Bayesian Inference

Chris Mathys

London SPM Course

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

A surprising piece of information

BBC Sign in News Sport Weather iPlayer TV Ra

NEWS MAGAZINE

Home | World | UK | England | N. Ireland | Scotland | Wales | Business | Politics | Health | Education | Sci/E

Video & Audio | Magazine | Editors' Blog | In Pictures | Also in the News | Have Your Say | Special Repor

19 November 2012 Last updated at 18:19

44K Share

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 surprising piece of information

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

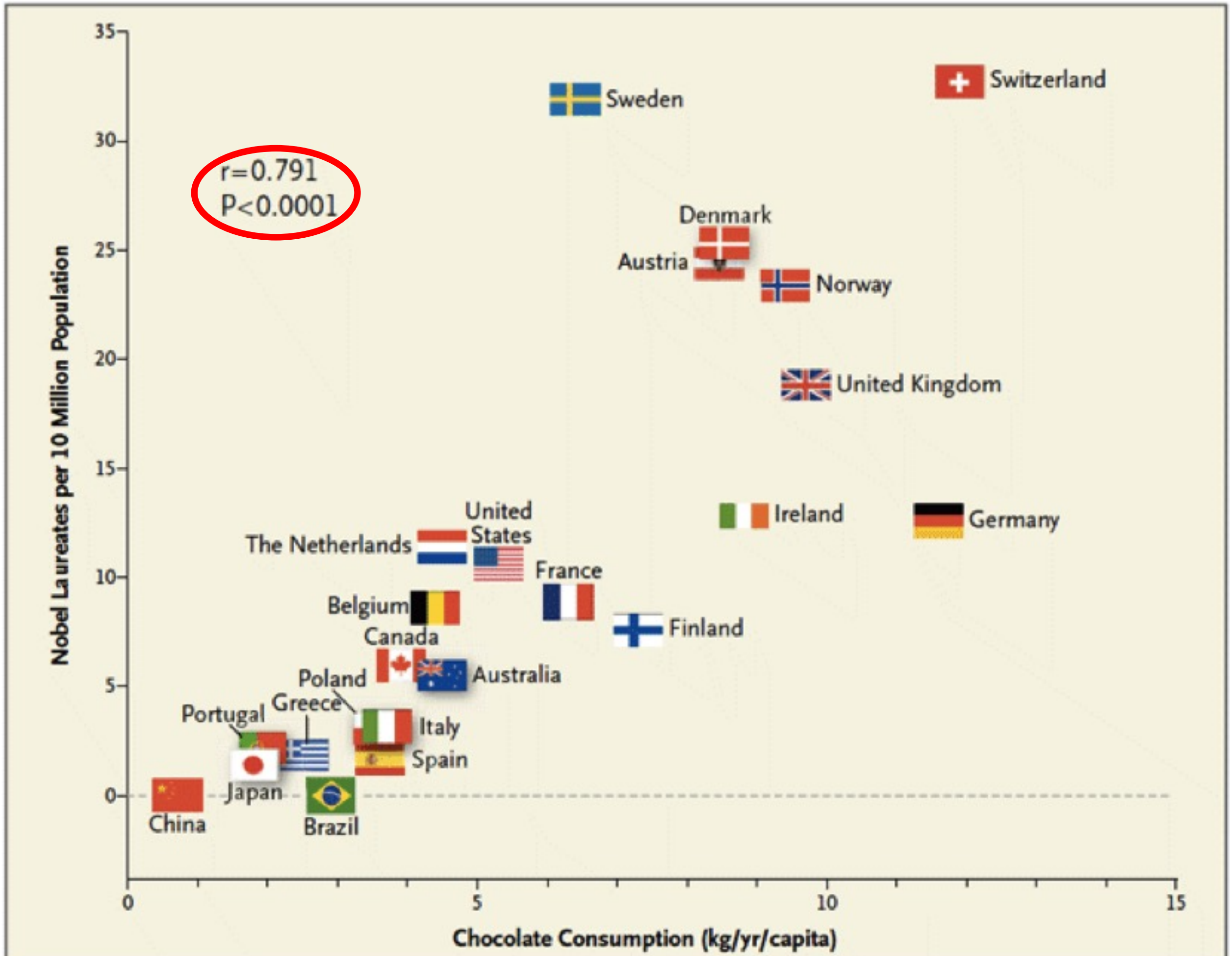


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

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*****.**

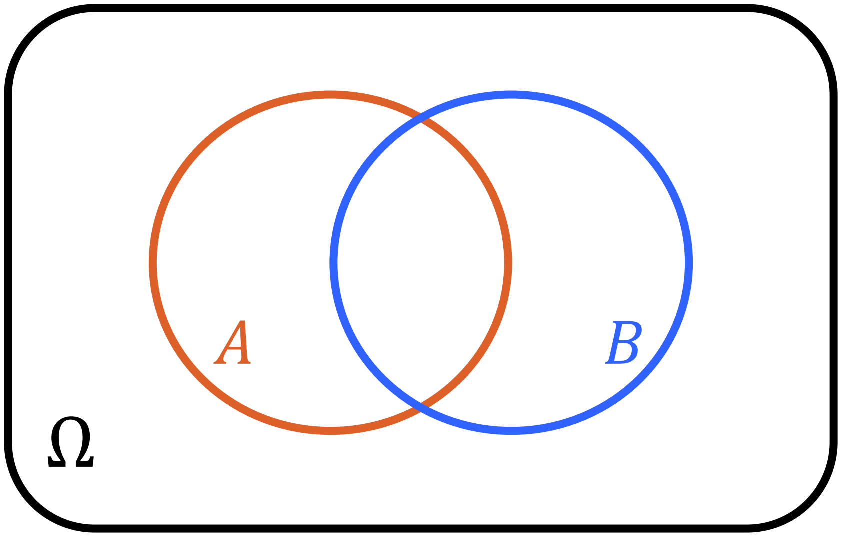
Our question here can be rephrased in terms of a conditional probability:

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

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

Calculating with probabilities: the setup

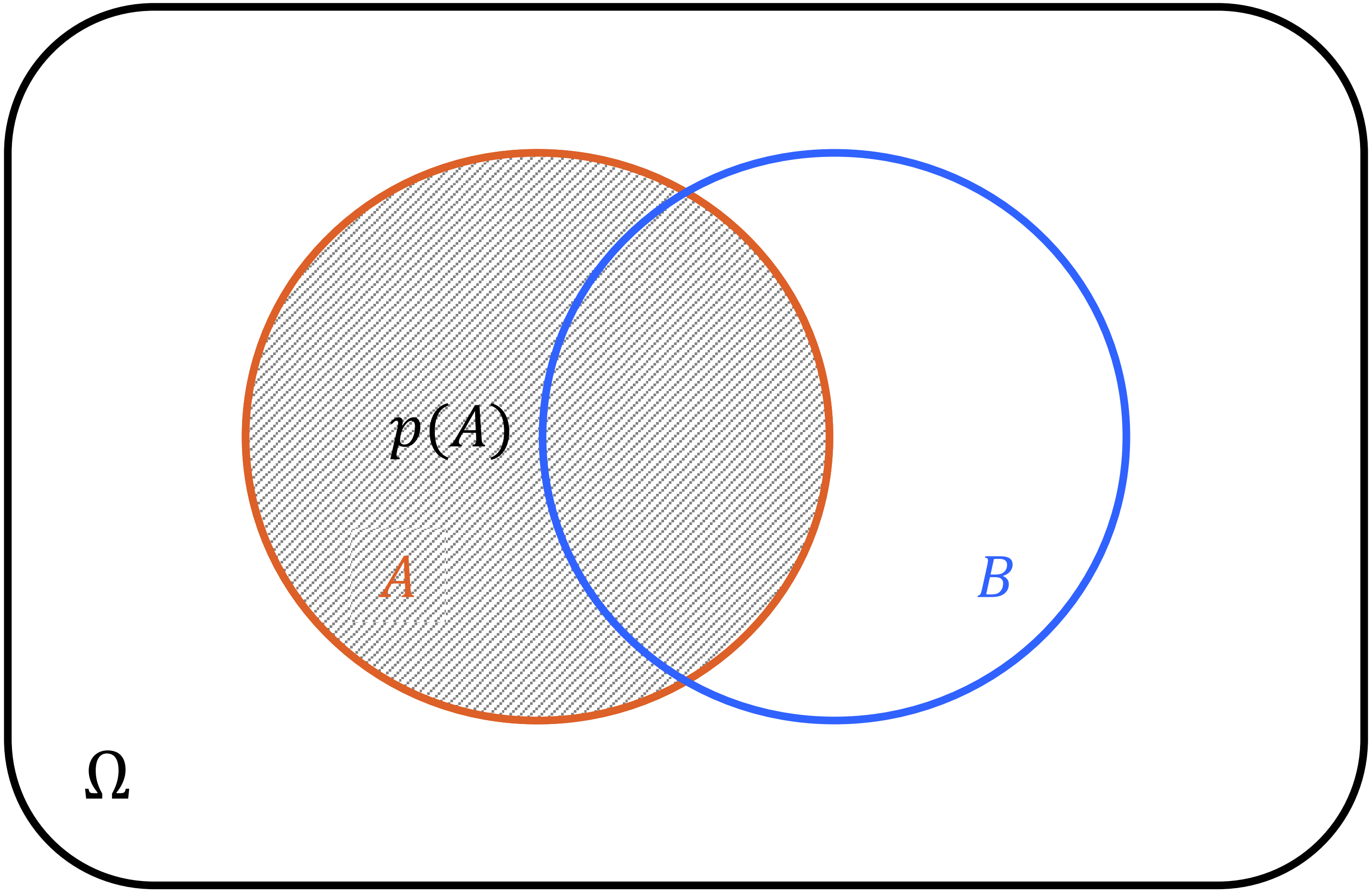
We assume a probability space Ω with subsets A and B



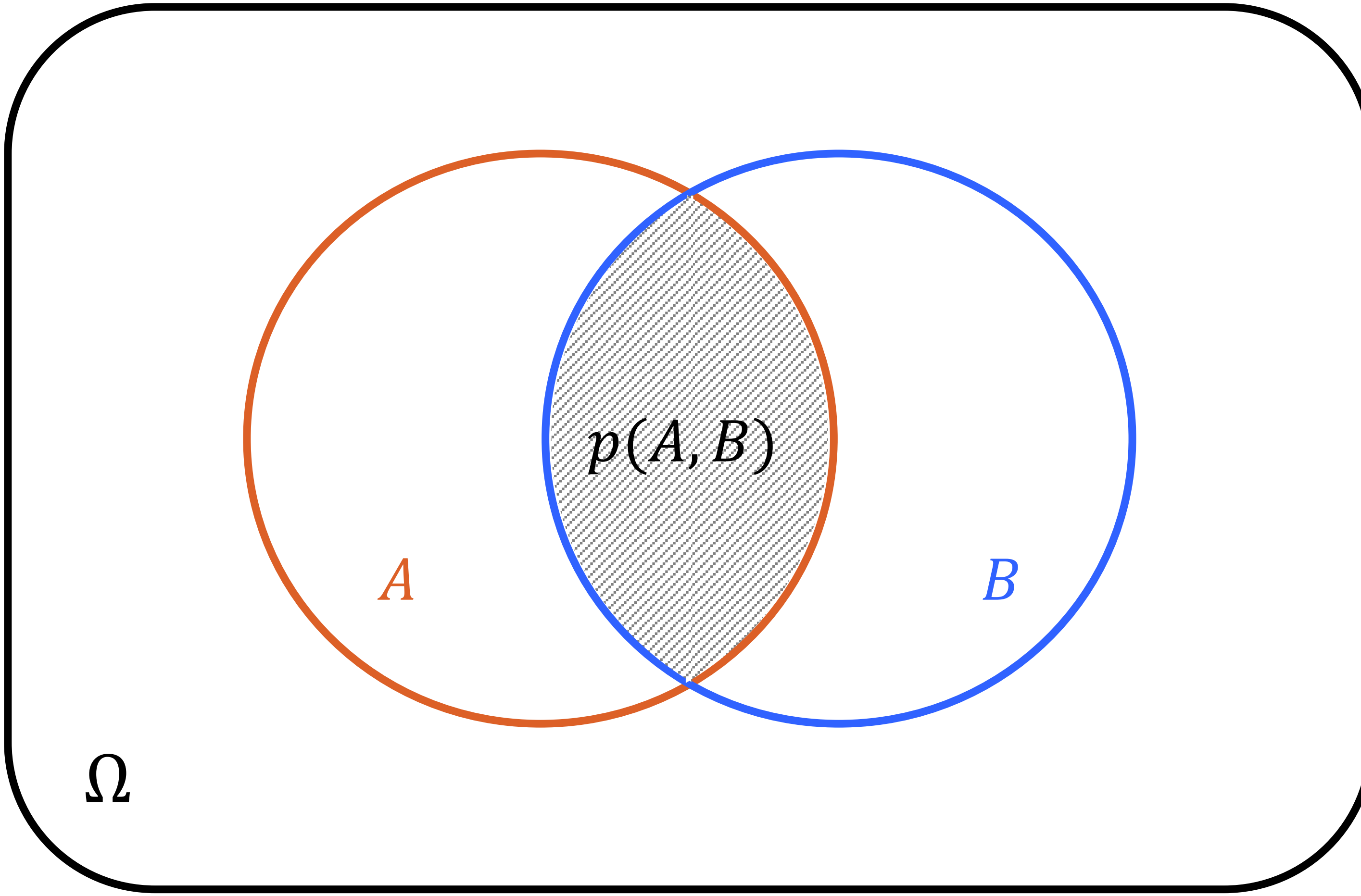
In order to understand *the rules of probability*, we need to understand **three kinds of probabilities**

- *Marginal* probabilities like $p(A)$
- *Joint* probabilities like $p(A, B)$
- *Conditional* probabilities like $p(B|A)$

Marginal probabilities



Joint probabilities



What is ‘marginal’ about marginal probabilities?

- Let A be the statement ‘the sun is shining’
- Let B be the statement ‘it is raining’
- $\bar{A}$ negates A, $\bar{B}$ negates B

Consider the following table of joint probabilities:

	B	$\bar{B}$	Marginal probabilities
A	$p(A, B) = 0.1$	$p(A, \bar{B}) = 0.5$	$p(A) = 0.6$
$\bar{A}$	$p(\bar{A}, B) = 0.2$	$p(\bar{A}, \bar{B}) = 0.2$	$p(\bar{A}\,) = 0.4$
Marginal probabilities	$p(B) = 0.3$	$p(\bar{B}) = 0.7$	Sum of all probabilities $\sum p(\cdot,\cdot) = 1$

Marginal probabilities get their name from being at the margins of tables such as this one.

Conditional probabilities

- In the previous example, what is the probability that the sun is shining given that it is not raining?
- This question refers to a conditional probability: $p(A|\bar{B})$
- You can find the answer by asking yourself: out of all times where it is not raining, which proportion of times will the sun be shining?

	B	$\bar{B}$	Marginal probabilities
A	$p(A,B) = 0.1$	$p(A,\bar{B}) = 0.5$	$p(A) = 0.6$
$\bar{A}$	$p(\bar{A},B) = 0.2$	$p(\bar{A},\bar{B}) = 0.2$	$p(\bar{A}\,) = 0.4$
Marginal probabilities	$p(B) = 0.3$	$p(\bar{B}) = 0.7$	Sum of all probabilities $\sum p(\cdot,\cdot) = 1$

- This means we have to divide the joint probability of 'sun shining, not raining' by the sum of all joint probabilities where it is not raining:

$$p(A|\bar{B}) = \frac{p(A,\bar{B})}{p(A,\bar{B}) + p(\bar{A},\bar{B})} = \frac{p(A,\bar{B})}{p(\bar{B})} = \frac{0.5}{0.7} \approx 0.71$$

The rules of probability

Considerations like the ones above led to the following definition of the **rules of probability:**

1. $\sum_a p(a) = 1$ (*Normalization*)

2. $p(B) = \sum_a p(a, B)$ (*Marginalization* – the **sum rule**)

3. $p(A, B) = p(A|B)p(B) = p(B|A)p(A)$ (*Conditioning* – the **product rule**)

These are **axioms**, ie they are assumed to be true. Therefore, we cannot test them the way we could test a theory. However, we can see if they turn out to be useful.

Bayes' rule

- The product rule of probability states that

$$p(A|B)p(B) = p(B|A)p(A)$$

- If we divide by $p(B)$, we get **Bayes' rule:**

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} = \frac{p(B|A)p(A)}{\sum_a p(B|a)p(a)}$$

- The last equality comes from unpacking $p(B)$ according to the product and sum rules:

$$p(B) = \sum_a p(B, a) = \sum_a p(B|a)p(a)$$

Bayes' rule: what problem does it solve?

- Why is Bayes' rule important?
- It allows us to invert conditional probabilities, ie to pass from $p(B|A)$ to $p(A|B)$:

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

- In other words, it allows us to update our belief about A in light of observation B

Bayes' rule and 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' rule:

likelihood model

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

prior

posterior

evidence

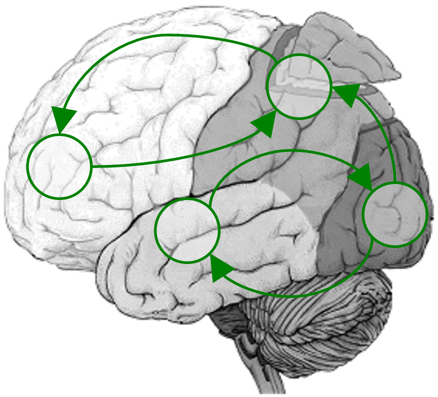
Inference on the quantities of interest in neuroimaging studies has exactly the same general structure.

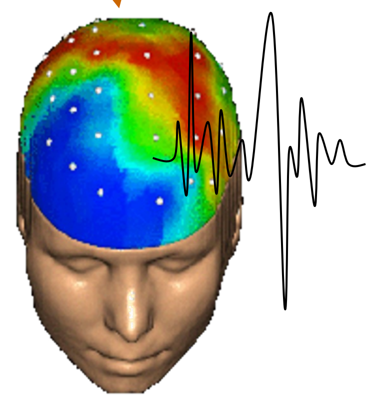
Inference in SPM

forward problem

$p(y|\vartheta, m)$

likelihood



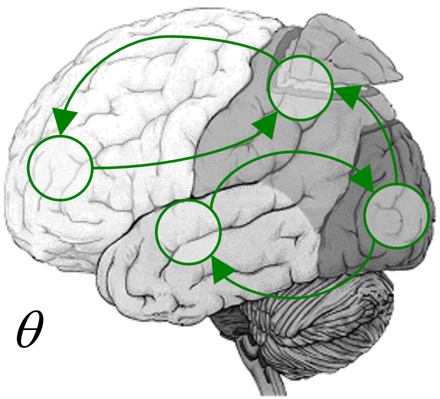


posterior distribution

$p(\vartheta|y, m)$

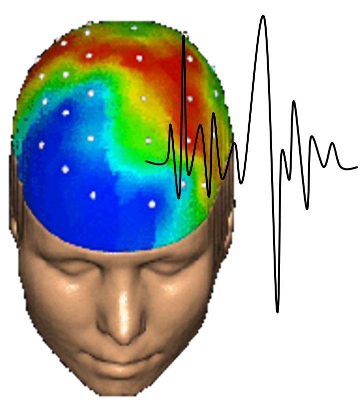
inverse problem

Inference in SPM



θ

generative model m

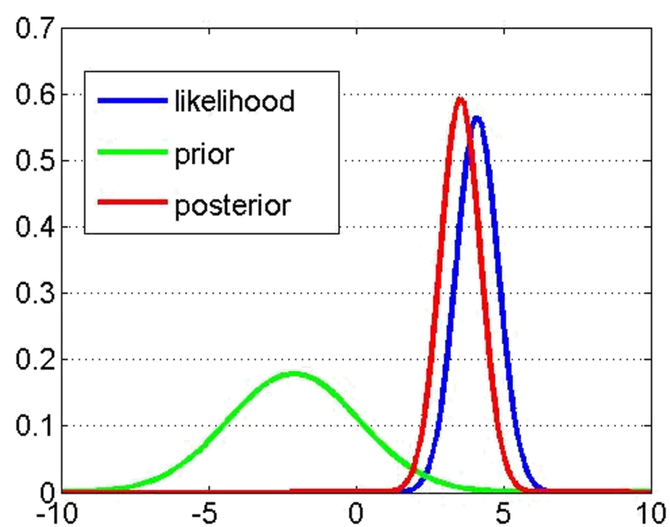


y

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))

This example comes with its own interactive Jupyter notebook:

https://github.com/chmathys/bayesian-inference-example

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

A:	B:
59.5814	48.8506
37.3953	48.7296
47.5956	59.1971
40.5607	51.8895
48.6468	
36.2789	
31.5110	
31.3606	
45.6517	

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

A simple example of Bayesian inference

First: how *not* to analyze these data – an illustration of the dangers of blindly applying recipes

- Let's do a *t*-test (but first, let's compare variances with an *F*-test):

```
>> [fh,fp,fci,fstats] = vartest2(xa,xb)

fh =          fp =          fci =           fstats =

     0           0.3297        0.2415          fstat: 3.5114
                              19.0173            df1: 8
                                                 df2: 3
```

Variances not significantly different!

```
>> [h, p, ci, stats]= ttest2(xa,xb)

h =           p =           ci =            stats =

     0           0.0665       -21.0191          tstat: -2.0367
                                0.8151             df: 11
                                                   sd: 8.2541
```

Means not significantly different!

Is this satisfactory? No, so what can we learn by turning to probability theory (i.e., Bayesian inference)?

A simple example of Bayesian inference

How to go about it:

- Determine your **question of interest** («What is the probability that...?»)
- **Specify** your model (likelihood and prior)
- **Justify** your model from first principles and/or **prior predictive simulation**
- Determine the **posterior distribution**
- Answer your question using **posterior predictive simulation**

All of this is illustrated in detail in the notebook:

https://github.com/chmathys/bayesian-inference-example

A simple example of Bayesian inference

The model:

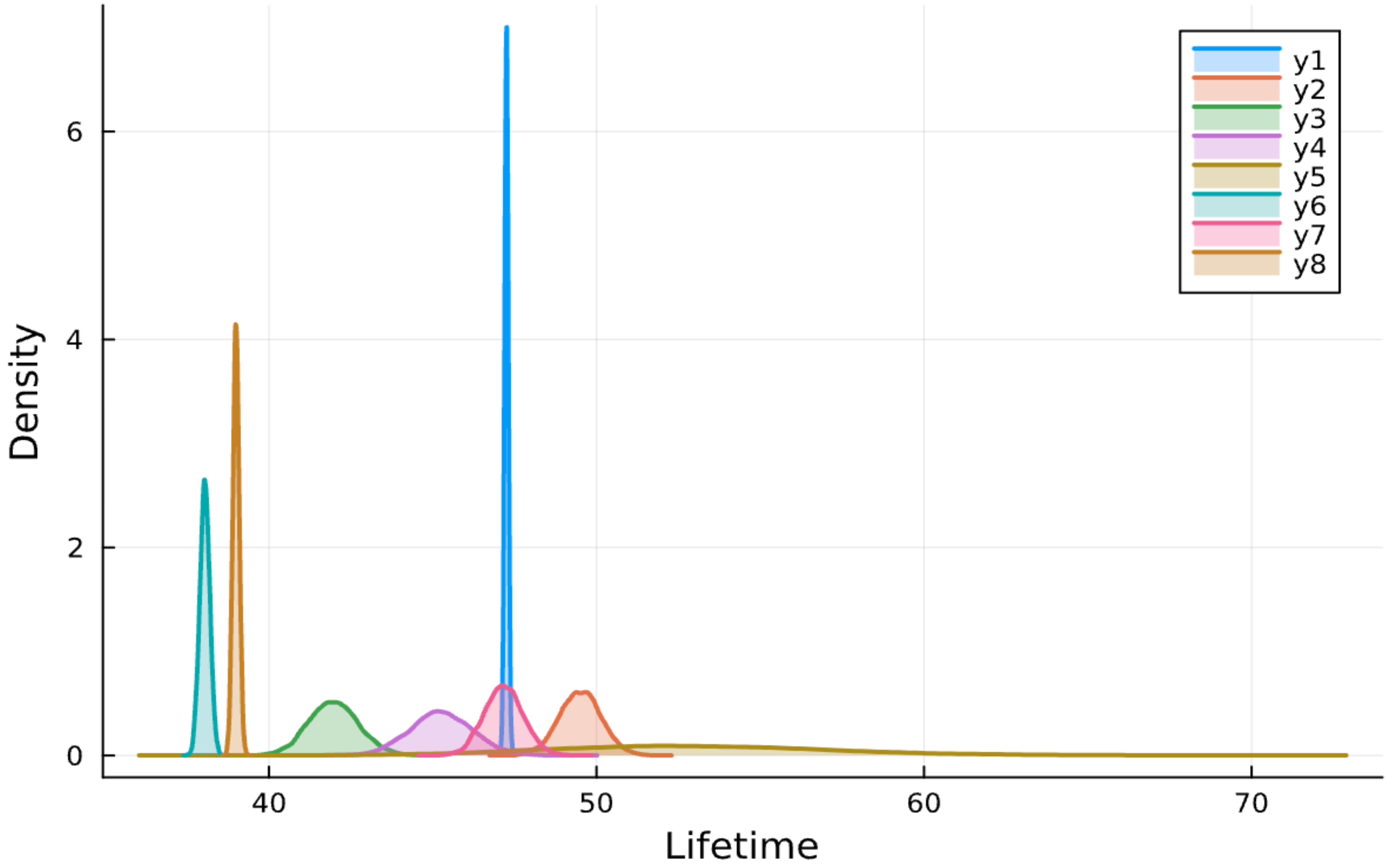
```julia
@model function gaussians(y, c, α_μ = 0, α_σ = 1, θ = 1)
    # Number of categories
    nc = length(unique(c))

    # Priors
    α ~ filldist(Normal(α_μ, α_σ), nc)
    σ ~ filldist(Exponential(θ), nc)

    # Observations
    # y .~ Normal.(α[c], σ[c])
    # The above works for inference, but not for predictive sampling.
    # For that to work, we need to use a loop.
    for i in eachindex(y)
        y[i] ~ Normal(α[c[i]], σ[c[i]])
    end
end
```

A simple example of Bayesian inference

Prior predictive simulation:



A simple example of Bayesian inference

After fitting the model to the data, we can do inference on means of lifetimes (as does the *t*-test):

```
# Probability that *median* lifetime from B is more than 3 hours greater
sum(mean_b - mean_a .> 3) / length(posterior_sample)

0.9515
```

The *t*-test recipe said that the difference of means was not significant, but probability theory (i.e., Bayesian inference) says that, under plausible assumptions, **there's a 95% probability that the median lifetime of parts from B is at least 3 days longer!**

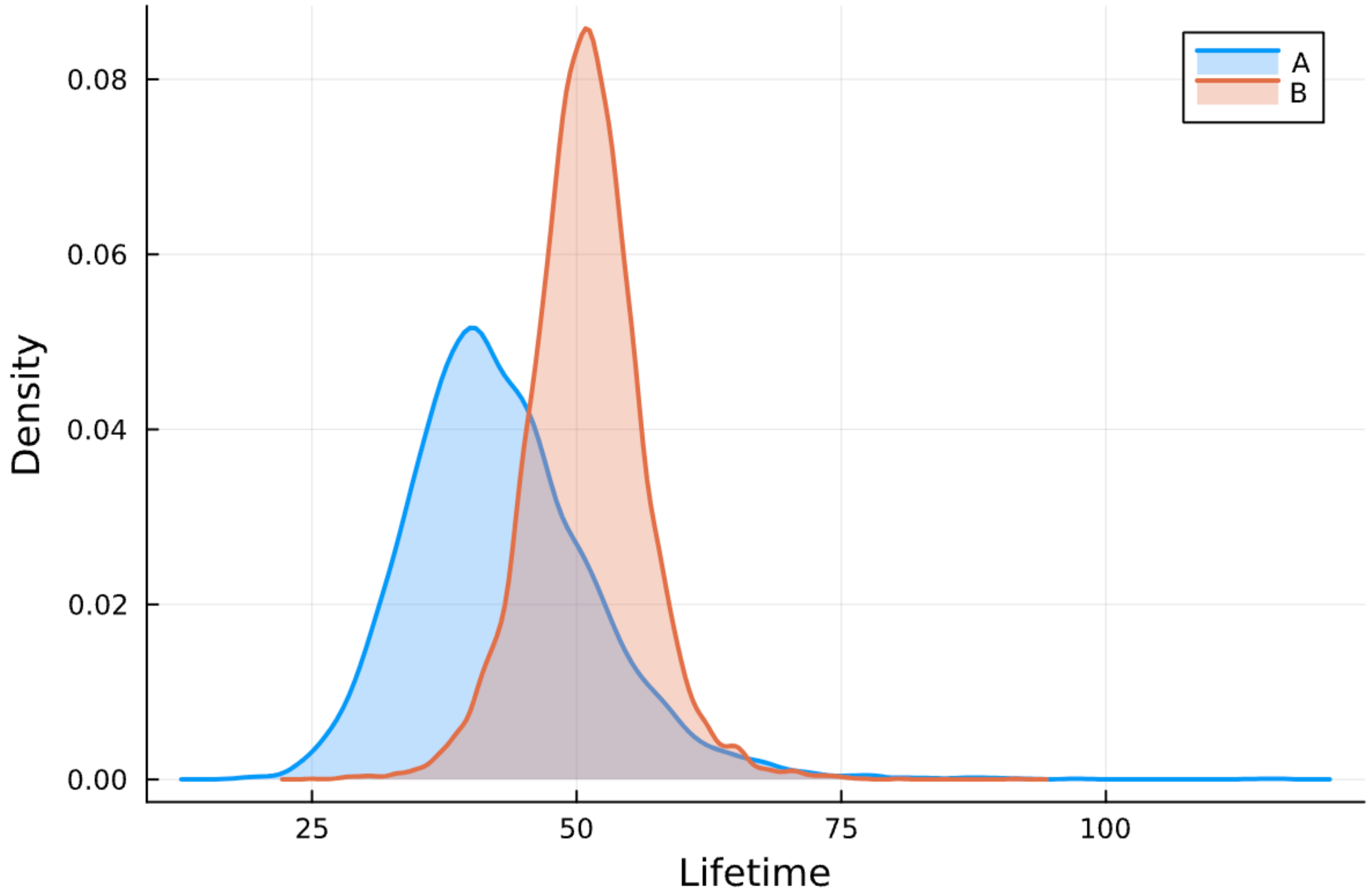
A simple example of Bayesian inference

The real 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 50%, I will choose those from B.**
- To determine this, we need to look at the **posterior predictive distribution**

A simple example of Bayesian inference

Posterior predictive simulation:



A simple example of Bayesian inference

Inference on lifetimes (answer to the real question):

```
# Probability that when randomly choosing a part from each manufacturer,
# the lifetime of that from B is more than 3 hours greater
sum(t_b - t_a .> 3) / length(posterior_predictive_sample)
```

```
0.713125
```

So our decision rule says: buy from B. (But we could have chosen another decision rule, and neither the data nor statistical procedures can give us decision rules. We have to reason about the real world to get them.)

Bayes' rule for odds

- The *odds* of A relate to the *probability* of A in the following way

$$o(A) = \frac{p(A)}{p(\bar{A})} = \frac{p(A)}{1 - p(A)}$$

$$p(A) = \frac{o(A)}{1 + o(A)}$$

- Bookmakers offer odds *against* events. For example, odds of 3:1 on a horse imply a probability of $\frac{3}{3+1} = 0.75$ for the horse *not* to win, ie a probability of $1 - 0.75 = 0.25$ for the horse to win.

Bayes' rule for odds

- In terms of odds, Bayes rule is

$$o(H|y) = \frac{p(H|y)}{p(\bar{H}|y)} = \frac{\frac{p(y|H)p(H)}{p(y)}}{\frac{p(y|\bar{H})p(\bar{H})}{p(y)}} = \frac{p(y|H)}{p(y|\bar{H})}\frac{p(H)}{p(\bar{H})} = \frac{p(y|H)}{p(y|\bar{H})}o(H)$$

- In sum:

$$\underbrace{o(H|y)}_{\substack{\textbf{posterior}\\\textbf{odds}}} = \underbrace{\frac{p(y|H)}{p(y|\bar{H})}}_{\substack{\textbf{likelihood}\\\textbf{ratio}}} \underbrace{o(H)}_{\substack{\textbf{prior}\\\textbf{odds}}}$$

- The *likelihood ratio* is sometimes called the ***Bayes factor***. This is because multiplying the prior odds with this factor gives the posterior odds.
- The Bayes factor is a measure for how much making observation y favours hypothesis H over hypothesis $\bar{H}$.

Model comparison

- The fact that the Bayes factor is a measure of strength of evidence can be used for model comparison

- Consider hypotheses (i.e., models) H_0 and H_1. Then Bayes' rule for the odds of H_1 over H_0 is

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

- The likelihood ratio is the ratio of **marginal likelihoods** (also called **model evidences**):

$$p(y|H_i) = \int p(y|\vartheta_i, H_i) p(\vartheta_i|H_i) \mathrm{d}\vartheta_i$$

- In terms of **log-model evidences**, the log-Bayes factor is simply the difference

$$\log \frac{p(y|H_1)}{p(y|H_0)} = \log p(y|H_1) - \log p(y|H_0)$$

Model comparison: negative variational free energy F

$$\textbf{log – model evidence} := \log\ p(y|H)$$

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

sum rule

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

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

a lower bound on the log-model evidence

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

Jensen's inequality

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

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

Kullback-Leibler divergence

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

product rule

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

Remarks on model comparison / model selection

- There is a range of scores that help in choosing a well-performing model: AIC (Akaike information criterion), BIC (Bayesian information criterion), Bayes factors, LME (log-model evidence), free energy, etc.
- Each model gets a particular score (which is on its own uninterpretable!)
- The difference in score between models is what counts
- However, model selection is not straightforward. AIC and BIC penalize complexity based on simple heuristics, which may not reflect complexity accurately. LME is better on that count, but is very sensitive to the modeller's choice of priors.
- **The three decisive considerations:**

    1. **Does the model allow me to answer my question of interest?**
    2. **Does the *prior predictive* distribution of observations make sense?**
    3. **Does the *posterior predictive* distribution of observations make sense?**

    When the answer to all three is yes, the model is fine.

A note on uninformative priors

- Using a flat or «uninformative» prior doesn't make lead to inferences that are more «data-driven». It's a modelling 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.

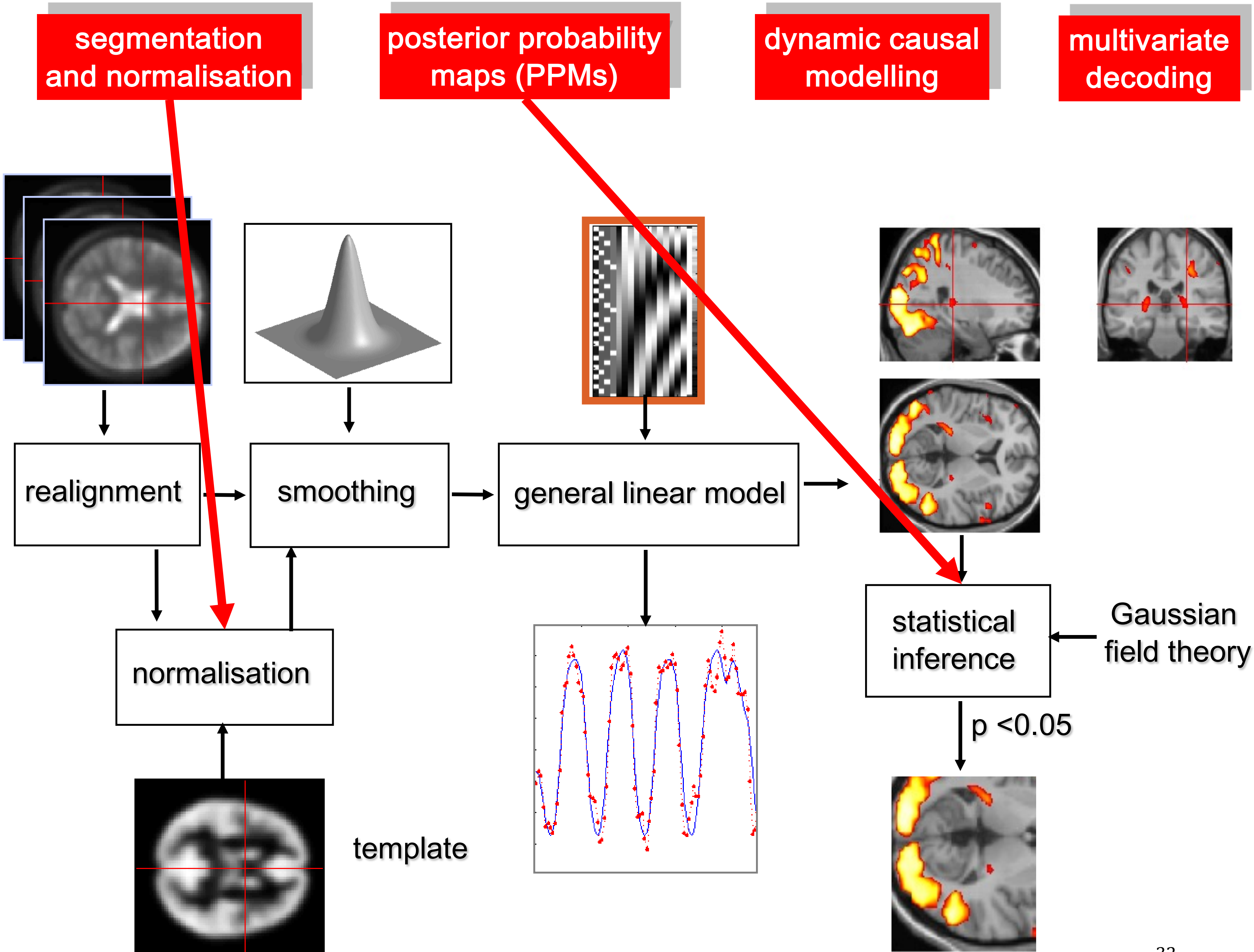
- Far from being unbiased, this amounts to a bias in favor of implausibly large effect sizes. Using flat priors is asking for a replicability crisis.

- Put another way, priors which are too uninformative amount to an implausible prior predictive distribution

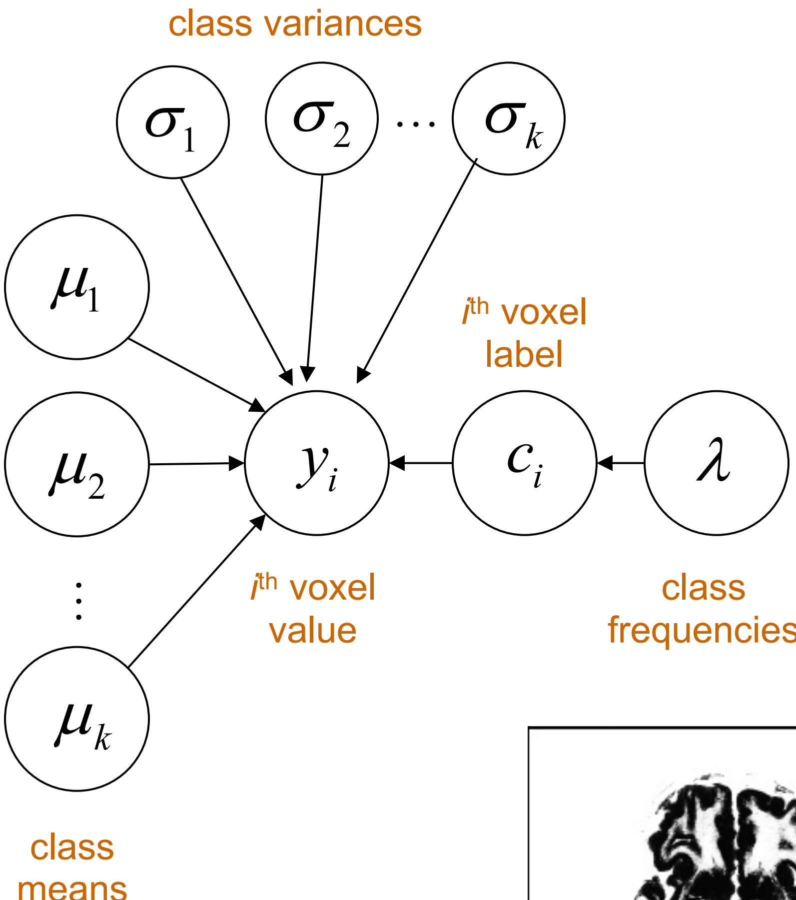
- One way to address this is to collect enough data to swamp the inappropriate priors. A cheaper way is to use more appropriate priors.

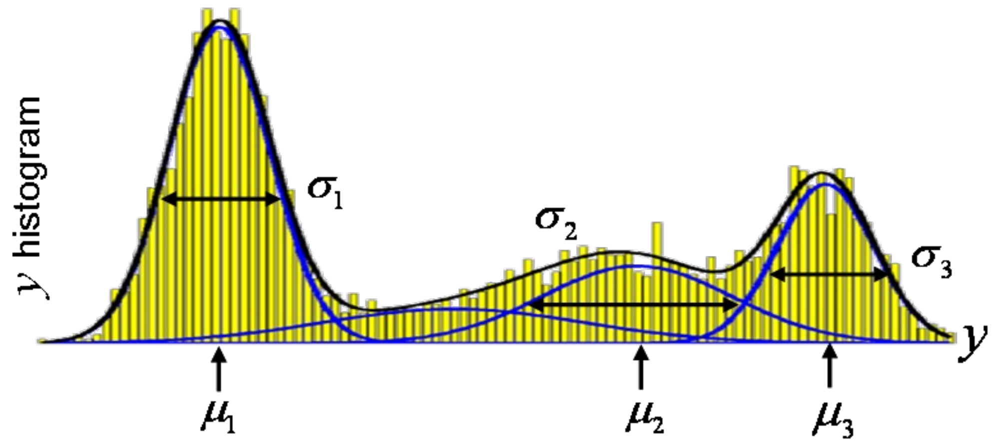
- Classical tests often imply flat priors. But also in a Bayesian context, priors which are too flat are common because they can give a higher model evidence (which is a limitation of the concept of model evidence).

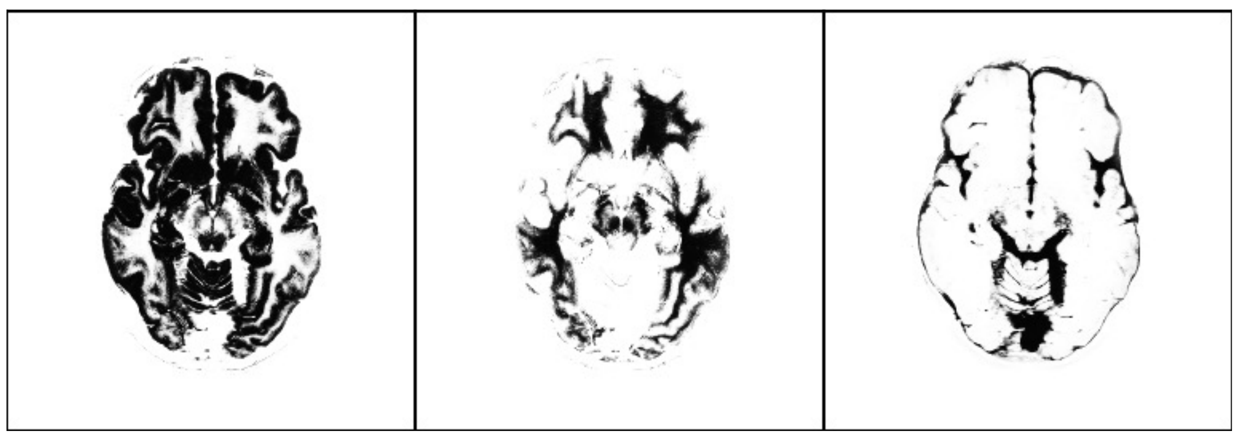
Applications of Bayesian inference in neuroimaging


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)

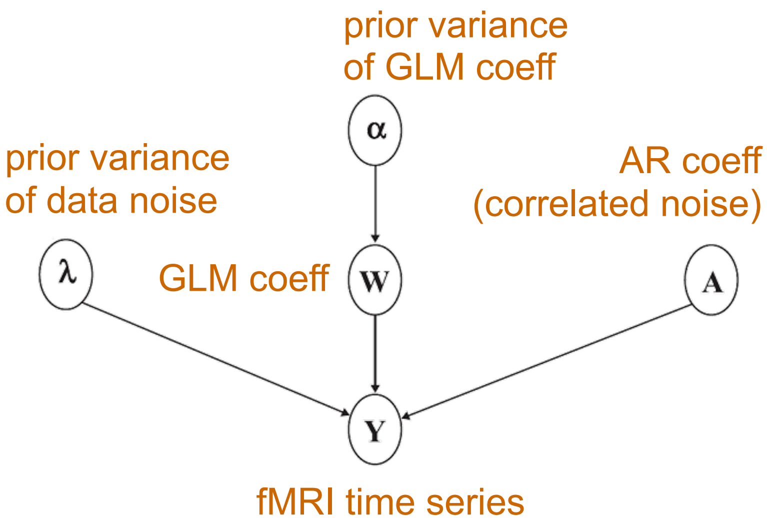




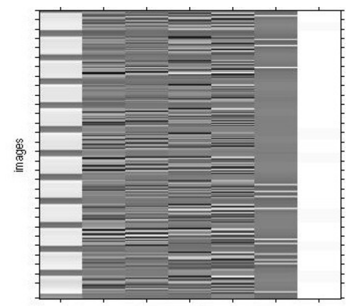


grey matter   white matter   CSF

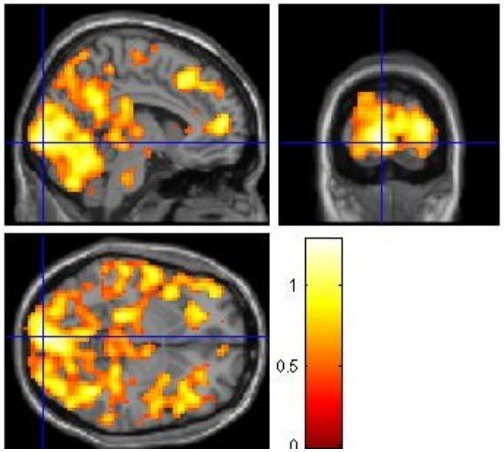
fMRI time series analysis



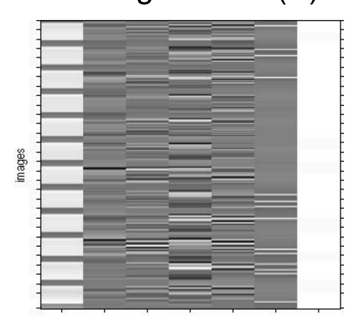
short-term memory
design matrix (X)



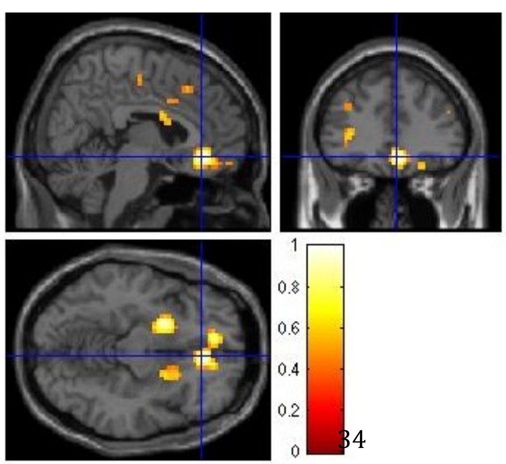
PPM: regions best explained
by short-term memory model



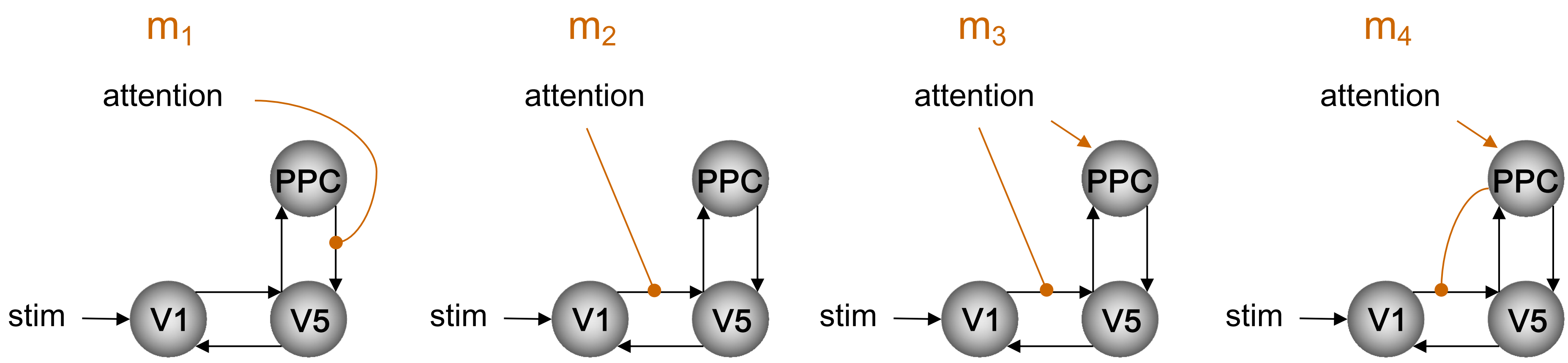
long-term memory
design matrix (X)

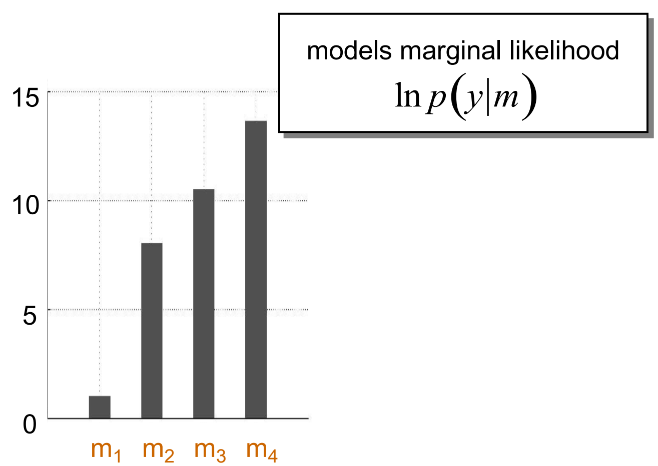


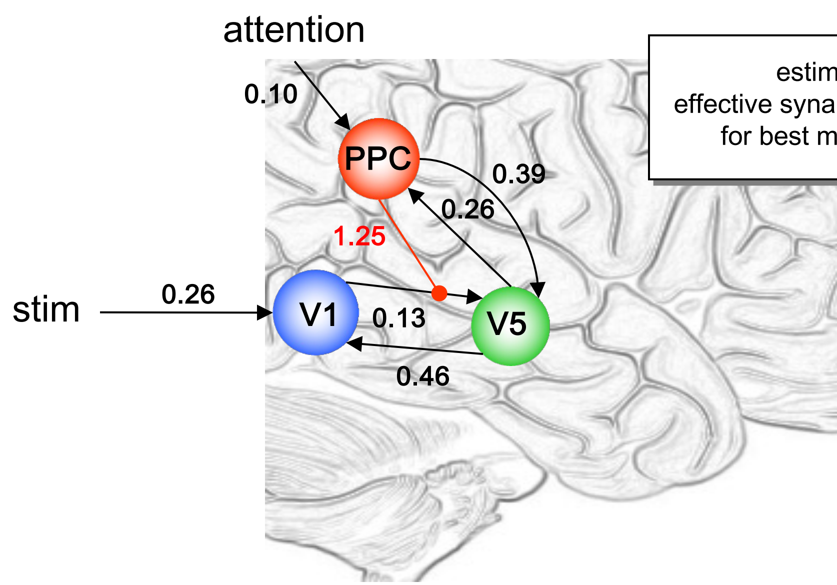
PPM: regions best explained
by long-term memory model



Dynamic causal modeling (DCM)



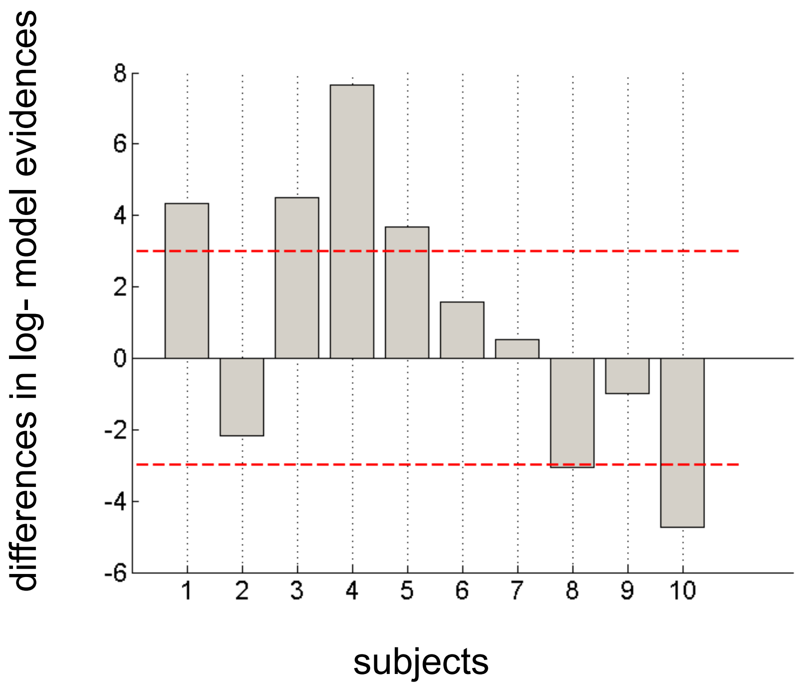




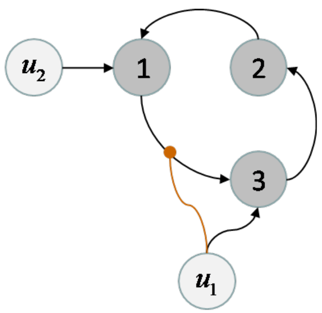
estimated
effective synaptic strengths
for best model (m_4)

Model comparison for group studies

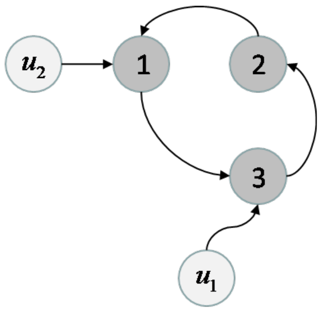
$$\ln p\left(y|m_1\right) - \ln p\left(y|m_2\right)$$



m_1



m_2



Fixed effect — Assume all subjects correspond to the same model

Random effect — Assume different subjects might correspond to different models

Thanks