Bayesian Inference

Chris Mathys

Wellcome Trust Centre for Neuroimaging

UCL

London SPM Course

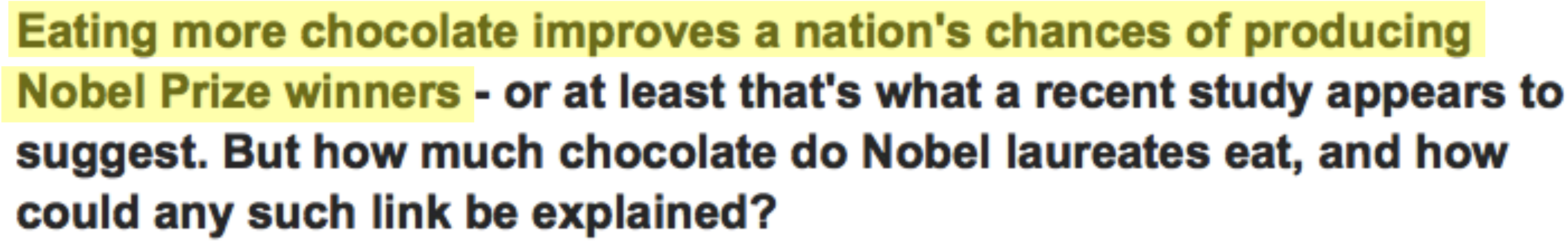
Thanks to Jean Daunizeau and Jérémie Mattout for previous versions of this talk

A spectacular piece of information

19 November 2012 Last updated at 18:19

Does chocolate make you clever?

By Charlotte Pritchard
BBC News



Eating more chocolate improves a nation's chances of producing Nobel Prize winners - or at least that's what a recent study appears to suggest. But how much chocolate do Nobel laureates eat, and how could any such link be explained?

A spectacular piece of information

Messerli, F. H. (2012). Chocolate Consumption, Cognitive Function, and Nobel Laureates. *New England Journal of Medicine*, *367*(16), 1562–1564.

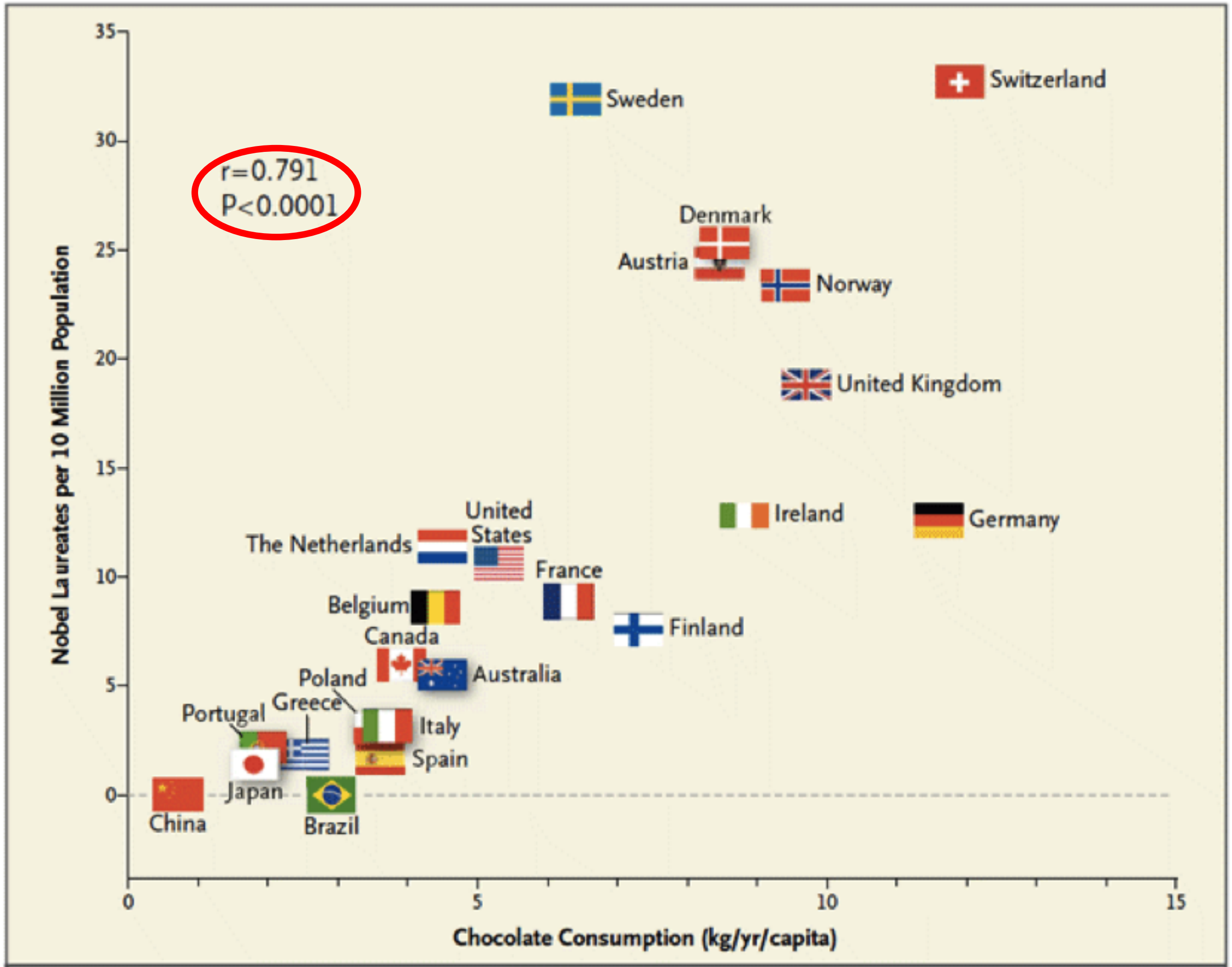


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

NEW ENGLAND JOURNAL OF MEDICINE

So will I win the Nobel prize if I eat lots of chocolate?

This is a question referring to uncertain quantities. Like almost all scientific questions, it cannot be answered by deductive logic. Nonetheless, quantitative answers can be given – but they can only be given in terms of probabilities.

Our question here can be rephrased in terms of a conditional probability:

$$p(Nobel \mid lots\ of\ chocolate) = ?$$

To answer it, we have to learn to calculate such quantities. The tool for this is Bayesian inference.

«Bayesian» = logical
and
logical = probabilistic

«The actual science of logic is conversant at present only with things either certain, impossible, or entirely doubtful, none of which (fortunately) we have to reason on. Therefore 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

«Bayesian» = logical
and
logical = probabilistic

But in what sense is probabilistic reasoning (i.e., reasoning about uncertain quantities according to the rules of probability theory) «logical»?

R. T. Cox showed in 1946 that the rules of probability theory can be derived from three basic desiderata:

1. Representation of degrees of plausibility by real numbers
2. Qualitative correspondence with common sense (in a well-defined sense)
3. Consistency

The rules of probability

By mathematical proof (i.e., by deductive reasoning) the three desiderata as set out by Cox imply the rules of probability (i.e., the rules of inductive reasoning).

This means that anyone who accepts the desiderata must accept the following rules:

1. $\sum_a p(a) = 1$ (Normalization)
2. $p(b) = \sum_a p(a, b)$ (Marginalization – also called the **sum rule**)
3. $p(a, b) = p(a|b)p(b) = p(b|a)p(a)$ (Conditioning – also called the **product rule**)

«Probability theory is nothing but common sense reduced to calculation.»

— Pierre-Simon Laplace, 1819

Conditional probabilities

The probability of ***a* given *b*** is denoted by

$$p(a|b).$$

In general, this is different from the probability of a alone (the *marginal* probability of a), as we can see by applying the sum and product rules:

$$p(a) = \sum_b p(a,b) = \sum_b p(a|b)p(b)$$

Because of the product rule, we also have the following rule (**Bayes' theorem**) for going from $p(a|b)$ to $p(b|a)$:

$$p(b|a) = \frac{p(a|b)p(b)}{p(a)} = \frac{p(a|b)p(b)}{\sum_{b'} p(a|b')p(b')}$$

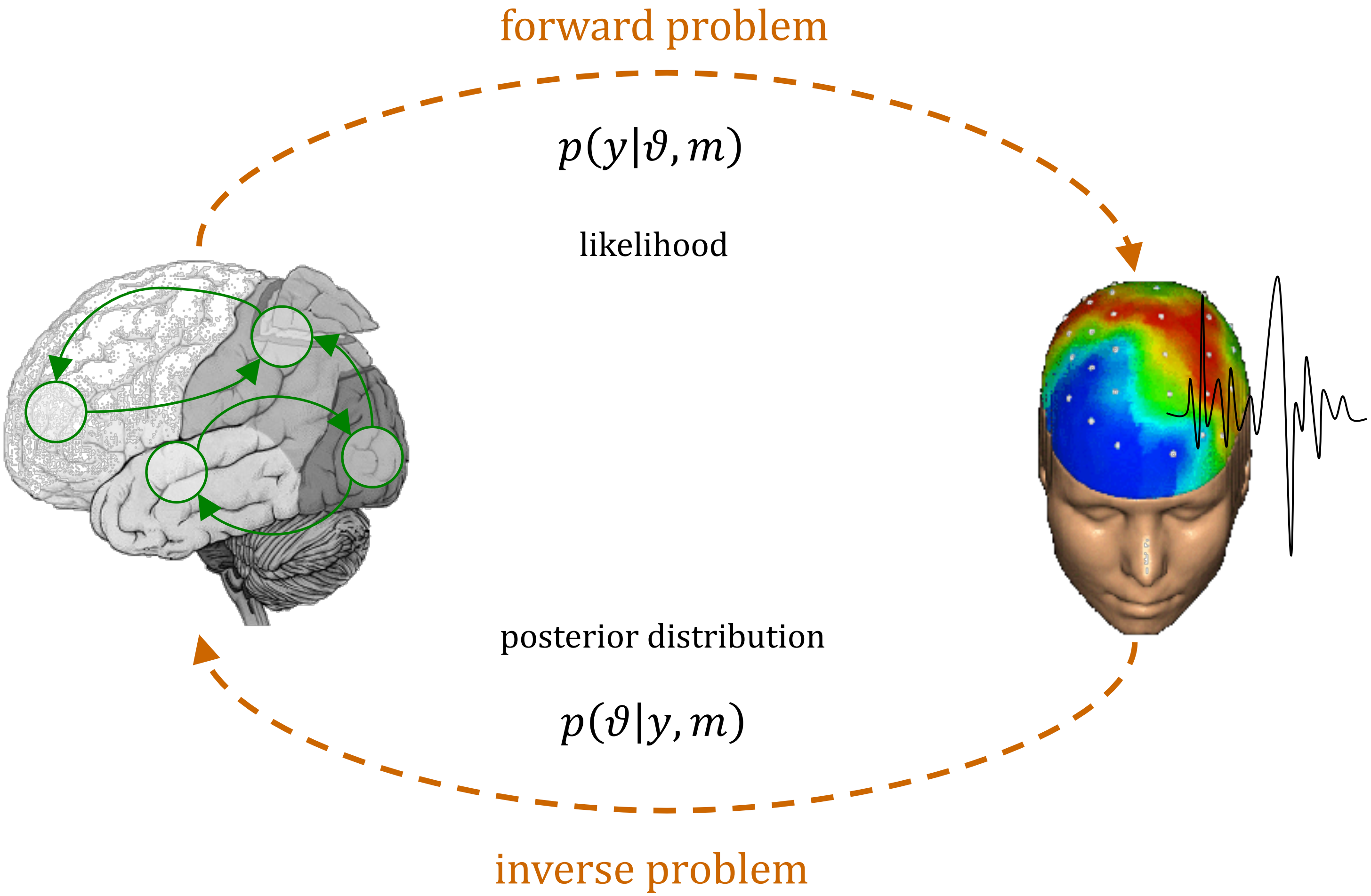
The chocolate example

In our example, it is immediately clear that $P(Nobel|chocolate)$ is very different from $P(chocolate|Nobel)$. While the first is hopeless to determine directly, the second is much easier to find out: ask Nobel laureates how much chocolate they eat. Once we know that, we can use Bayes' theorem:

$$\underbrace{p(Nobel|chocolate)}_{\text{posterior}} = \frac{\overbrace{\overbrace{p(chocolate|Nobel)}^{\text{likelihood}}\overbrace{P(Nobel)}^{\text{prior}}}^{\text{model}}}{\underbrace{p(chocolate)}_{\text{evidence}}}$$

Inference on the quantities of interest in neuroimaging studies has exactly the same general structure.

Inference in SPM



forward problem

$p(y|\vartheta, m)$

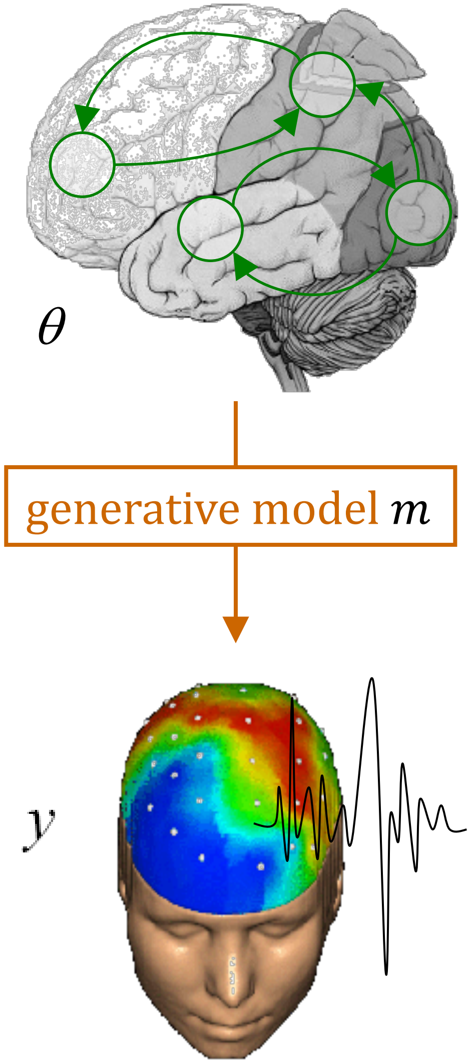
likelihood

posterior distribution

$p(\vartheta|y, m)$

inverse problem

Inference in SPM

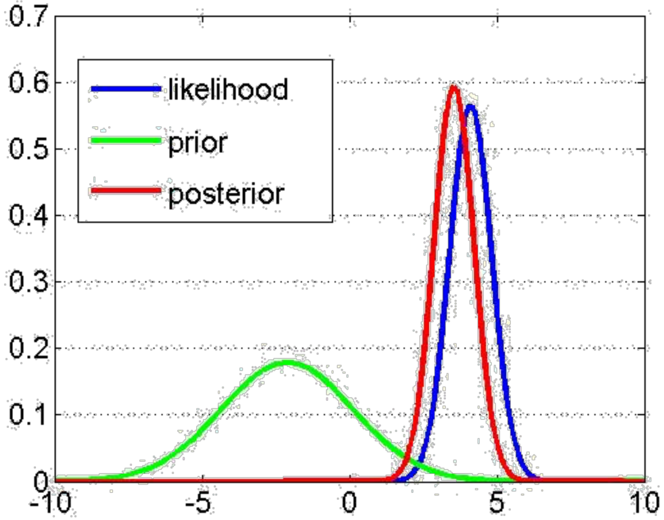


generative model m

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

Prior: $p(\vartheta|m)$

Bayes' theorem: $p(\vartheta|y, m) = \dfrac{p(y|\vartheta, m)p(\vartheta|m)}{p(y|m)}$

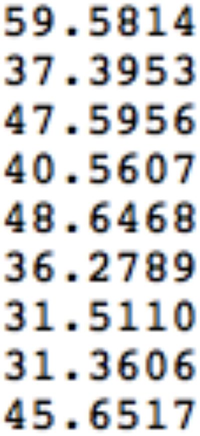


A simple example of Bayesian inference

(adapted from Jaynes (1976))

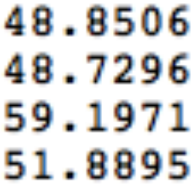
Two manufacturers, A and B, deliver the same kind of components that turn out to have the following lifetimes (in hours):

A:



```
59.5814
37.3953
47.5956
40.5607
48.6468
36.2789
31.5110
31.3606
45.6517
```

B:



```
48.8506
48.7296
59.1971
51.8895
```

Assuming prices are comparable, from which manufacturer would you buy?

A simple example of Bayesian inference

How do we compare such samples?

A simple example of Bayesian inference

What next?

Is this satisfactory?

A simple example of Bayesian inference

The procedure in brief:

- Determine your question of interest («What is the probability that...?»)
- Specify your model (likelihood and prior)
- Calculate the full posterior using Bayes' theorem
- [Pass to the uninformative limit in the parameters of your prior]
- Integrate out any nuisance parameters
- Ask your question of interest of the posterior

All you need is the rules of probability theory.

(Ok, sometimes you'll encounter a nasty integral – but that's a technical difficulty, not a conceptual one).

A simple example of Bayesian inference

The question:

- What is the probability that the components from manufacturer B have a longer lifetime than those from manufacturer A?
- More specifically: given how much more expensive they are, how much longer do I require the components from B to live.
- Example of a decision rule: if the components from B live 3 hours longer than those from A with a probability of at least 80%, I will choose those from B.

A simple example of Bayesian inference

The model (bear with me, this **will** turn out to be simple):

- Likelihood (Gaussian):

$$p(\{x_i\}|\mu,\lambda) = \prod_{i=1}^{n} \left(\frac{\lambda}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\lambda}{2}(x_i-\mu)^2\right)$$

- Prior (Gaussian-gamma):

$$p(\mu,\lambda|\mu_0,\kappa_0 a_0,b_0) = \mathcal{N}(\mu|\mu_0,(\kappa_0\lambda)^{-1})\mathrm{Gam}(\lambda|a_0,b_0)$$

A simple example of Bayesian inference

The posterior (Gaussian-gamma):

$$p(\mu, \lambda | \{x_i\}) = \mathcal{N}(\mu | \mu_n, (\kappa_n \lambda)^{-1}) \mathrm{Gam}(\lambda | a_n, b_n)$$

Parameter updates:

$$\mu_n = \mu_0 + \frac{n}{\kappa_0 + n}(\bar{x} - \mu_0), \qquad \kappa_n = \kappa_0 + n, \qquad a_n = a_0 + \frac{n}{2}$$

$$b_n = b_0 + \frac{n}{2}\left(s^2 + \frac{\kappa_0}{\kappa_0 + n}(\bar{x} - \mu_0)^2\right)$$

with

$$\bar{x} \coloneqq \frac{1}{n}\sum_{i=1}^{n} x_i\,, \qquad s^2 \coloneqq \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

A simple example of Bayesian inference

The limit for which the prior becomes uninformative:

- For $\kappa_0 = 0$, $a_0 = 0$, $b_0 = 0$, the updates reduce to:

$$\mu_n = \bar{x} \qquad \kappa_n = n \qquad a_n = \frac{n}{2} \qquad b_n = \frac{n}{2} s^2$$

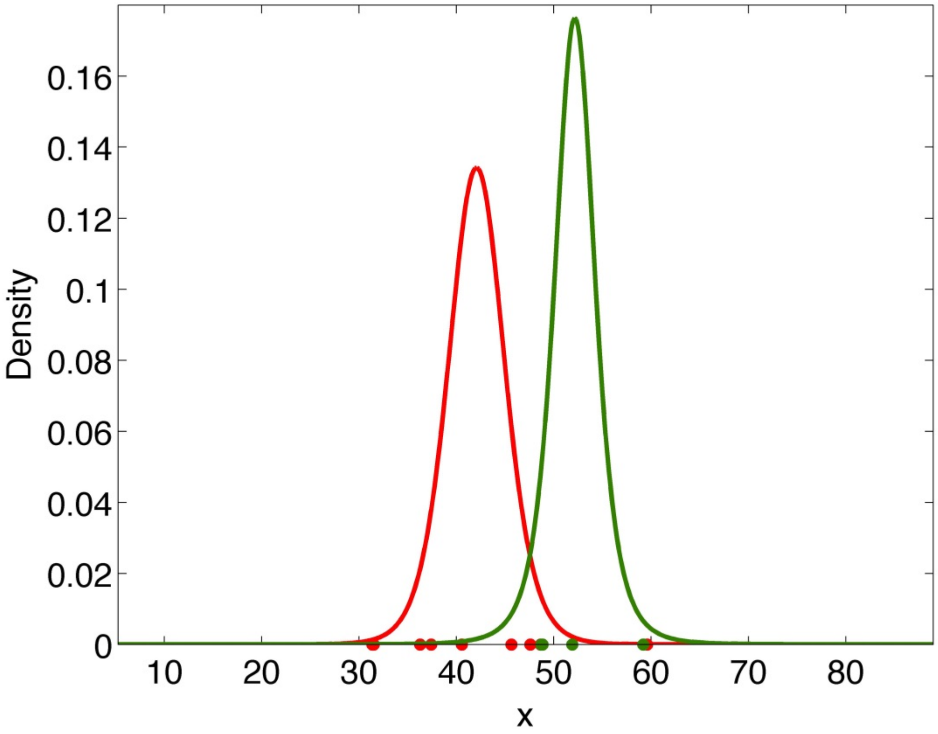
- As promised, this is really simple: **all you need is $\boldsymbol{n}$, the number of datapoints; $\boldsymbol{\bar{x}}$, their mean; and $\boldsymbol{s^2}$, their variance**.

- This means that only the data influence the posterior and all influence from the parameters of the prior has been eliminated. **This is normally not what you want.** The prior contains important information that regularizes your inferences. Often, inference only works with informative priors.

- In any case, the uninformative limit should only ever be taken **after** the calculation of the posterior using a proper prior.

A simple example of Bayesian inference

Integrating out the nuisance parameter λ gives rise to a t-distribution:



A simple example of Bayesian inference

The joint posterior $p(\mu_A, \mu_B|\{x_i\}_A, \{x_k\}_B)$ is simply the product of our two independent posteriors $p(\mu_A|\{x_i\}_A)$ and $p(\mu_B|\{x_k\}_B)$. It will now give us the answer to our question:

$$p(\mu_B - \mu_A > 3) = \int_{-\infty}^{\infty} \mathrm{d}\mu_A\, p(\mu_A|\{x_i\}_A) \int_{\mu_A+3}^{\infty} \mathrm{d}\mu_B\, p(\mu_B|\{x_k\}_B) = 0.9501$$

Note that the t-test told us that there was «no significant difference» even though there is a >95% probability that the parts from B will last at least 3 hours longer than those from A.

Bayesian inference

The procedure in brief:

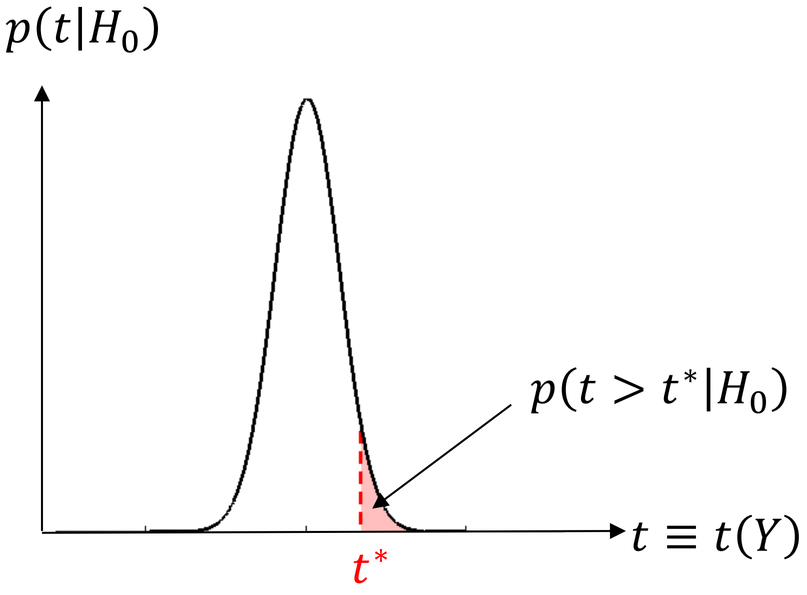
- Determine your question of interest («What is the probability that...?»)
- Specify your model (likelihood and prior)
- Calculate the full posterior using Bayes' theorem
- [Pass to the uninformative limit in the parameters of your prior]
- Integrate out any nuisance parameters
- Ask your question of interest of the posterior

All you need is the rules of probability theory.

Frequentist (or: orthodox, classical) versus Bayesian inference: hypothesis testing

Classical

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

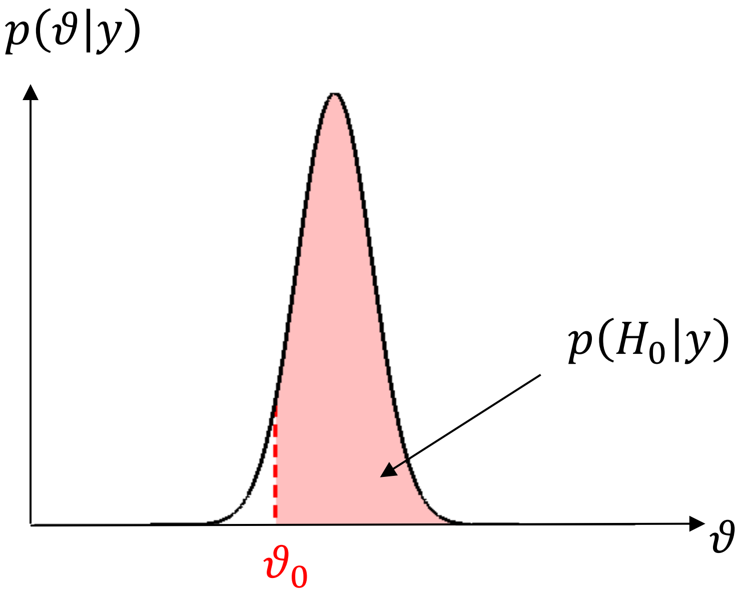


- estimate parameters (obtain test stat. t^*)
- apply decision rule, i.e.:

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

Bayesian

- invert model (obtain posterior pdf)

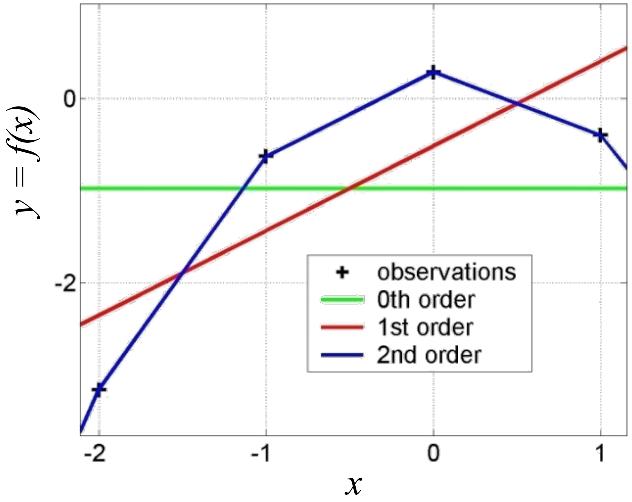


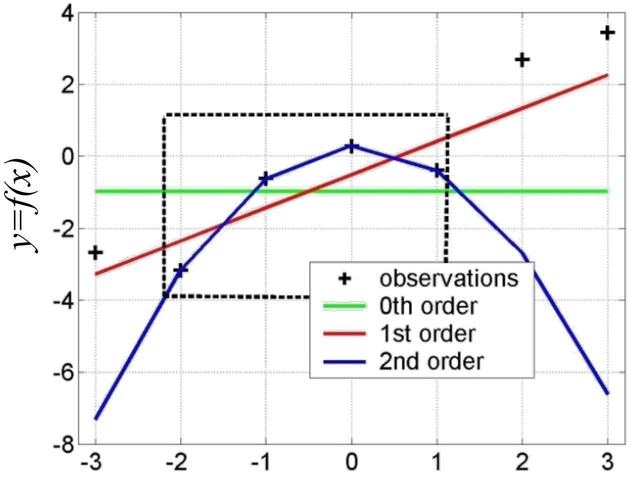
- define the null, e.g.: $H_0: \vartheta > \vartheta_0$
- apply decision rule, i.e.:

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

Model comparison: general principles

- *Principle of parsimony*: «plurality should not be assumed without necessity»
- Automatically enforced by Bayesian model comparison



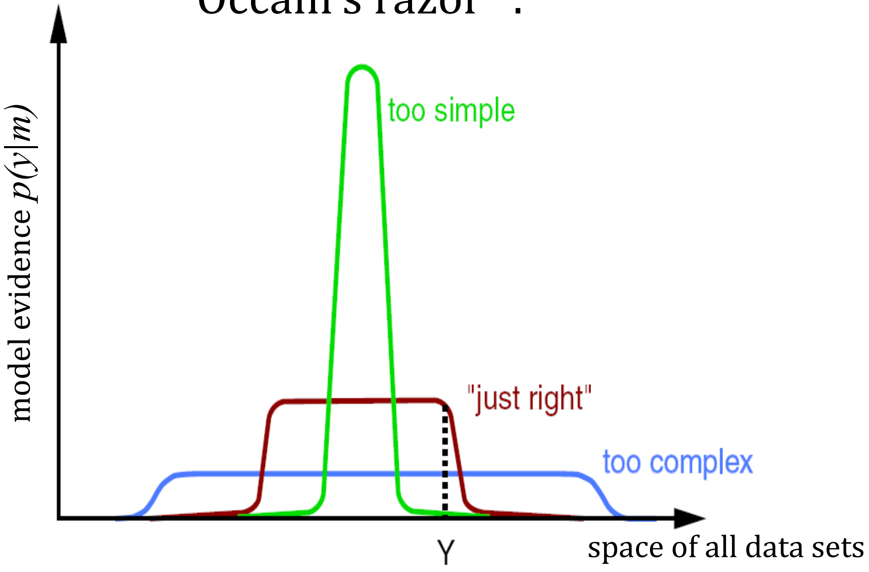


Model evidence:

$$p(y|m) = \int p(y|\vartheta, m)p(\vartheta|m)\mathrm{d}\vartheta$$

$$\approx \exp(accuracy \ - complexity)$$

“Occam’s razor” :



Model comparison: negative variational free energy *F*

$$\textbf{log} - \textbf{model evidence} := \log\ p(y|m)$$

$$= \log \int p(y,\vartheta|m)\mathrm{d}\vartheta$$

sum rule

$$= \log \int q(\vartheta)\frac{p(y,\vartheta|m)}{q(\vartheta)}\mathrm{d}\vartheta$$

multiply by $1 = \frac{q(\vartheta)}{q(\vartheta)}$

a lower bound on the log-model evidence

$$\geq \int q(\vartheta)\log\frac{p(y,\vartheta|m)}{q(\vartheta)}\mathrm{d}\vartheta$$

Jensen's inequality

$$=: \boldsymbol{F} = \textbf{negative variational free energy}$$

$$F := \int q(\vartheta)\log\frac{p(y,\vartheta|m)}{q(\vartheta)}\mathrm{d}\vartheta$$

Kullback-Leibler divergence

$$= \int q(\vartheta)\log\frac{p(y|\vartheta,m)p(\vartheta|m)}{q(\vartheta)}\mathrm{d}\vartheta$$

product rule

$$= \underbrace{\int q(\vartheta)\log p(y|\vartheta,m)\,\mathrm{d}\vartheta}_{\textbf{Accuracy (expected log}-\textbf{likelihood)}} \quad \underbrace{-KL[q(\vartheta),p(\vartheta|m)]}_{\textbf{Complexity}}$$

Model comparison: *F* in relation to Bayes factors, AIC, BIC

$$\textbf{Bayes factor} := \frac{p(y|m_1)}{p(y|m_0)} = \exp\left(\log\frac{p(y|m_1)}{p(y|m_0)}\right) = \exp(\log p(y|m_1) - \log p(y|m_0))$$

$$\approx \exp(F_1 - F_0)$$

[Meaning of the Bayes factor: $\overbrace{\frac{p(m_1|y)}{p(m_0|y)}}^{\text{Posterior odds}} = \overbrace{\frac{p(y|m_1)}{p(y|m_0)}}^{\text{Bayes factor}} \overbrace{\frac{p(m_1)}{p(m_0)}}^{\text{Prior odds}}$]

$$\boldsymbol{F} = \int q(\vartheta)\log p(y|\vartheta,m)\,\mathrm{d}\vartheta - KL[q(\vartheta),p(\vartheta|m)]$$

$$= \text{Accuracy} - \text{Complexity}$$

$$\textbf{AIC} := \text{Accuracy} - \underbrace{p}_{\text{Number of parameters}}$$

$$\textbf{BIC} := \text{Accuracy} - \frac{p}{2}\log \underbrace{N}_{\text{Number of data points}}$$

A note on informative priors

- Any model consists of two parts: likelihood and prior.
- The choice of likelihood requires as much justification as the choice of prior because it is just as «subjective» as that of the prior.
- The data never speak for themselves. They only acquire meaning when seen through the lens of a model. However, this does not mean that all is subjective because models differ in their validity.
- In this light, the widespread concern that informative priors might bias results (while the form of the likelihood is taken as a matter of course requiring no justification) is misplaced.
- Informative priors are an important tool and their use can be justified by establishing the validity (face, construct, and predictive) of the resulting model as well as by model comparison.

A note on uninformative priors

- Using a flat or «uninformative» prior doesn't make you more «data-driven» than anybody else. It's a choice that requires just as much justification as any other.

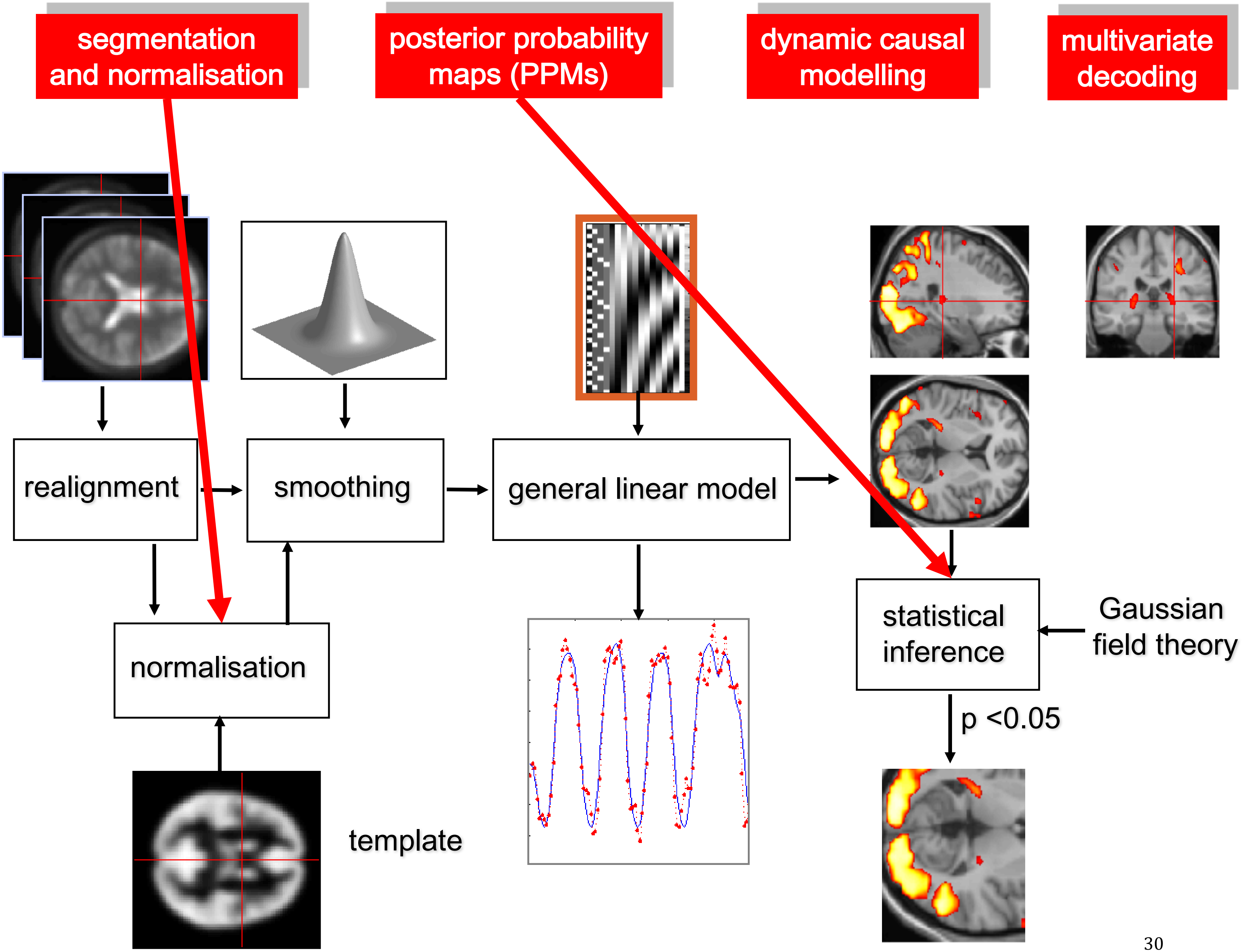
- For example, if you're studying a small effect in a noisy setting, using a flat prior means assigning the same prior probability mass to the interval covering effect sizes -1 to +1 as to that covering effect sizes +999 to +1001.

- Far from being unbiased, this amounts to a bias in favor of implausibly large effect sizes. Using flat priors is asking for a replicability crisis.

- One way to address this is to collect enough data to swamp the inappropriate priors. A cheaper way is to use more appropriate priors.

- Disclaimer: if you look at my papers, you will find flat priors.

Applications of Bayesian inference



segmentation and normalisation

posterior probability maps (PPMs)

dynamic causal modelling

multivariate decoding

realignment

smoothing

general linear model

normalisation

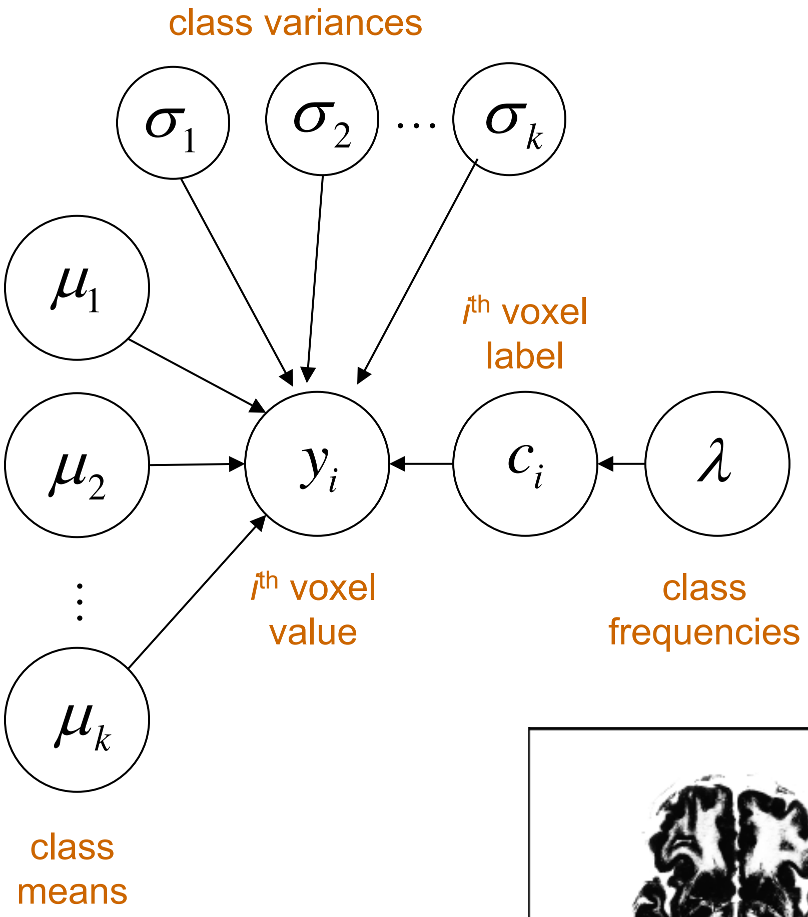
template

statistical inference

Gaussian field theory

$p < 0.05$

Segmentation (mixture of Gaussians-model)



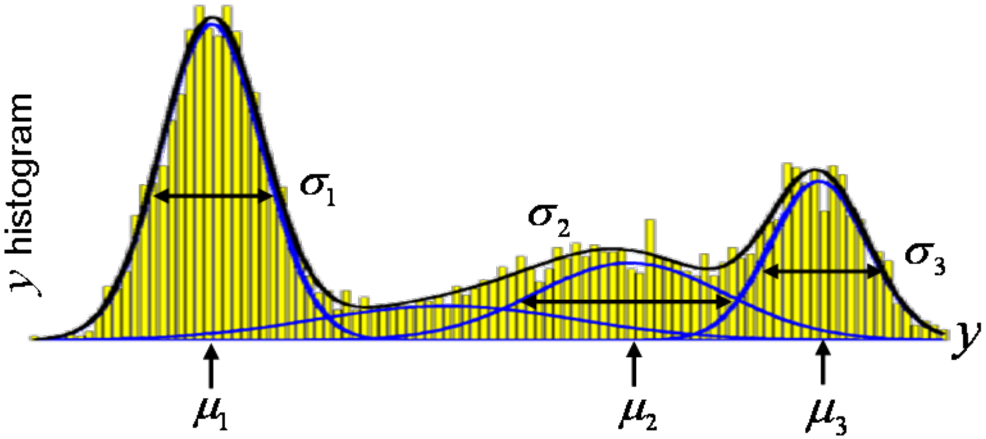
class variances

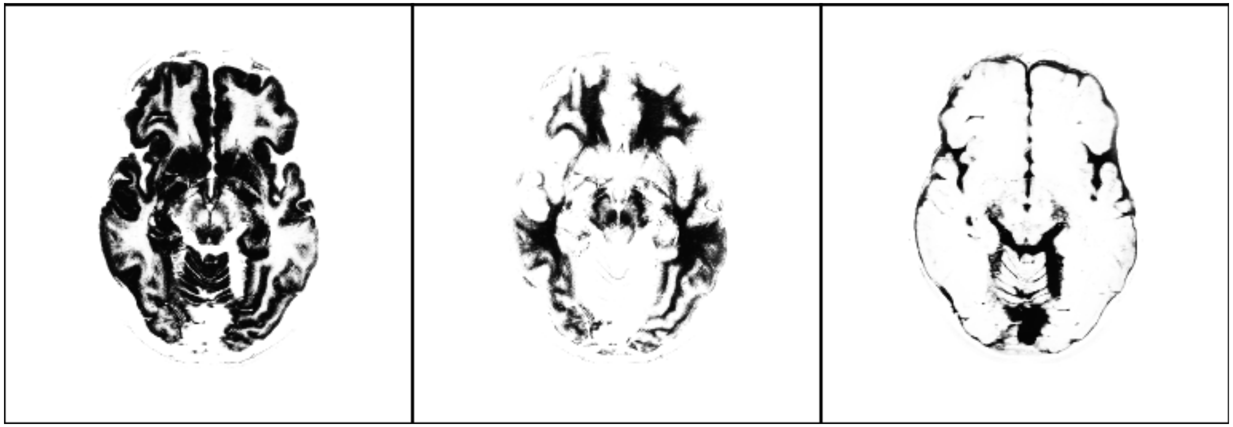
i^{th} voxel label

i^{th} voxel value

class frequencies

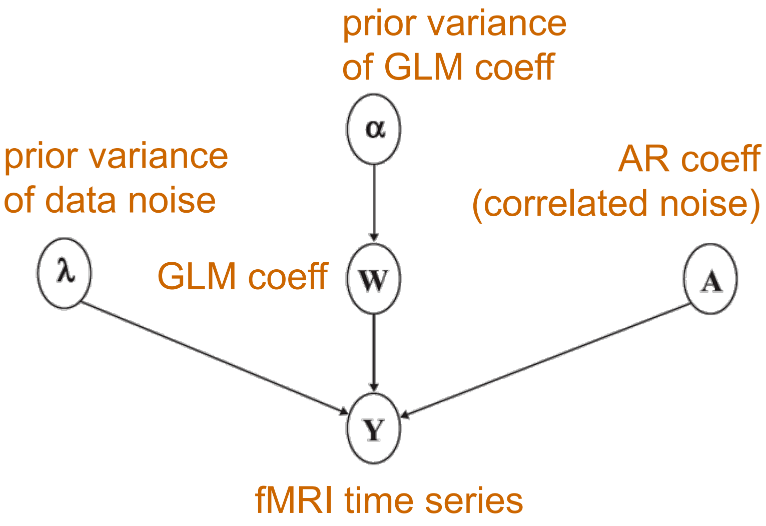
class means





grey matter white matter CSF

fMRI time series analysis



prior variance
of GLM coeff

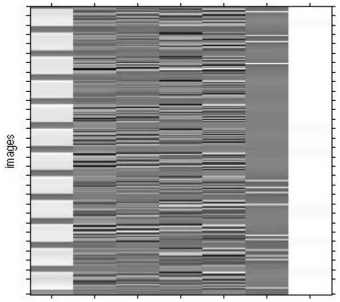
prior variance
of data noise

AR coeff
(correlated noise)

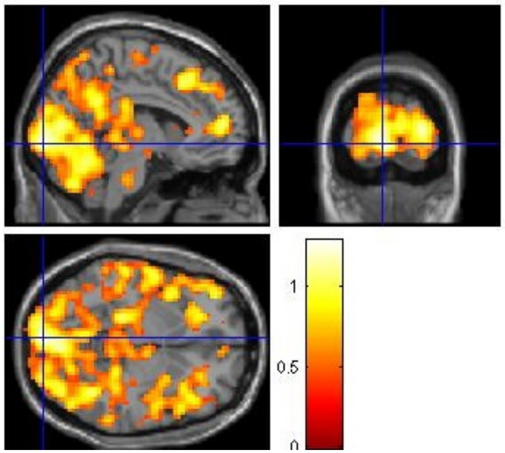
GLM coeff

fMRI time series

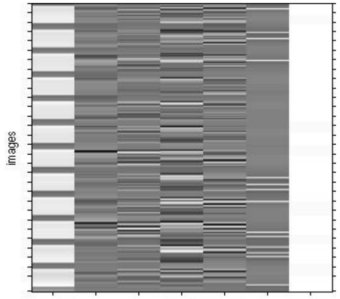
short-term memory
design matrix (X)



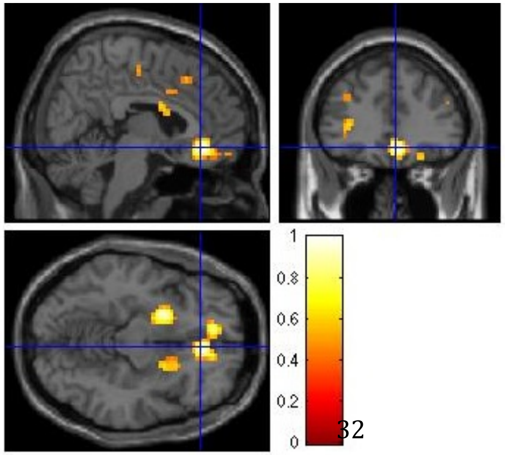
PPM: regions best explained
by short-term memory model



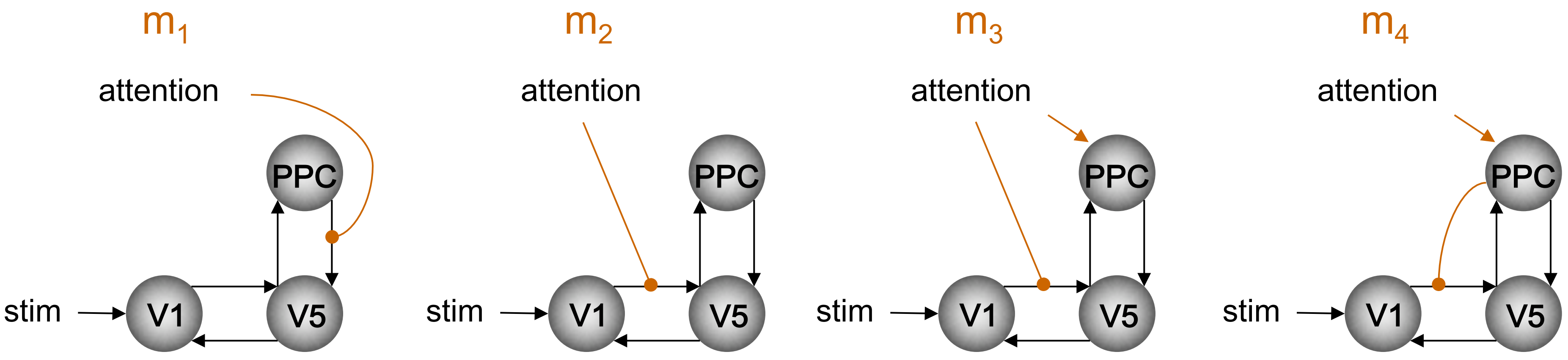
long-term memory
design matrix (X)

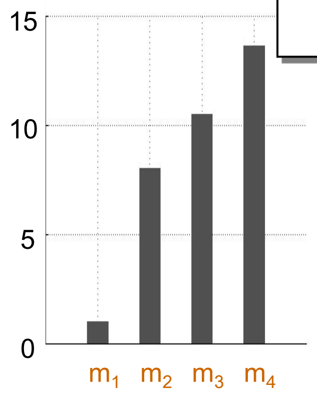


PPM: regions best explained
by long-term memory model

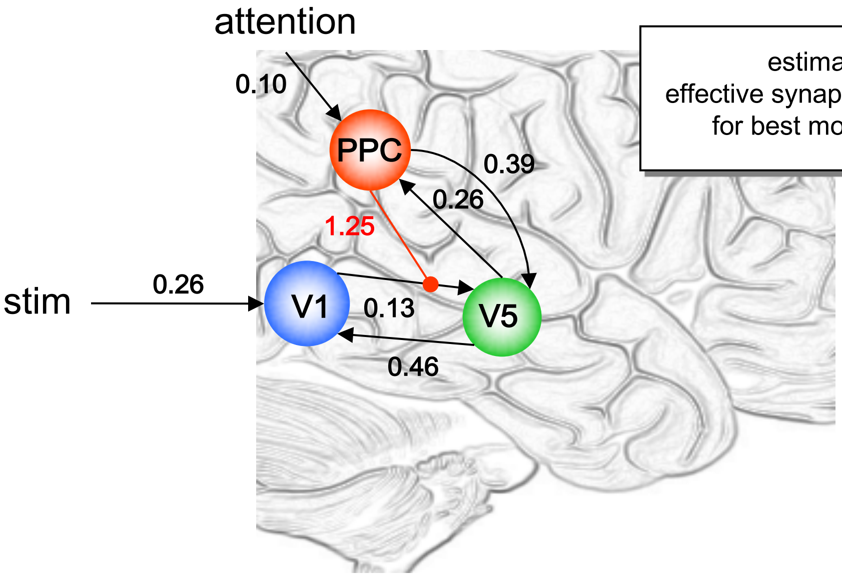


Dynamic causal modeling (DCM)



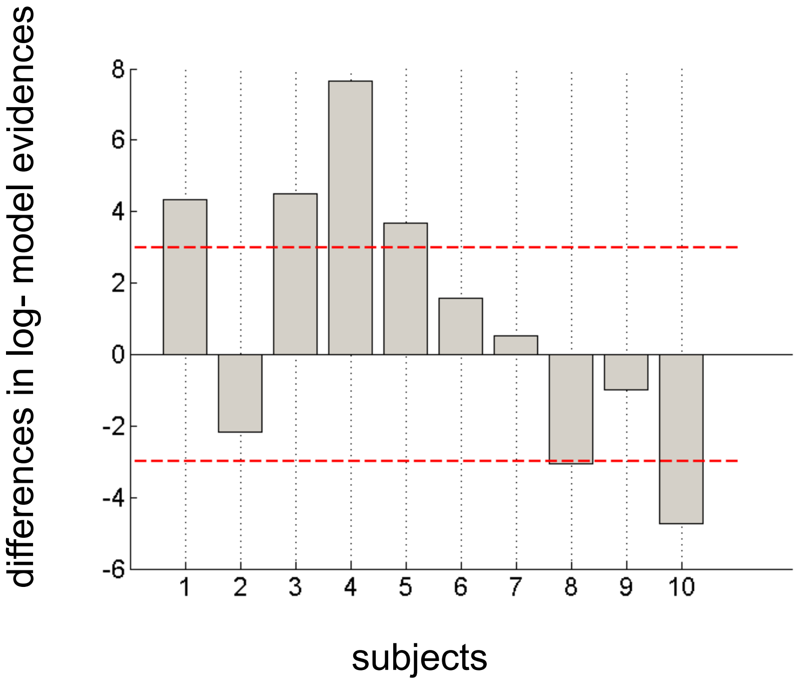


models marginal likelihood
$\ln p(y|m)$

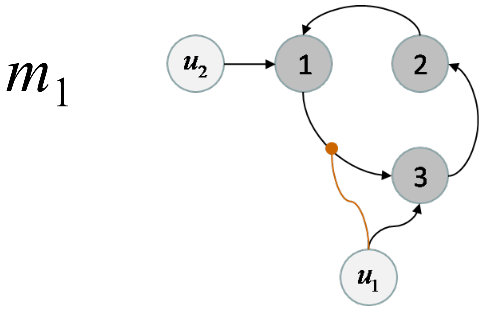


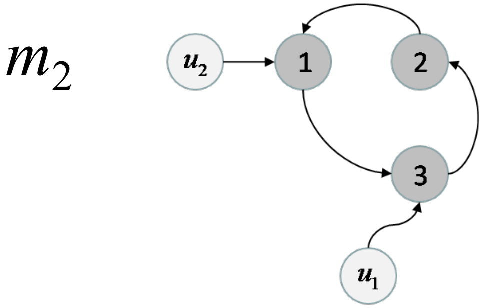
estimated
effective synaptic strengths
for best model (m_4)

Model comparison for group studies



$$\ln p\left(y|m_1\right) - \ln p\left(y|m_2\right)$$





Fixed effect

Assume all subjects correspond to the same model

Random effect

Assume different subjects might correspond to different models

Thanks