An introduction to Bayesian inference and model comparison

J. Daunizeau

ICM, Paris, France
TNU, Zurich, Switzerland

Overview of the talk

- ✓ An introduction to probabilistic modelling
- ✓ Bayesian model comparison
- ✓ SPM applications

Overview of the talk

- ✓ An introduction to probabilistic modelling
- ✓ Bayesian model comparison
- ✓ SPM applications

Probability theory: basics

***Degree of plausibility** desiderata*:
- should be represented using real numbers (D1)
- should conform with intuition (D2)
- should be consistent (D3)

a=2

• normalization: $$\sum_a P(a) = 1$$

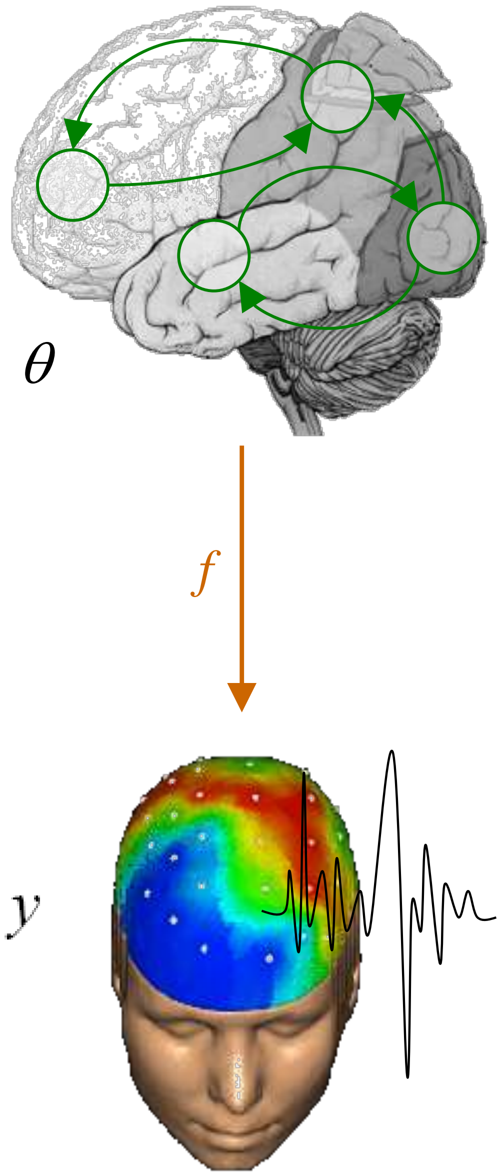
a=2 b=5

• marginalization: $$P(b) = \sum_a P(a,b)$$

• conditioning : (*Bayes rule*)

$$P(a,b) = P(a|b)\,P(b) = P(b|a)\,P(a)$$

Deriving the likelihood function



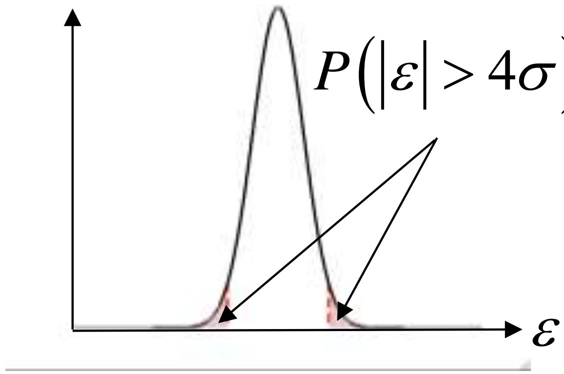
- Model of data with unknown parameters:

$$y = f(\theta) \qquad \text{e.g., GLM:} \quad f(\theta) = X\theta$$

- But data is noisy: $y = f(\theta) + \varepsilon$

- Assume noise/residuals is 'small':

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



$$P(|\varepsilon| > 4\sigma) \approx 0.05$$

→ Distribution of data, *given fixed parameters*:

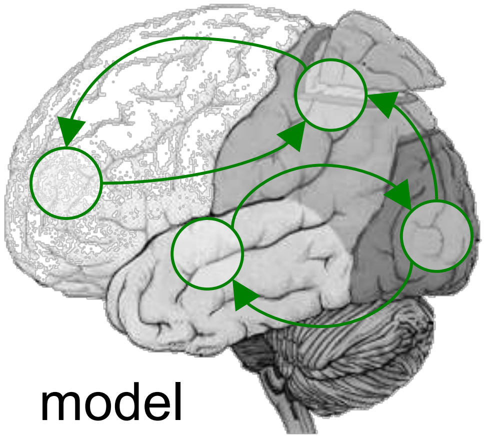
$$p(y|\theta) \propto \exp\left(-\frac{1}{2\sigma^2}\left(y - f(\theta)\right)^2\right)$$

Forward and inverse problems

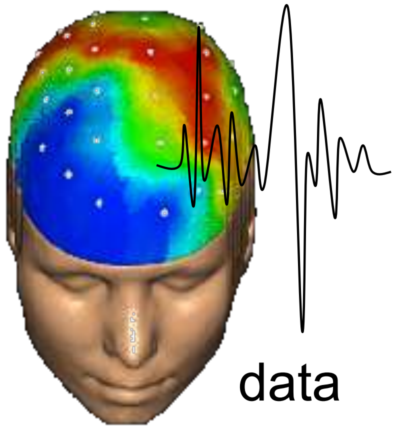
forward problem

$$p\left(y \middle| \vartheta, m\right)$$

likelihood



model



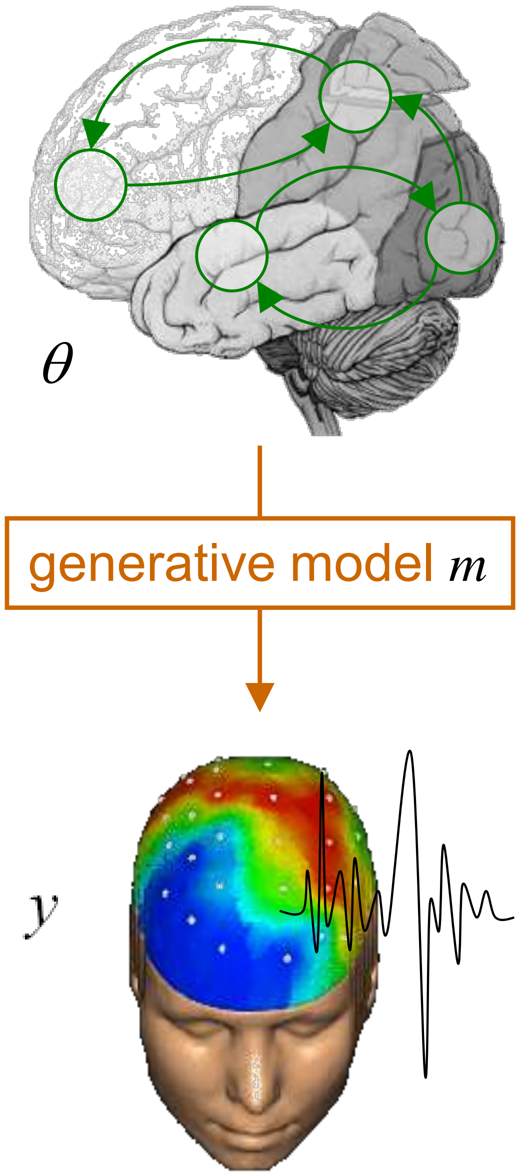
data

posterior distribution

$$p\left(\vartheta \middle| y, m\right)$$

inverse problem

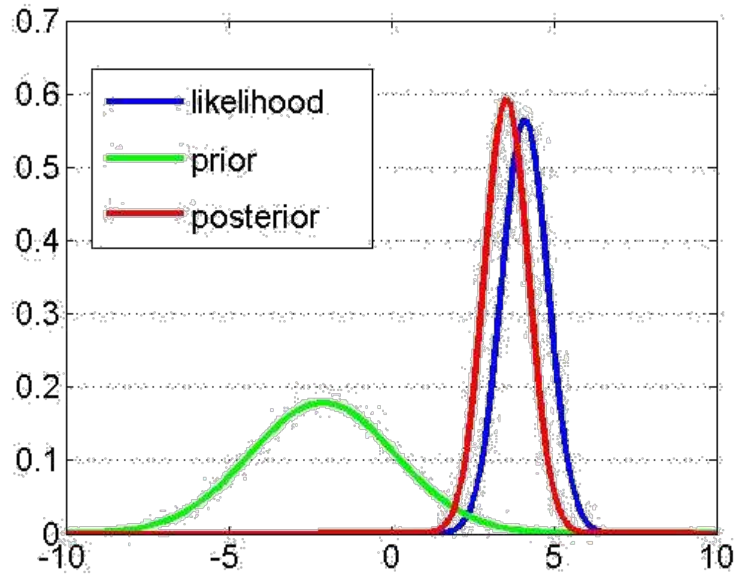
Likelihood, priors and the model evidence



Likelihood: $p(y|\theta, m)$

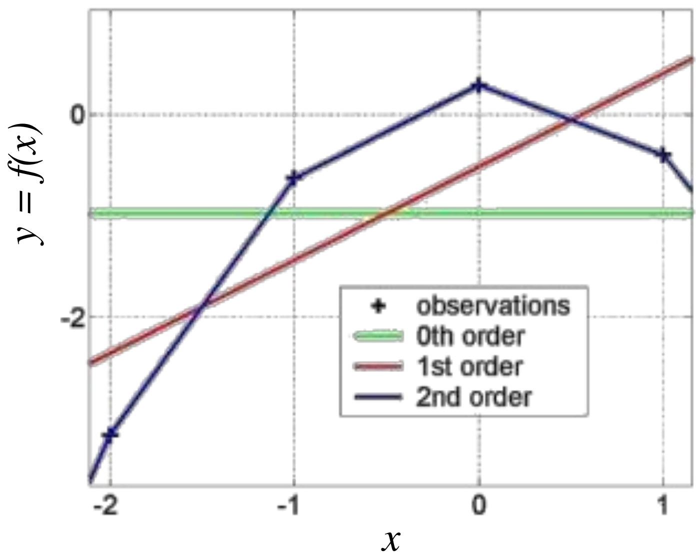
Prior: $p(\theta|m)$

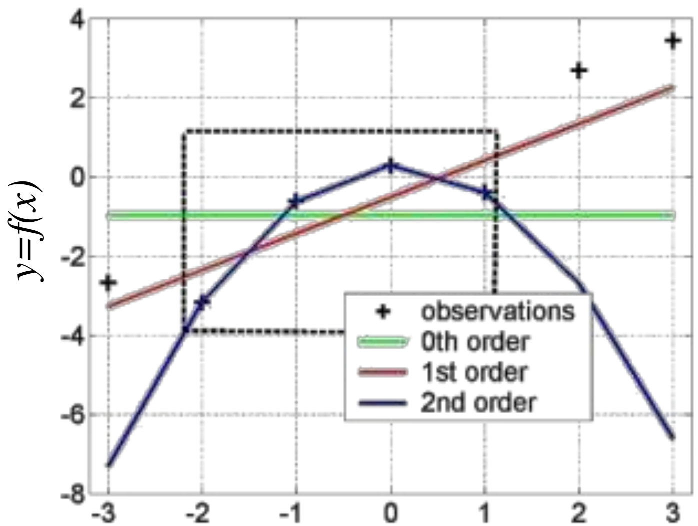
Bayes rule: $$p(\theta|y,m) = \frac{p(y|\theta,m)\,p(\theta|m)}{p(y|m)}$$



Bayesian model comparison

Principle of parsimony :
« plurality should not be assumed without necessity »

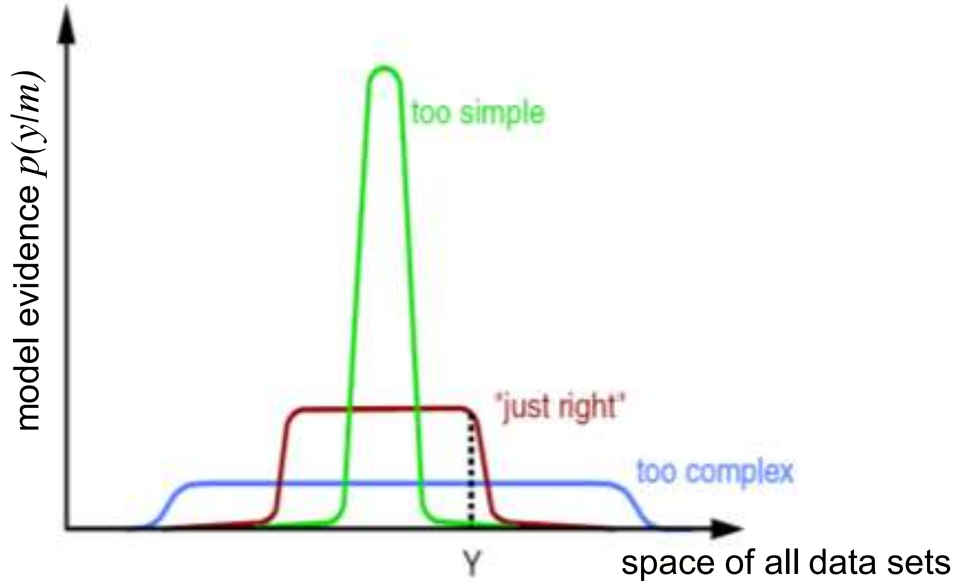




Model evidence:

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

"Occam's razor" :

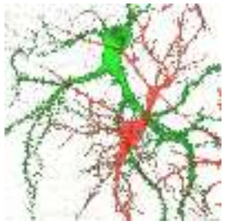


Hierarchical models

$\vdots$

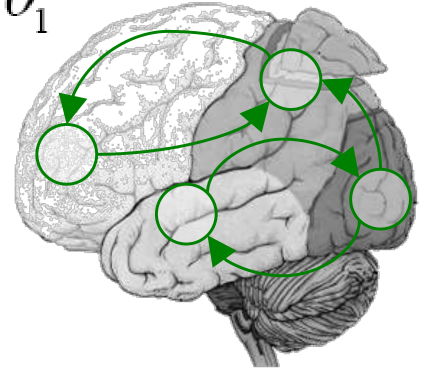
inference

θ_2



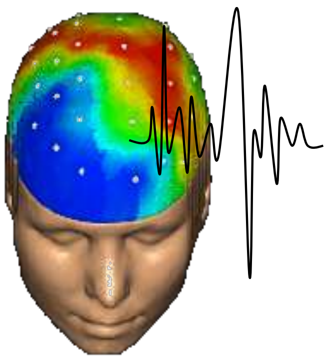
$$p(\theta_2 | \theta_3, m)$$

θ_1



$$p(\theta_1 | \theta_2, m)$$

y

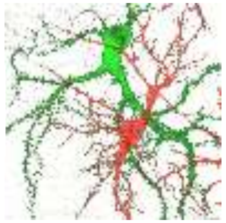


$$p(y | \theta_1, m)$$

causality

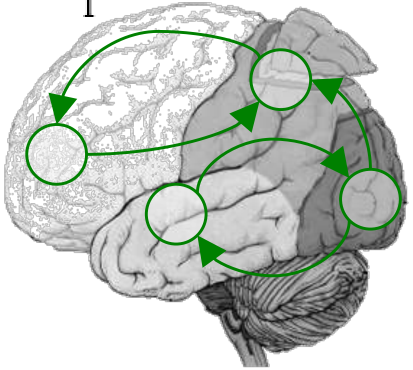
Directed acyclic graphs (DAGs)

θ_2

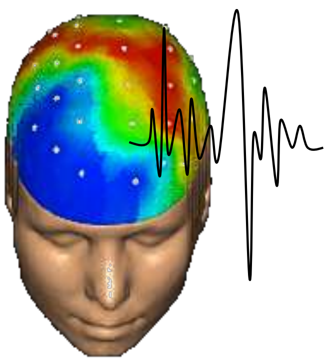


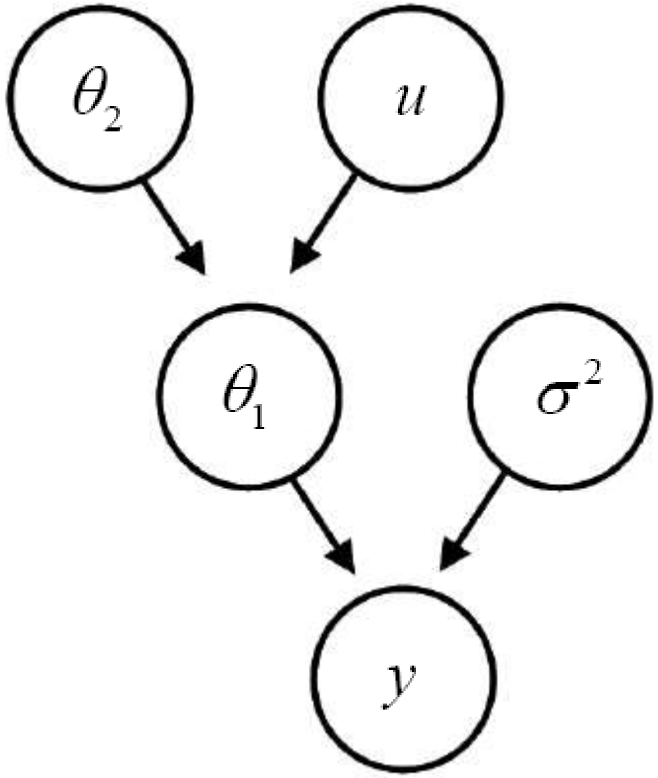
u

θ_1



y





$$p(\theta_1|\theta_2,u,m)$$

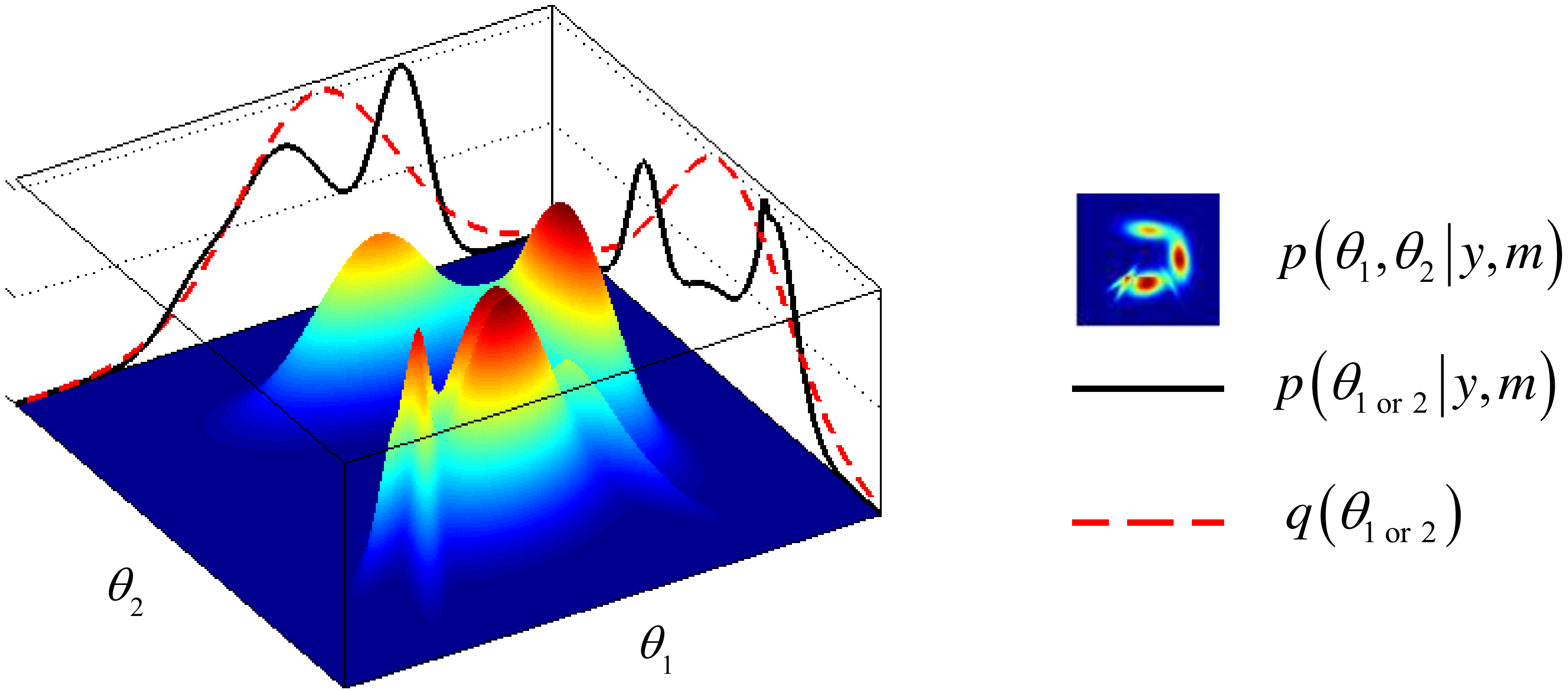
$$p(y|\theta_1,\sigma^2,m)$$

$$p(\theta|m)=\prod_j p(\theta_j|par(\theta_j),m)$$

Variational approximations (VB, EM, ReML)

$$\ln p(y|m) = \underbrace{\left\langle \ln p(y,\theta|m) \right\rangle_q + S(q)}_{\text{Free energy } F(q)} + KL\left(p(\theta|y,m);q(\theta)\right)$$

→ **VB** : maximize the free energy $F(q)$ w.r.t. the approximate posterior $q(\theta)$ under some (e.g., *mean field, Laplace*) simplifying constraint

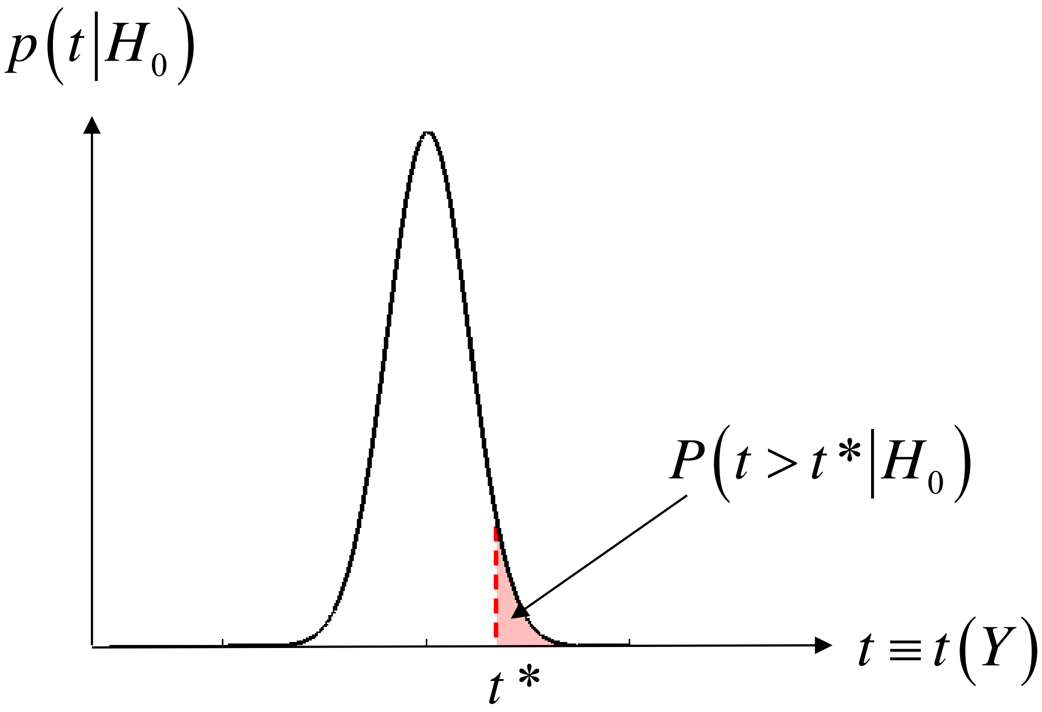


Overview of the talk

- ✓ An introduction to probabilistic modelling
- ✓ Bayesian model comparison
- ✓ SPM applications

Frequentist versus Bayesian inference

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



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

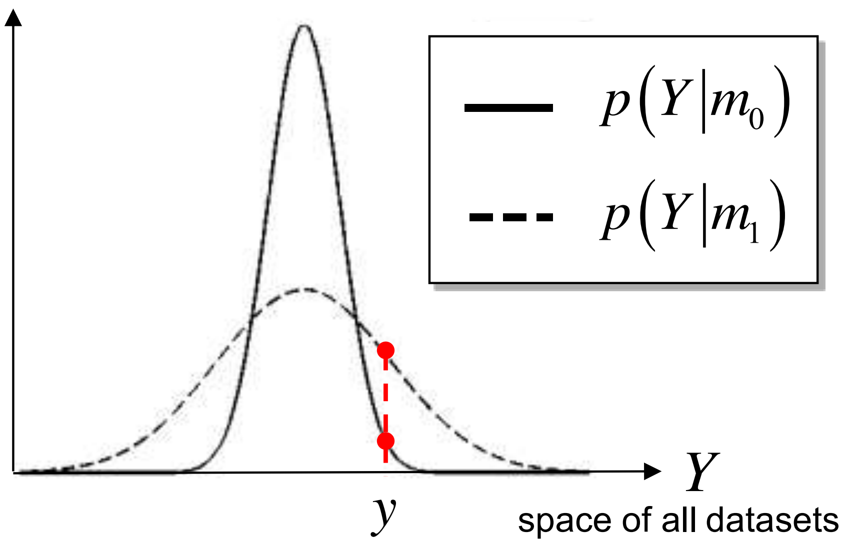
if $P(t > t^* | H_0) \leq \alpha$ then reject H0

classical (null) hypothesis testing

- define two alternative models, e.g.:

$$m_0 : p(\theta|m_0) = \begin{cases} 1 & \text{if } \theta = 0 \\ 0 & \text{otherwise} \end{cases}$$

$$m_1 : p(\theta|m_1) = N(0, \Sigma)$$



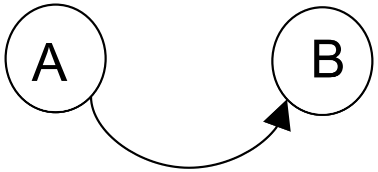
- apply decision rule, e.g.:

if $\frac{P(m_0|y)}{P(m_1|y)} \geq \alpha$ then accept m_0

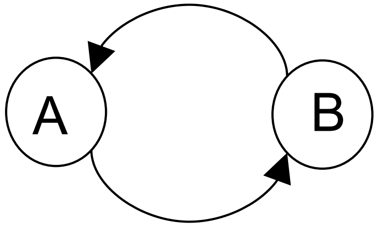
Bayesian Model Comparison

Family-level inference

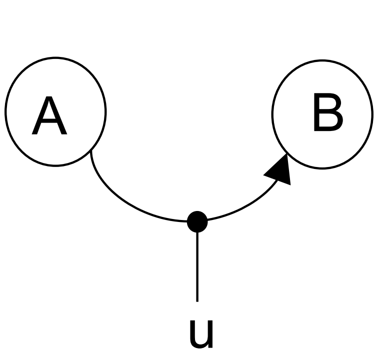
$P(m_1|y) = 0.04$



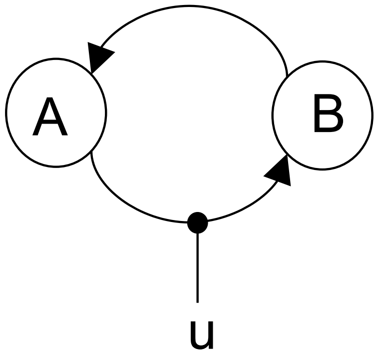
$P(m_2|y) = 0.25$



$P(m_2|y) = 0.01$



$P(m_2|y) = 0.7$

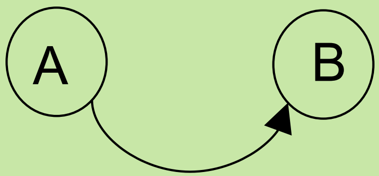


model selection error risk:

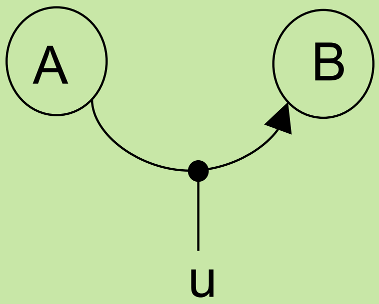
$$P(e=1|y) = 1 - \max_m P(m|y)$$
$$= 0.3$$

Family-level inference

$P(m_1|y) = 0.04$

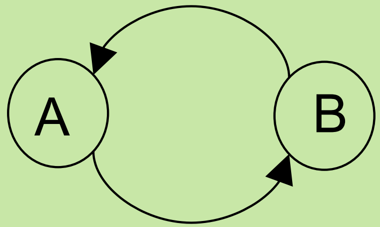


$P(m_2|y) = 0.01$

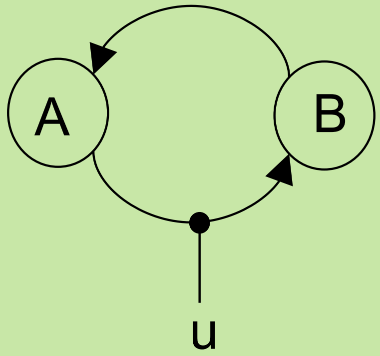


$P(f_1|y) = 0.05$

$P(m_2|y) = 0.25$



$P(m_2|y) = 0.7$



$P(f_2|y) = 0.95$

model selection error risk:

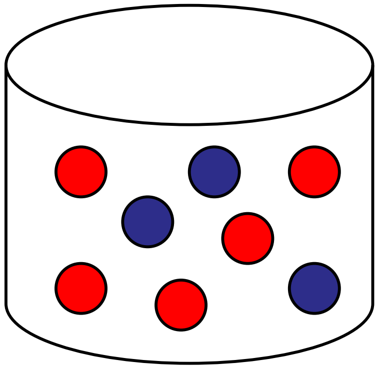
$$P(e=1|y) = 1 - \max_m P(m|y) = 0.3$$

family inference
(pool statistical evidence)

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

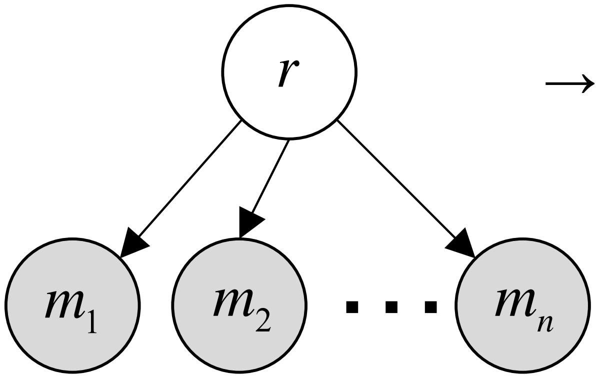
$$P(e=1|y) = 1 - \max_f P(f|y) = 0.05$$

Sampling subjects as marbles in an urn



$$\begin{cases} m_i = 1 & \rightarrow i^{\text{th}} \text{ marble is blue} \\ m_i = 0 & \rightarrow i^{\text{th}} \text{ marble is purple} \end{cases}$$

r = proportion of blue marbles in the urn



→ (binomial) probability of drawing a set of n marbles:

$$p(m|r) = \prod_{i=1}^{n} r^{m_i} (1-r)^{1-m_i}$$

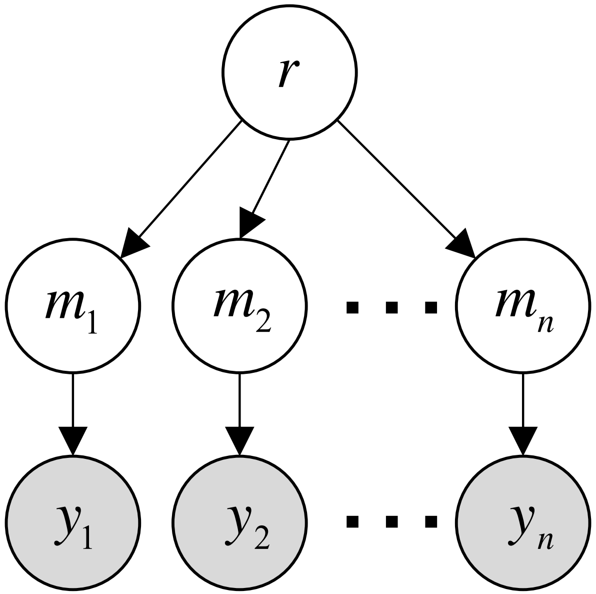
Thus, our belief about the proportion of blue marbles is:

$$p(r|m) \propto p(r) \prod_{i=1}^{n} r^{m_i} (1-r)^{1-m_i} \quad \overset{p(r)\propto 1}{\Rightarrow} \quad E[r|m] = \frac{1}{n} \sum_{i=1}^{n} m_i$$

Group-level model comparison

At least, we can measure how likely is the i^{th} subject's data under each model!

$$p\left(y_1|m_1\right) \quad p\left(y_2|m_2\right) \qquad \cdots \qquad p\left(y_i|m_i\right) \qquad \cdots \qquad p\left(y_n|m_n\right)$$



$$p\left(r,m|y\right) \propto p\left(r\right)\prod_{i=1}^{n} p\left(y_i|m_i\right)p\left(m_i|r\right)$$

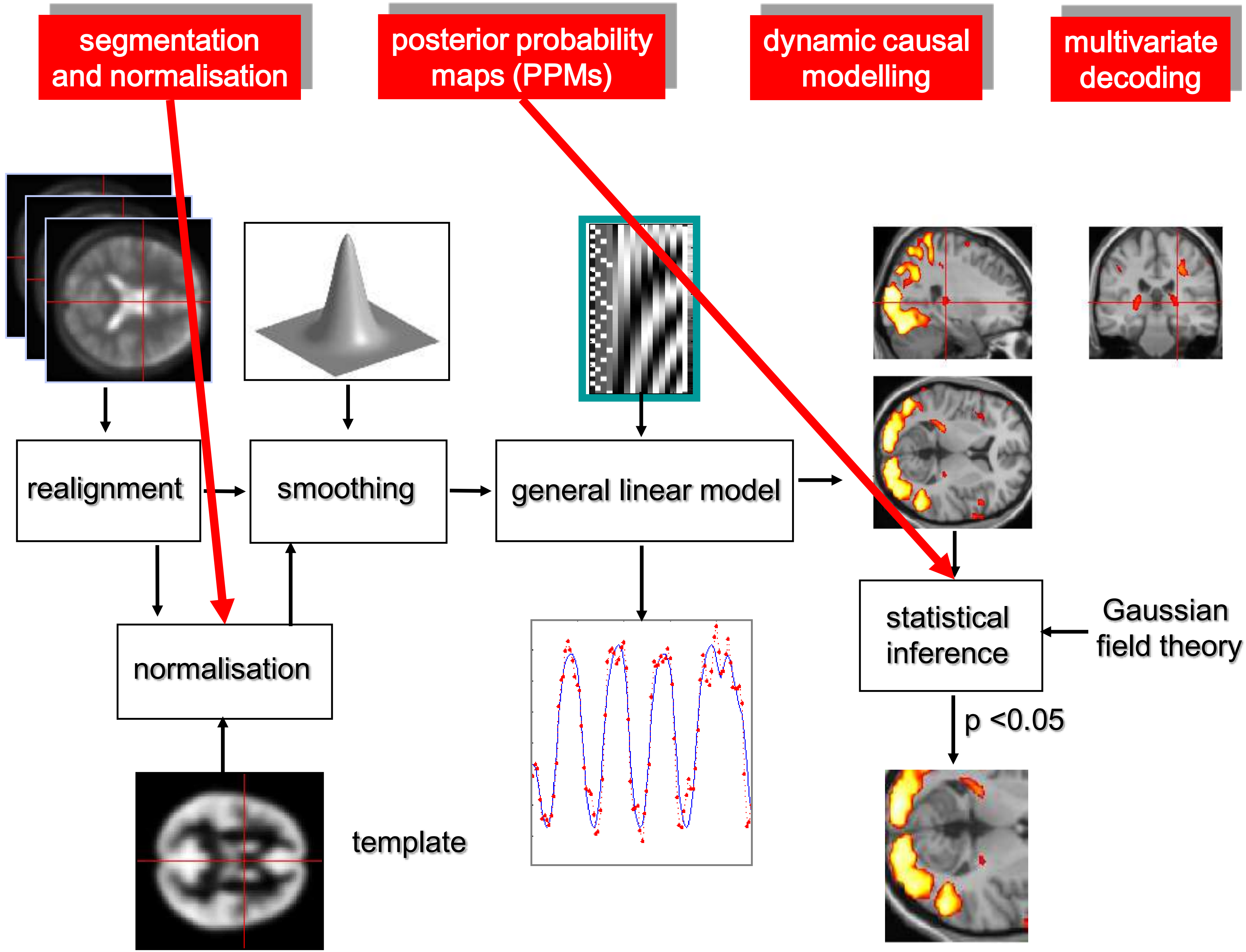
Our belief about the proportion of models is:

$$p\left(r|y\right) = \sum_{m} p\left(r,m|y\right)$$

Exceedance probability: $\varphi_k = P\left(r_k > r_{k'\neq k}|y\right)$

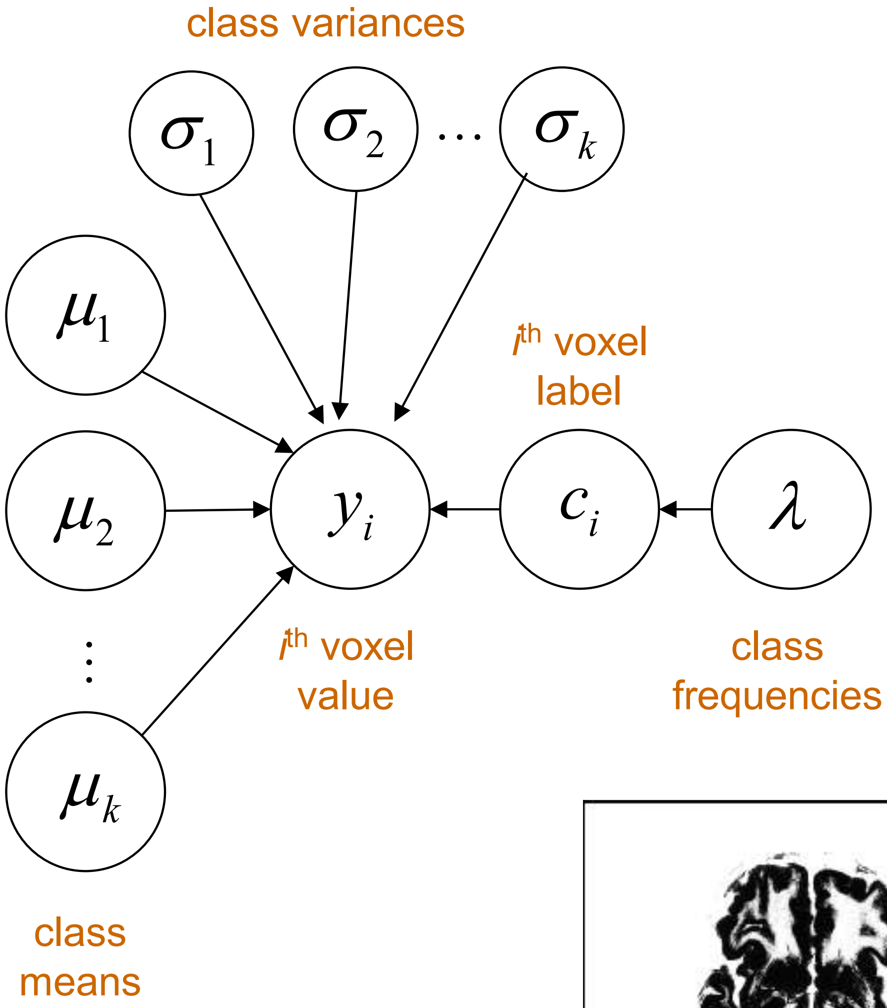
Overview of the talk

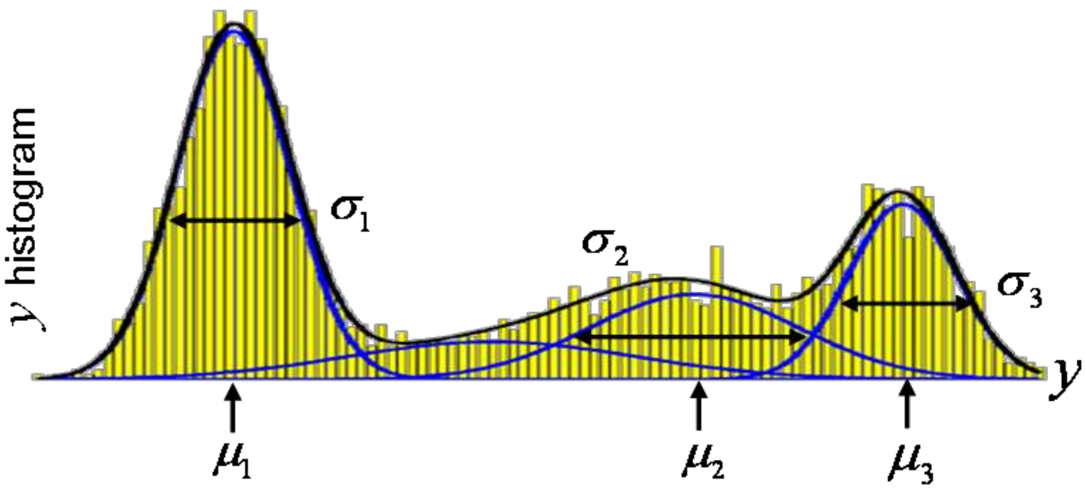
- ✓ An introduction to probabilistic modelling
- ✓ Bayesian model comparison
- ✓ SPM applications

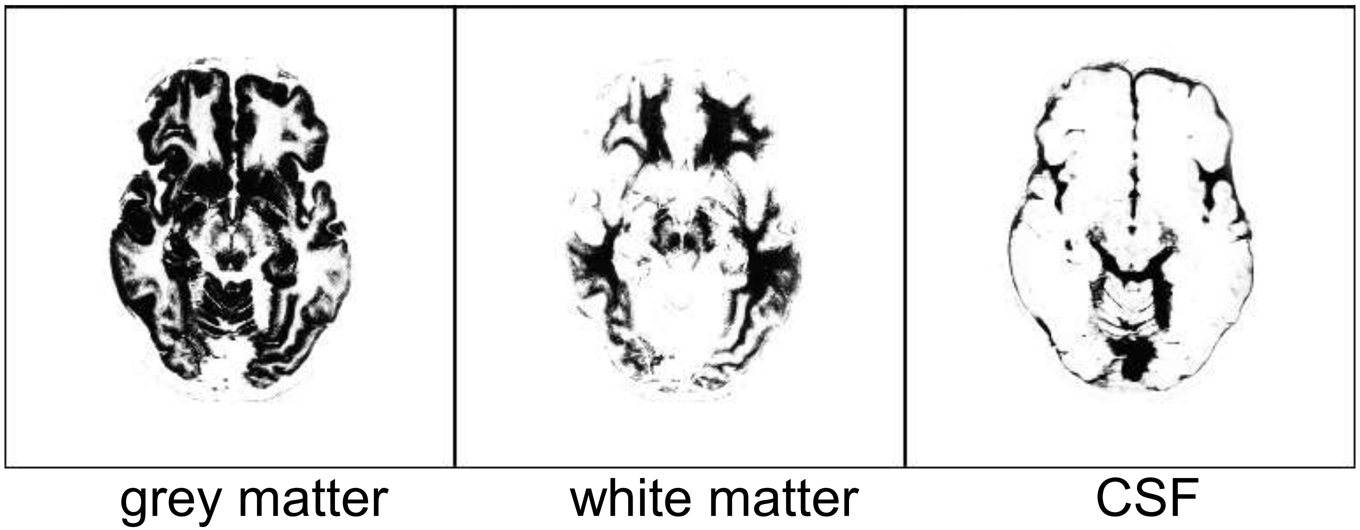

segmentation and normalisation
posterior probability maps (PPMs)
dynamic causal modelling
multivariate decoding
realignment
smoothing
general linear model
normalisation
template
statistical inference
Gaussian field theory
p <0.05

aMRI segmentation

mixture of Gaussians (MoG) model

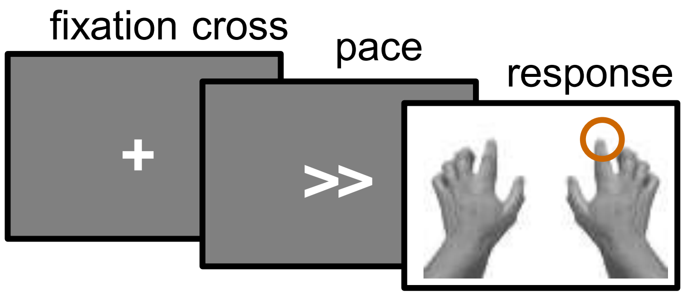


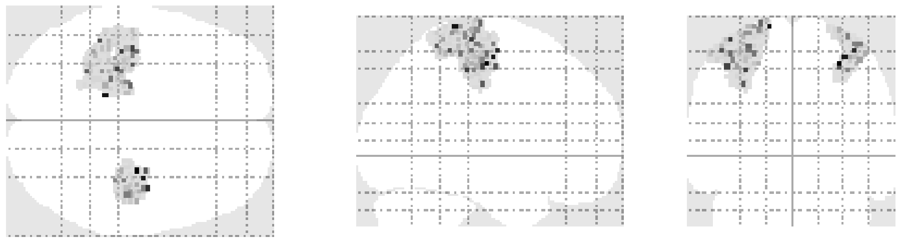




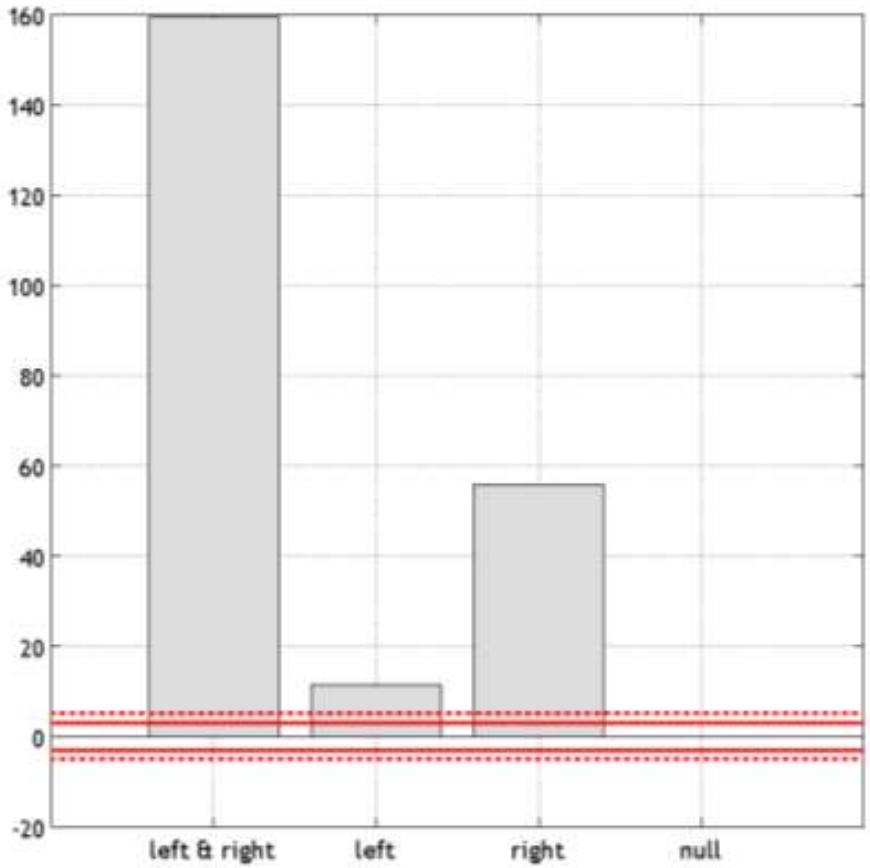
Decoding of brain images

recognizing brain states from fMRI

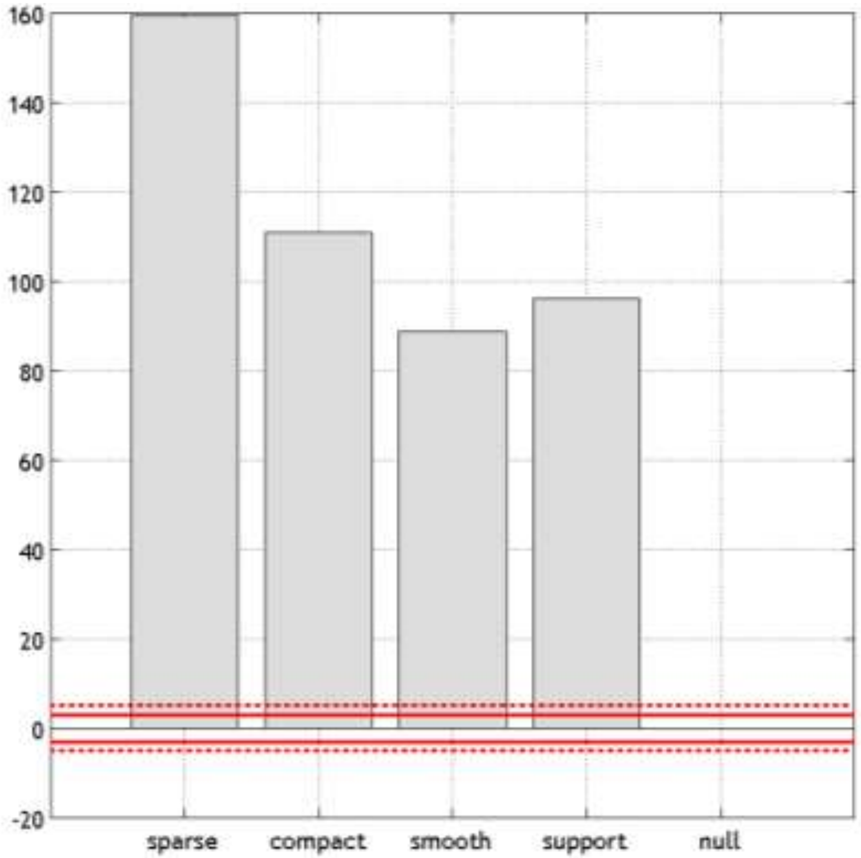




log-evidence of X-Y sparse mappings:
effect of lateralization



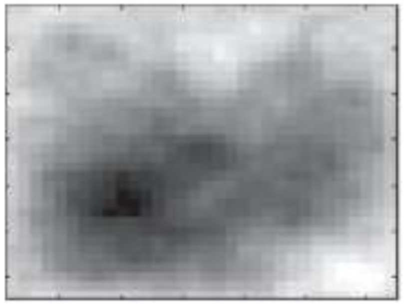
log-evidence of X-Y bilateral mappings:
effect of spatial deployment

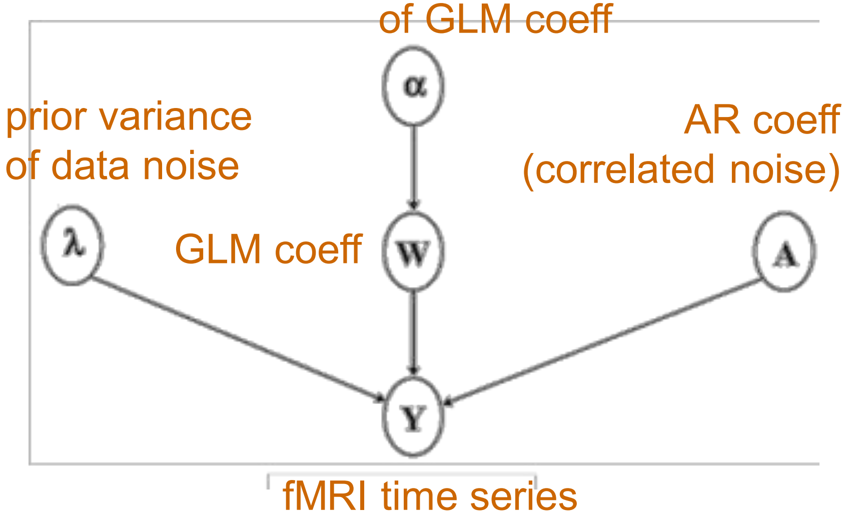


fMRI time series analysis

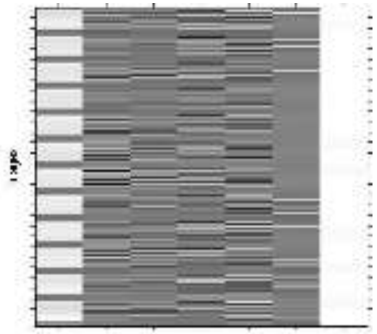
spatial priors and model comparison

		1		
	2	-8	2	
1	-8	20	-8	1
	2	-8	2	
		1		

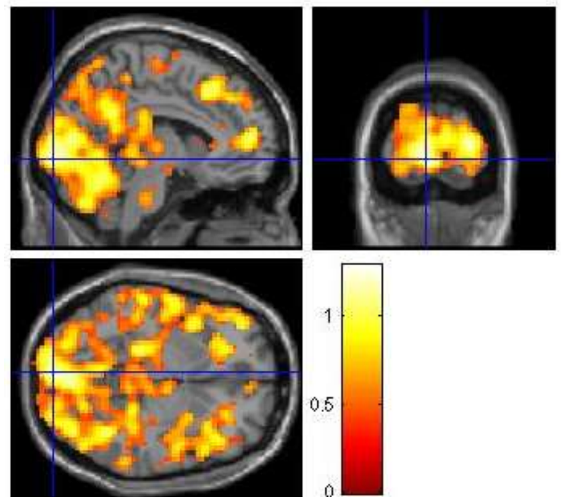




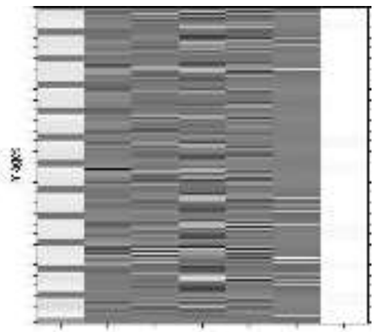
short-term memory
design matrix (X)



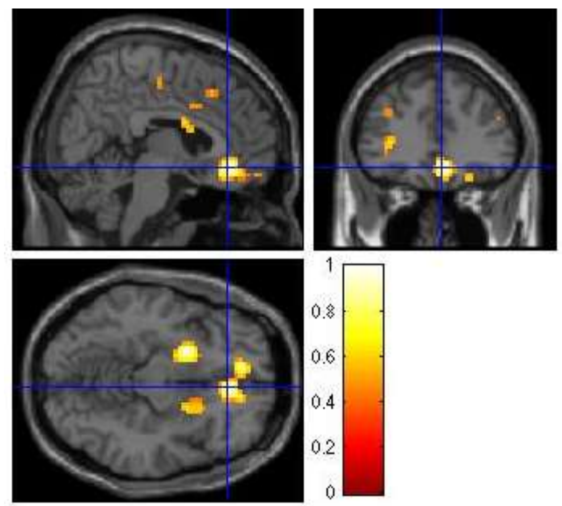
PPM: regions best explained
by short-term memory model



long-term memory
design matrix (X)

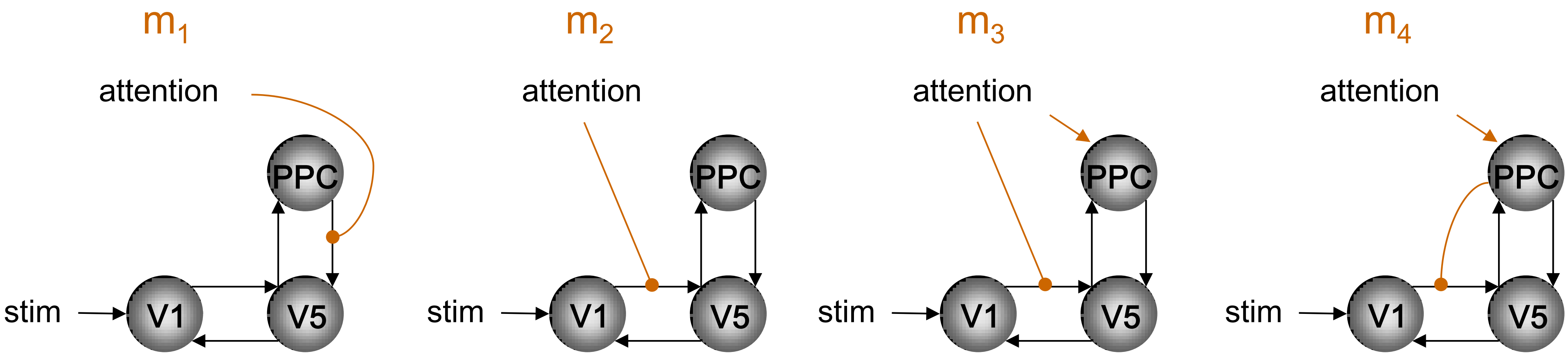


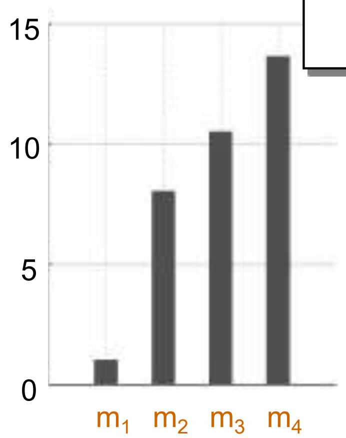
PPM: regions best explained
by long-term memory model



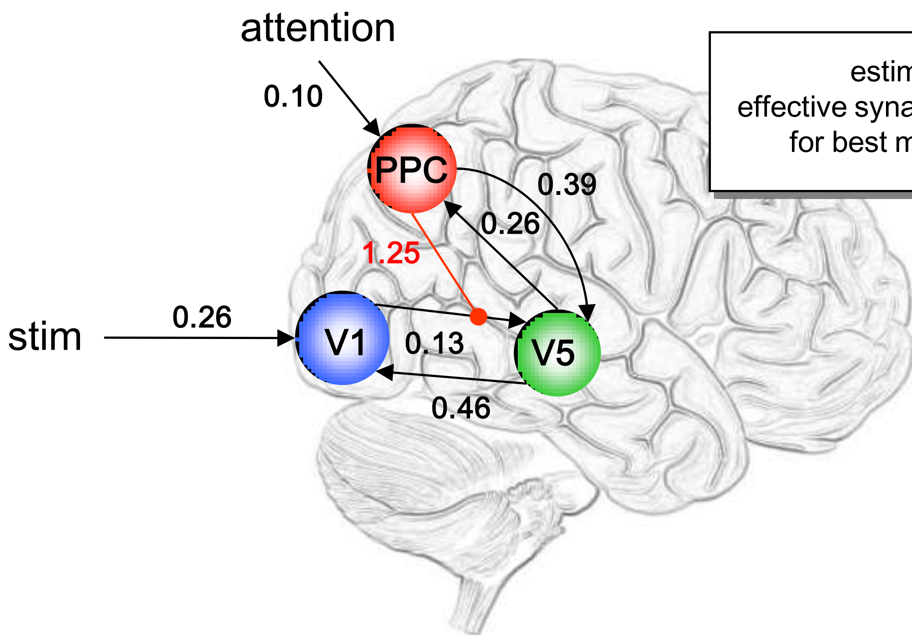
Dynamic Causal Modelling

network structure identification



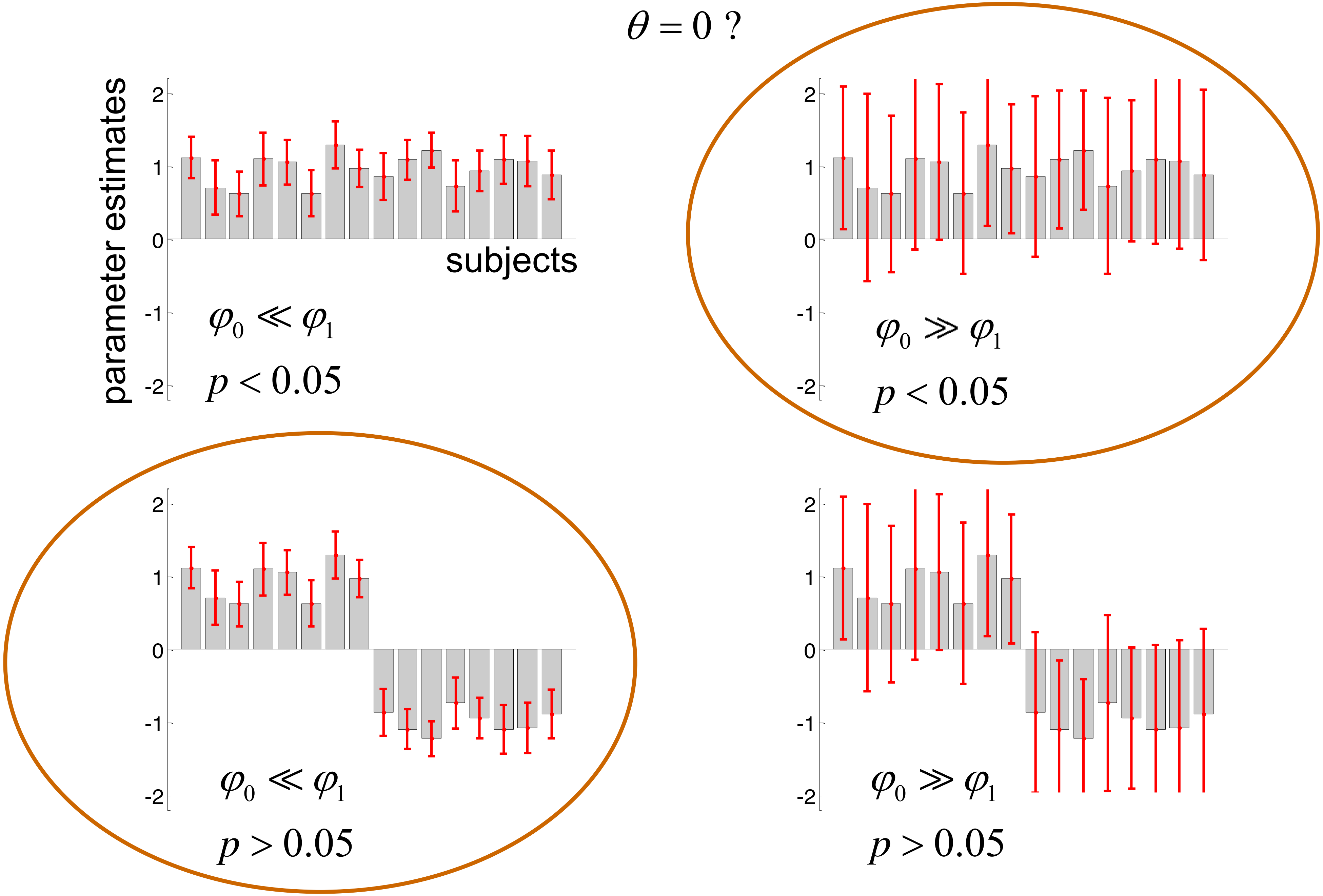


models marginal likelihood
$\ln p(y|m)$



estimated
effective synaptic strengths
for best model (m_4)

SPM: frequentist vs Bayesian RFX analysis



I thank you for your attention.

A note on statistical significance

lessons from the Neyman-Pearson lemma

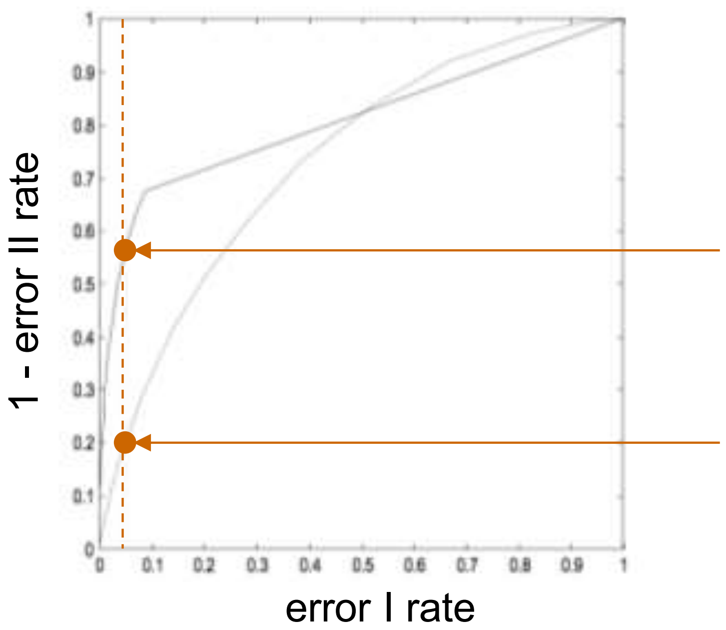
- Neyman-Pearson lemma: the likelihood ratio (or Bayes factor) test

$$\Lambda = \frac{p(y|H_1)}{p(y|H_0)} \geq u$$

is the most powerful test of size $\alpha = p(\Lambda \geq u | H_0)$ to test the null.

- what is the threshold u, above which the Bayes factor test yields a error I rate of 5%?

ROC analysis



MVB (Bayes factor)
u=1.09, power=56%

CCA (F-statistics)
F=2.20, power=20%