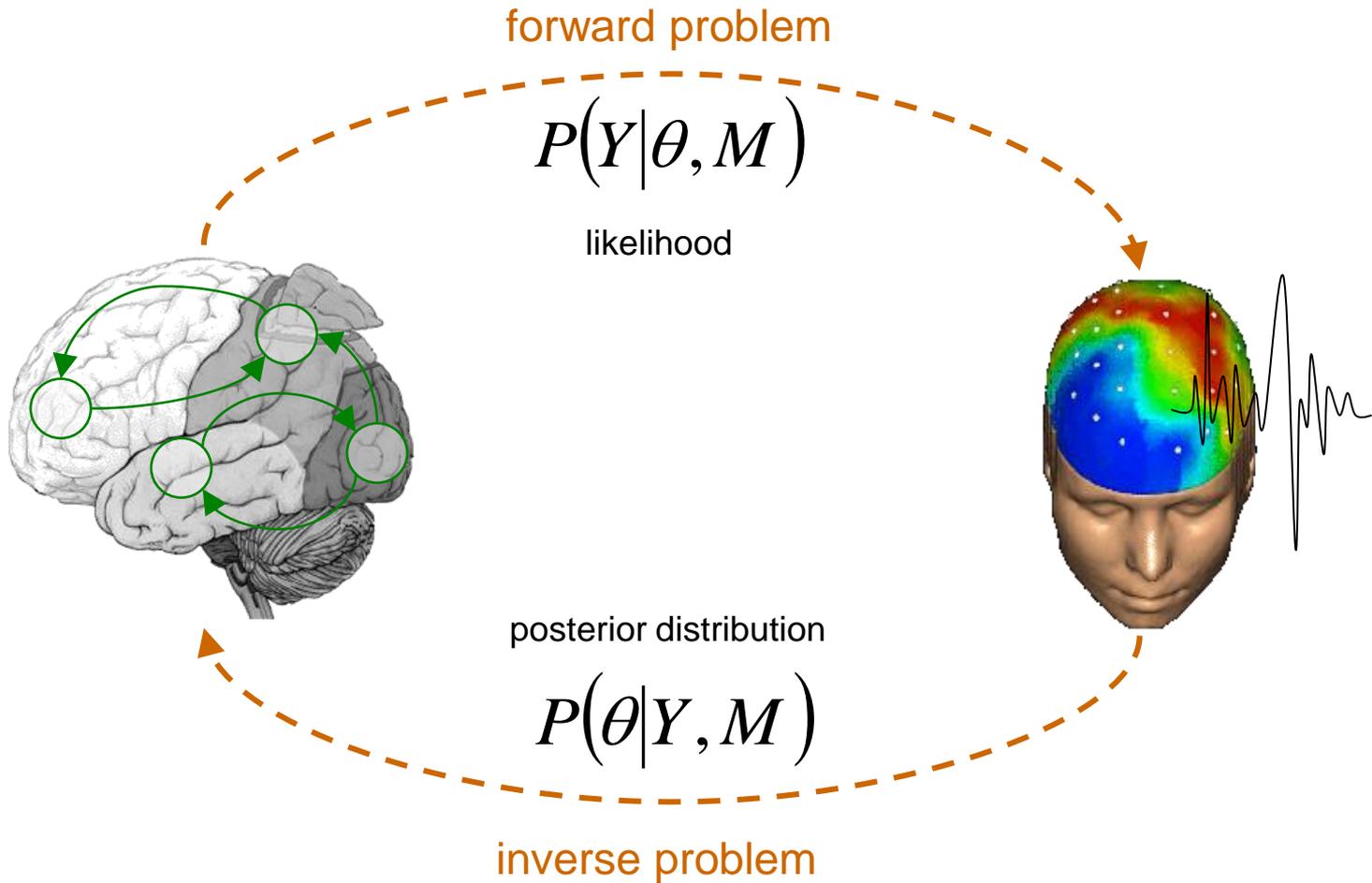
Posterior or conditional $P(\theta|Y, M)$

Marginal likelihood or evidence $P(Y|M)$

To be inferred

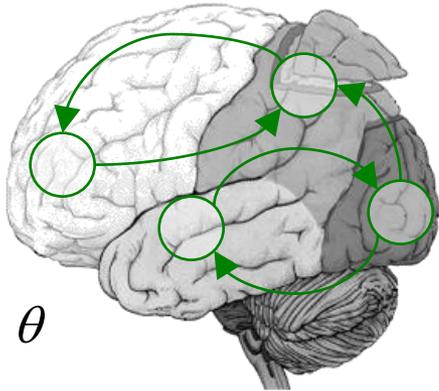
The diagram illustrates Bayes' rule with several components highlighted in colored boxes and labeled. The numerator of the fraction, $P(Y|\theta, M)P(\theta|M)$, is enclosed in a light green box labeled "Model/Hypothesis" above it. Within this green box, $P(Y|\theta, M)$ is labeled "Likelihood" and $P(\theta|M)$ is labeled "Prior". The denominator, $P(Y|M)$, is enclosed in a light blue box labeled "Marginal likelihood or evidence" below it. The entire fraction is enclosed in a larger light blue box labeled "Posterior or conditional" below it. The text "To be inferred" is positioned below the "Posterior or conditional" label.

Another look at Bayes rule

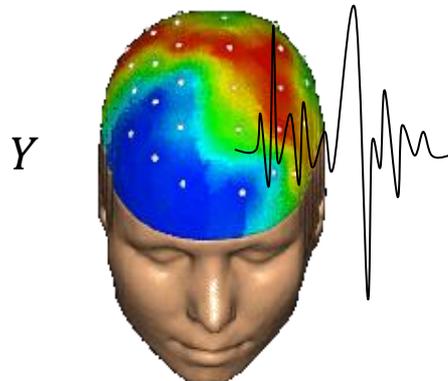


A simple example

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

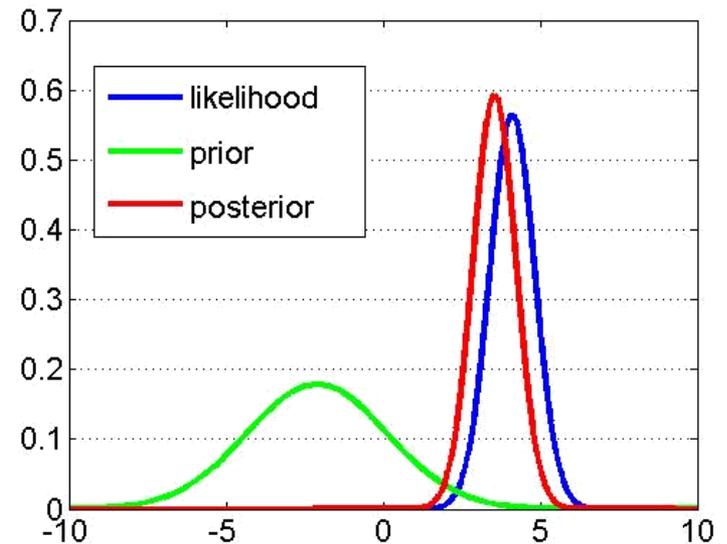


generative model M



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

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



Qualifying priors

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

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.

Binomiale

Multinomiale

Gaussian

Gamma

Conjugate prior dist.

Beta

Dirichlet

Gaussian

Gamma

Hierarchical models and empirical priors

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

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

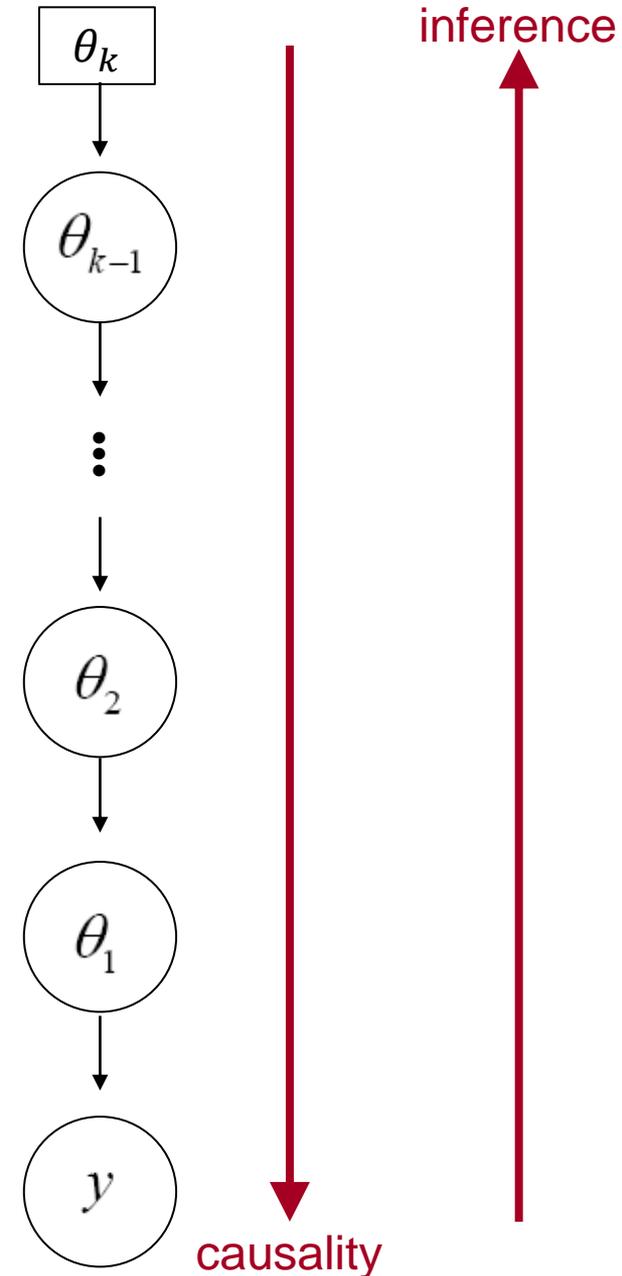
$$\theta_1 \sim N(\theta_2, \sigma_2)$$

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

\vdots

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

Graphical representation



Model comparison

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

Bayes rule again...

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

And with no prior in favor of one particular model...

$$P(M|Y) \propto P(Y|M)$$

Model comparison

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

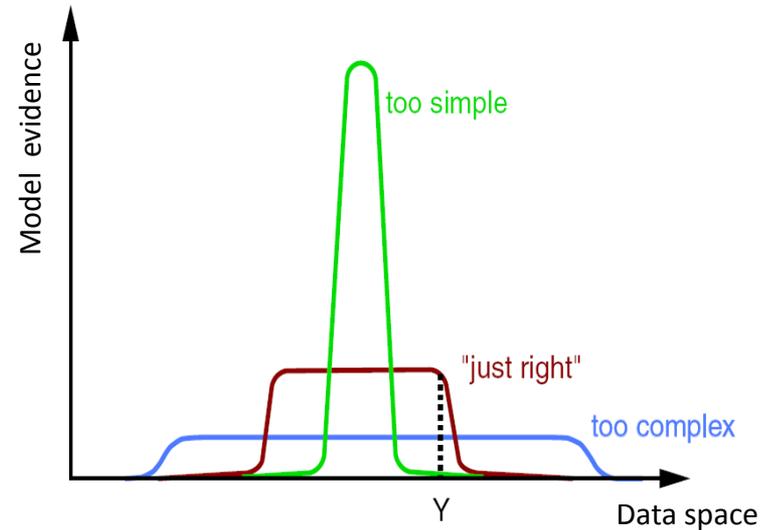
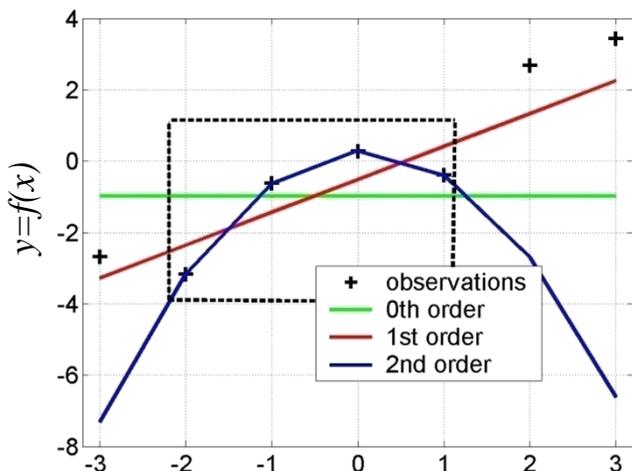
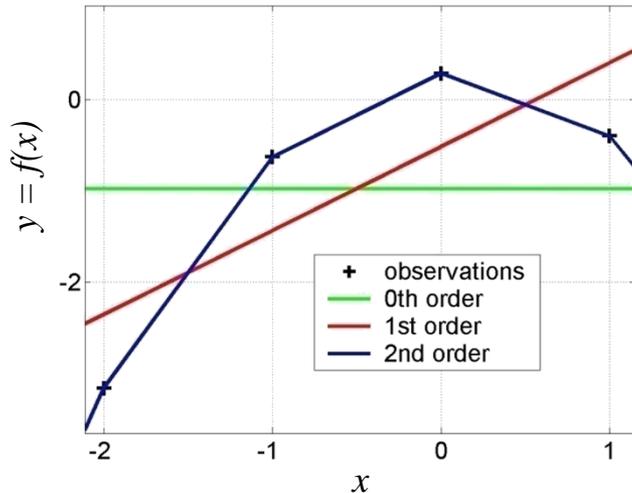
B_{12}	Evidence
1 to 3	Weak
3 to 20	Positive
20 to 150	Strong
≥ 150	Very strong

Principle of parsimony

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

Occam's razor

Complex models should not be considered without necessity



$$p(Y | M) = \int p(Y | \theta, M) p(\theta | M) d\theta$$



Usually no exact analytic solution !!

Approximations to the (log-)evidence

$$\Delta BIC = -2 \log \left[\frac{\sup P(Y|\theta, M_1)}{\sup P(Y|\theta, M_2)} \right] - (n_2 - n_1) \log N$$

$$\Delta AIC = -2 \log \left[\frac{\sup P(Y|\theta, M_1)}{\sup P(Y|\theta, M_2)} \right] - 2(n_2 - n_1)$$

Free energy **F**



Obtained from the Variational Bayes inference

Variational Bayes Inference

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

Main features

- Yields a twofold inference on parameters θ and models M
- Iterative optimization procedure
- 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

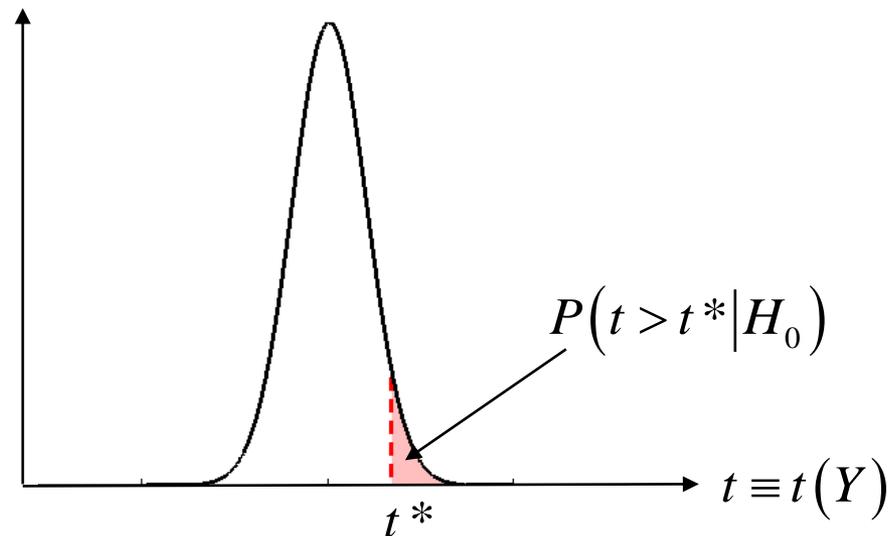
$$\begin{aligned} \mathbf{F} &= \ln P(Y|M) - D_{KL}(Q(\theta); P(\theta|Y, M)) \\ &= \langle \ln P(Y|\theta, M) \rangle_Q - D_{KL}(Q(\theta); P(\theta|M)) \end{aligned}$$

F is a lower bound to the log-evidence

F = accuracy - complexity

Hypothesis testing (classical way)

- given a null hypothesis, e.g.: $H_0 : \theta = 0$



- apply decision rule, i.e.:
if $P(t > t^* | H_0) \leq \alpha$ then reject H_0

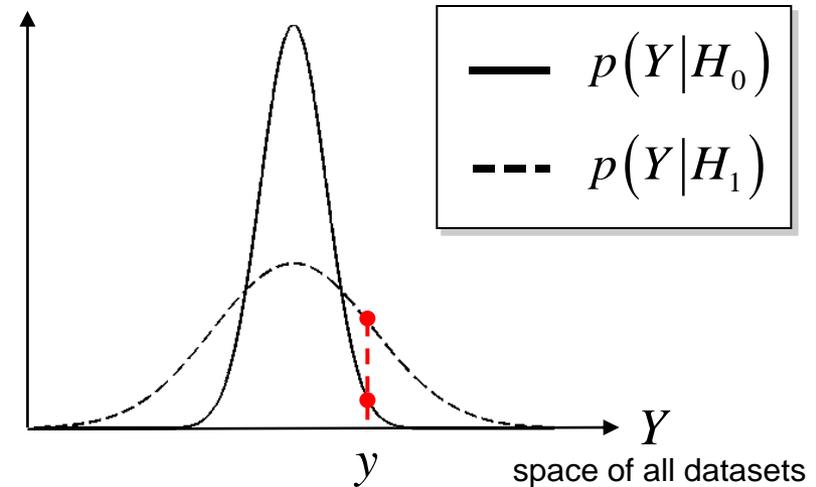
Statistical Parametric Map (SPM)

Hypothesis testing (bayesian way)

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

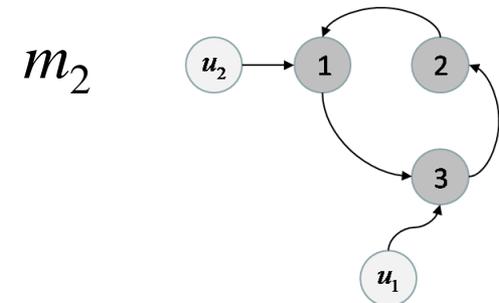
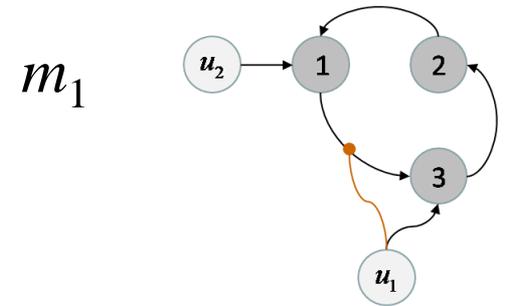
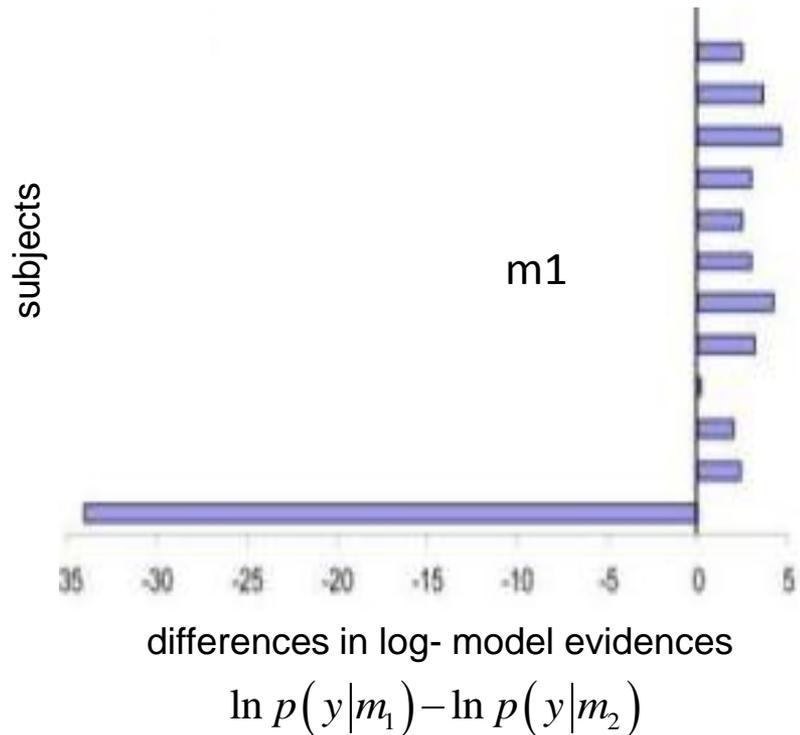
$$H_0 : p(\theta|H_0) = \begin{cases} 1 & \text{if } \theta = 0 \\ 0 & \text{otherwise} \end{cases}$$

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



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

Model comparison for group studies



fixed effect

assume all subjects correspond to the same model

-> posterior model probability (sum of evidences over subjects) might favour model 1

random effect

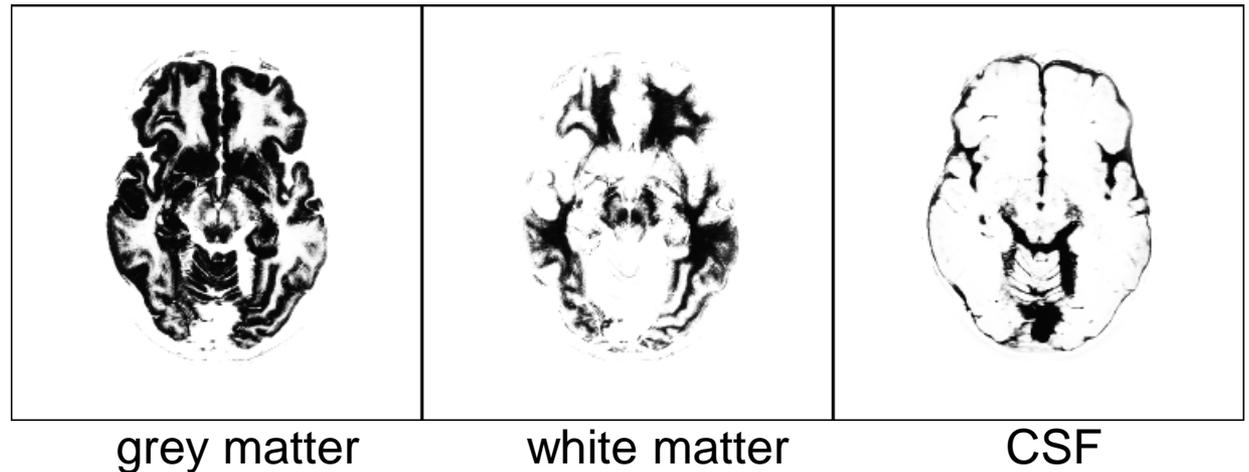
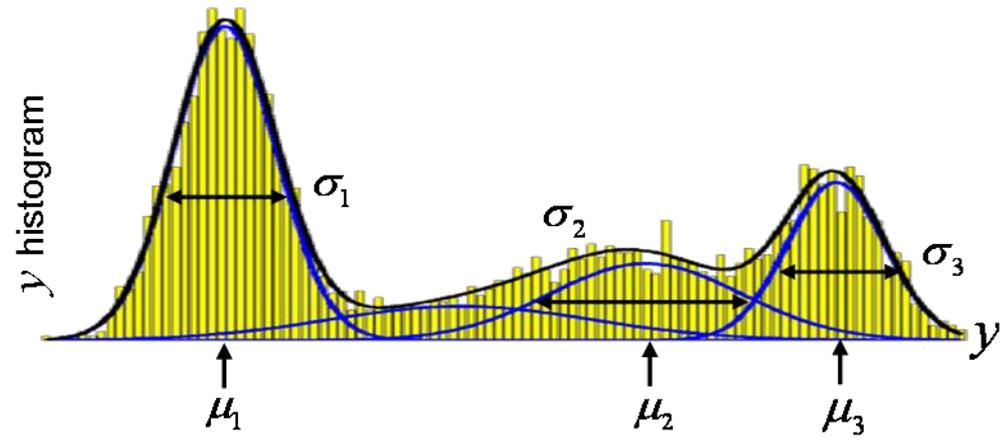
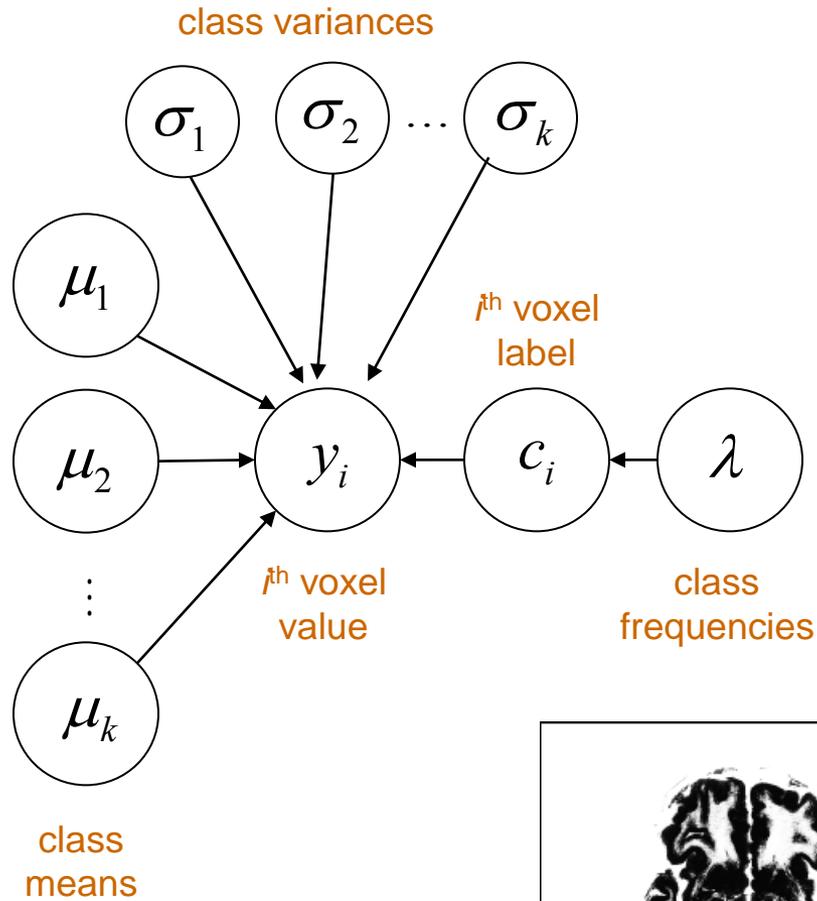
assume different subjects might correspond to different models

-> estimates how likely is a given model at the origin of the data of any randomly selected subject

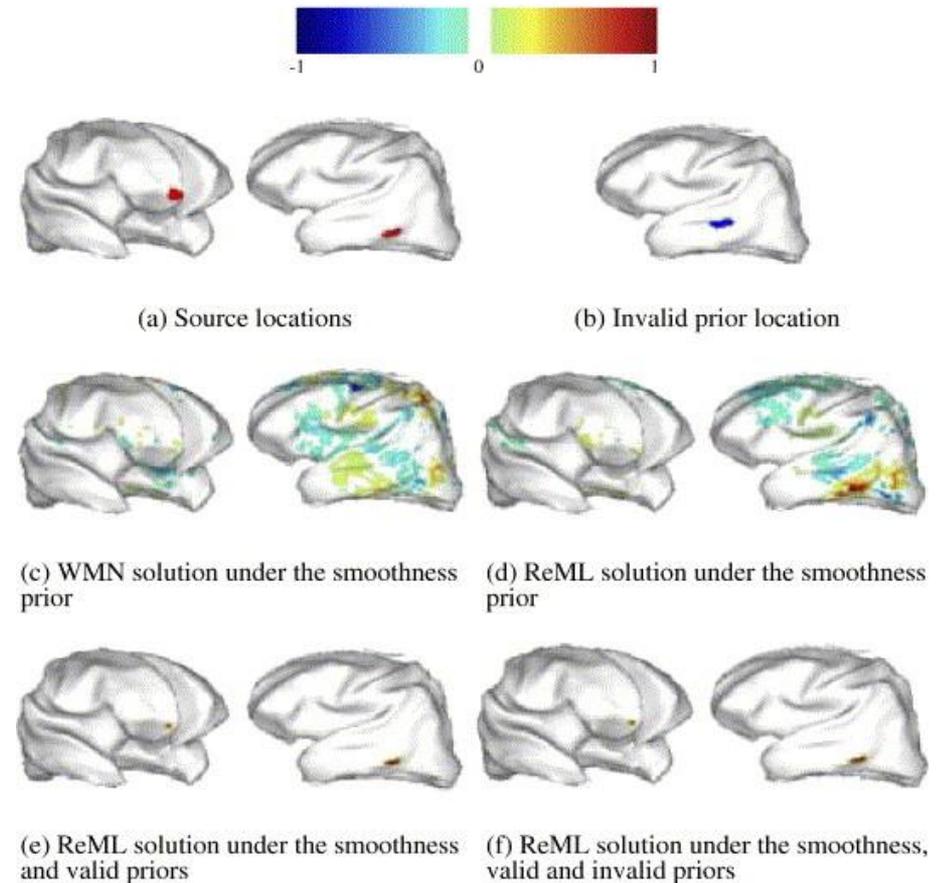
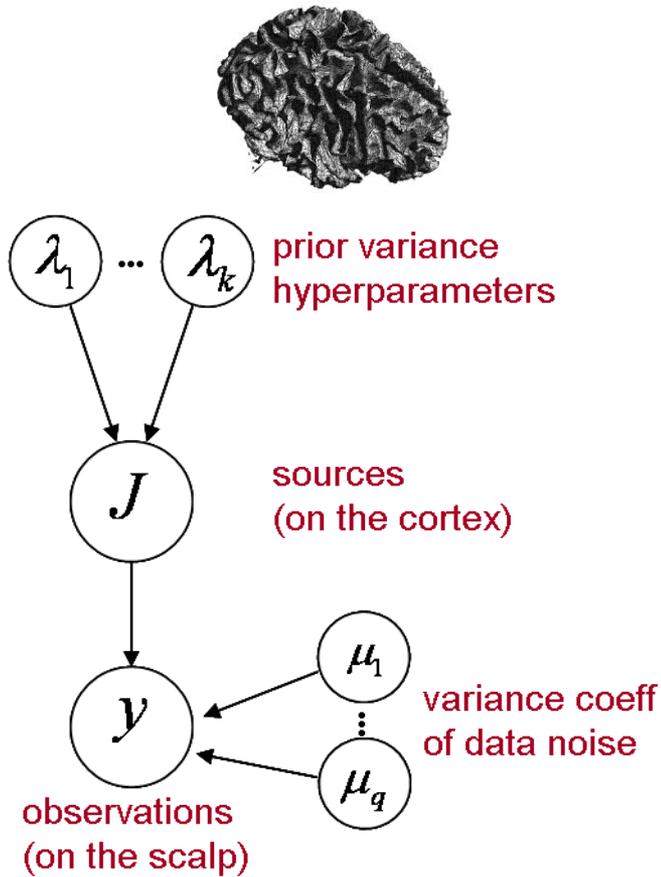
-> exceedance probability: how more likely is one model compare to any other, as the generator of the group's data

- General principles
- The Bayesian way
- **SPM examples**

Segmentation of anatomical MRI

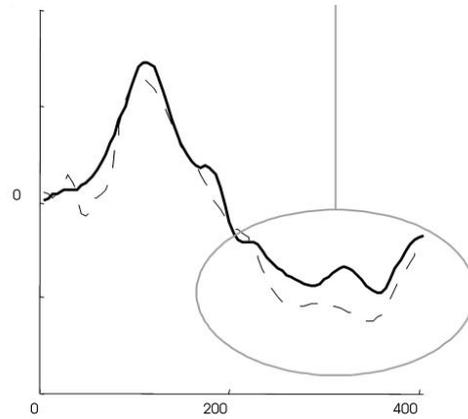
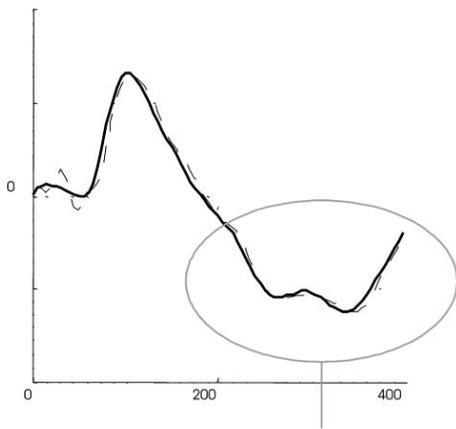


EEG/MEG source reconstruction

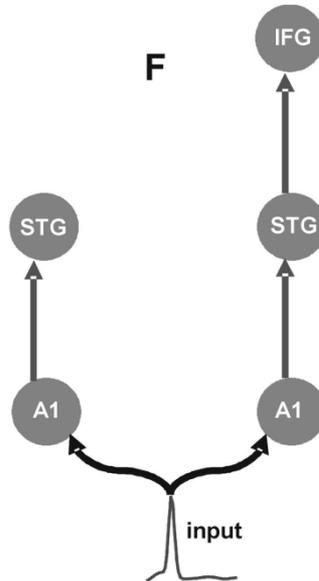
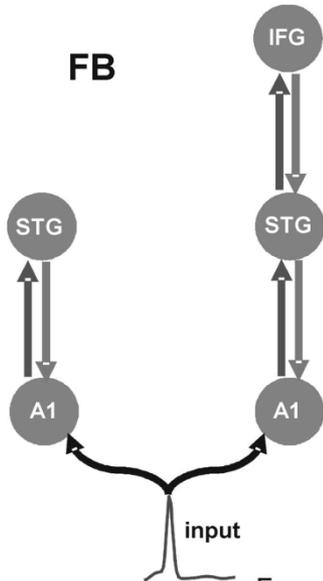
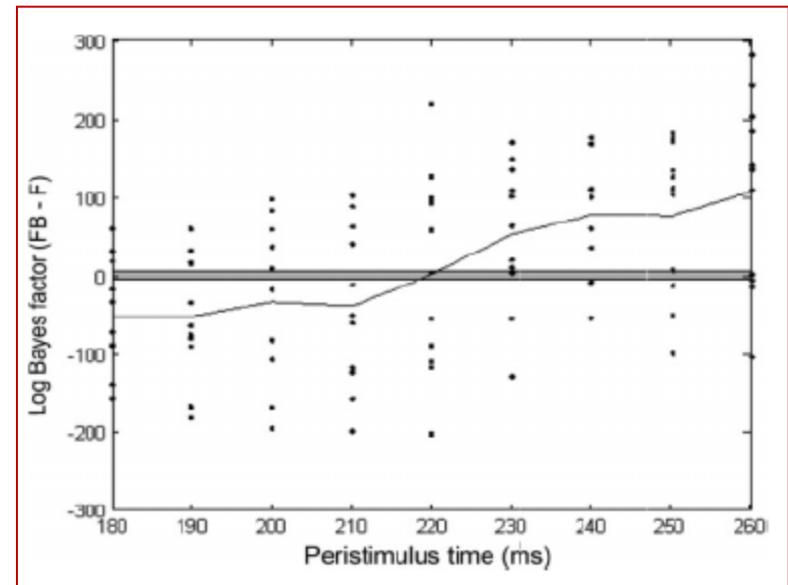


Dynamic causal modelling of EEG data

Evidence for feedback loops (MMN paradigm)



Devient condition



Suggestions for further reading



<http://www.fil.ion.ucl.ac.uk/~wpenny/>

