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

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.

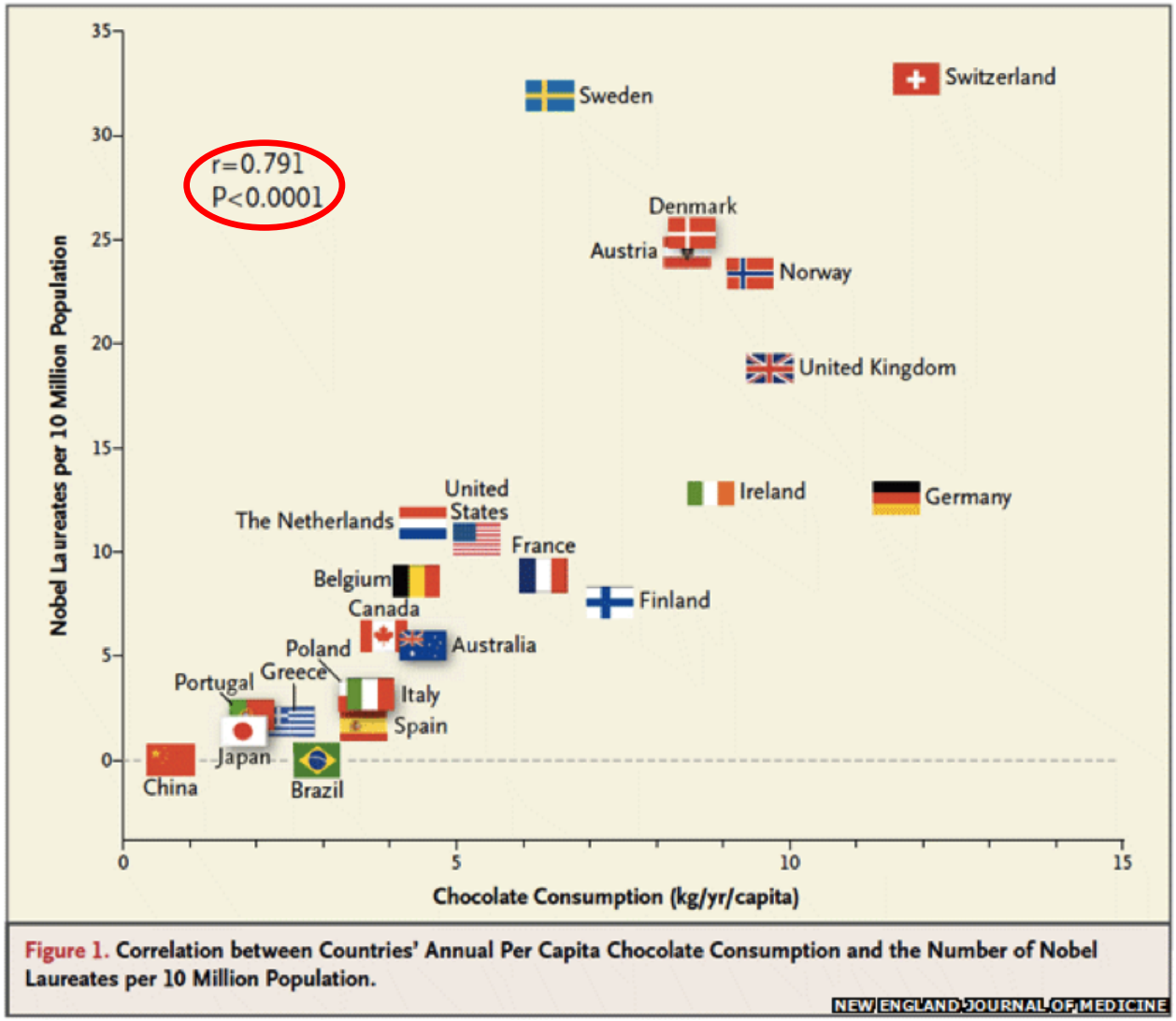


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

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*.**

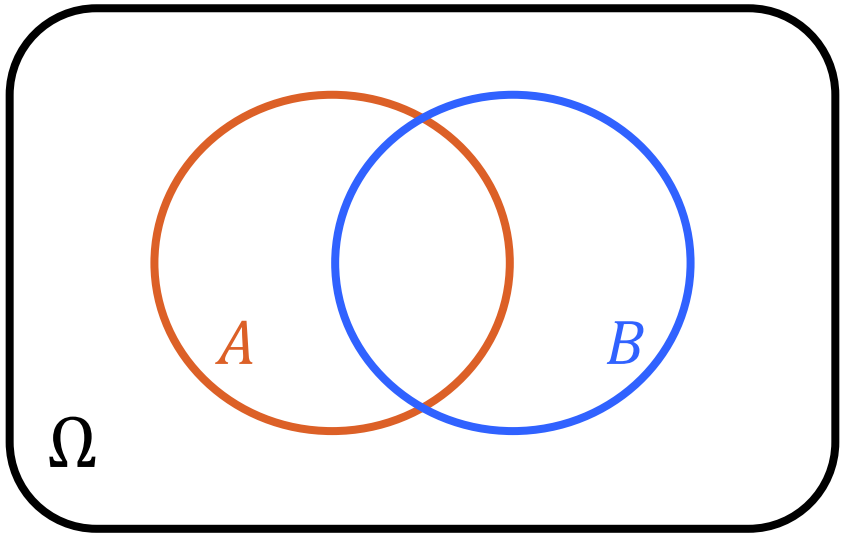
Our question here can be rephrased in terms of a conditional probability:

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

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

Calculating with probabilities: the setup

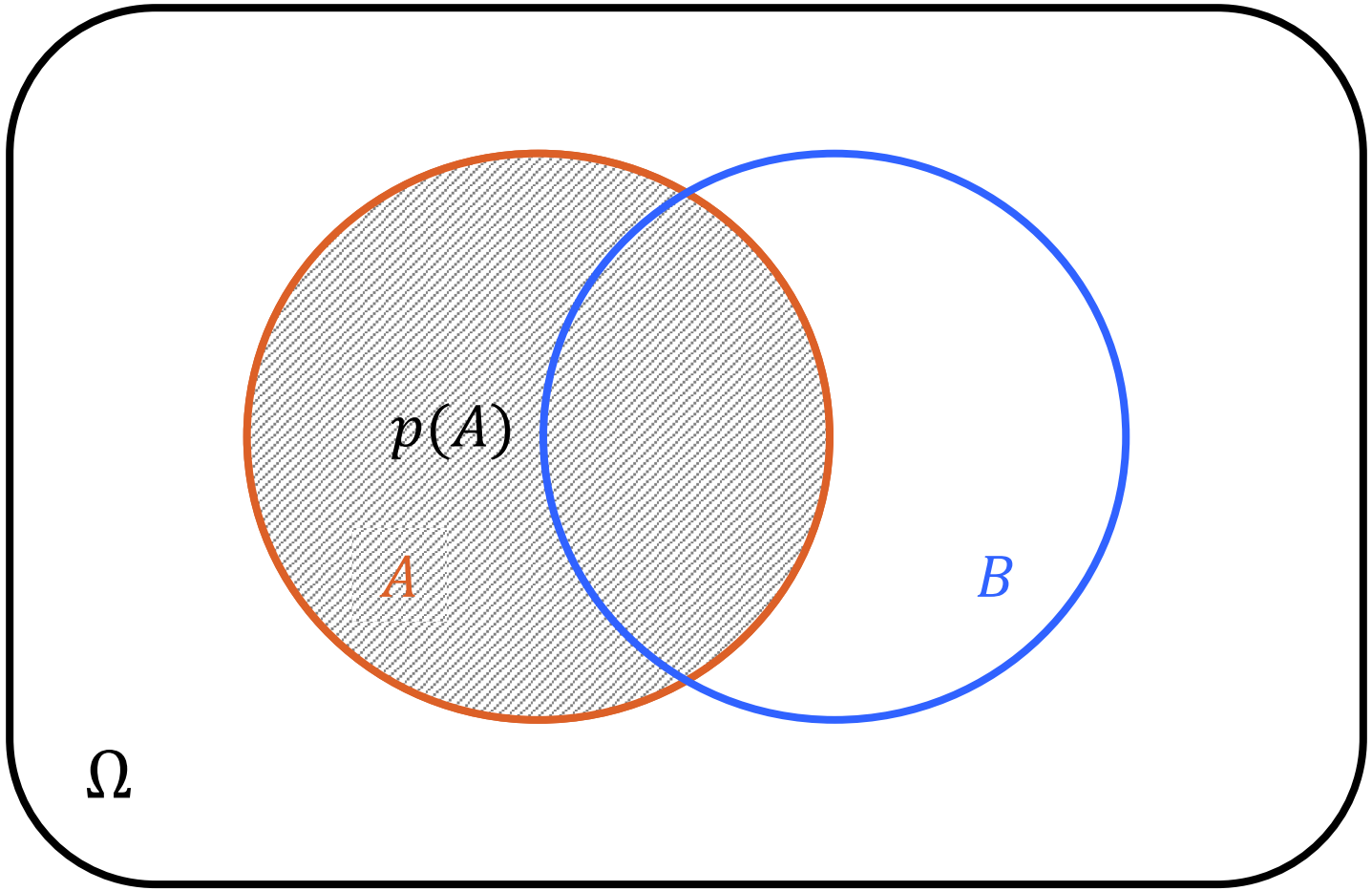
We assume a probability space Ω with subsets A and B



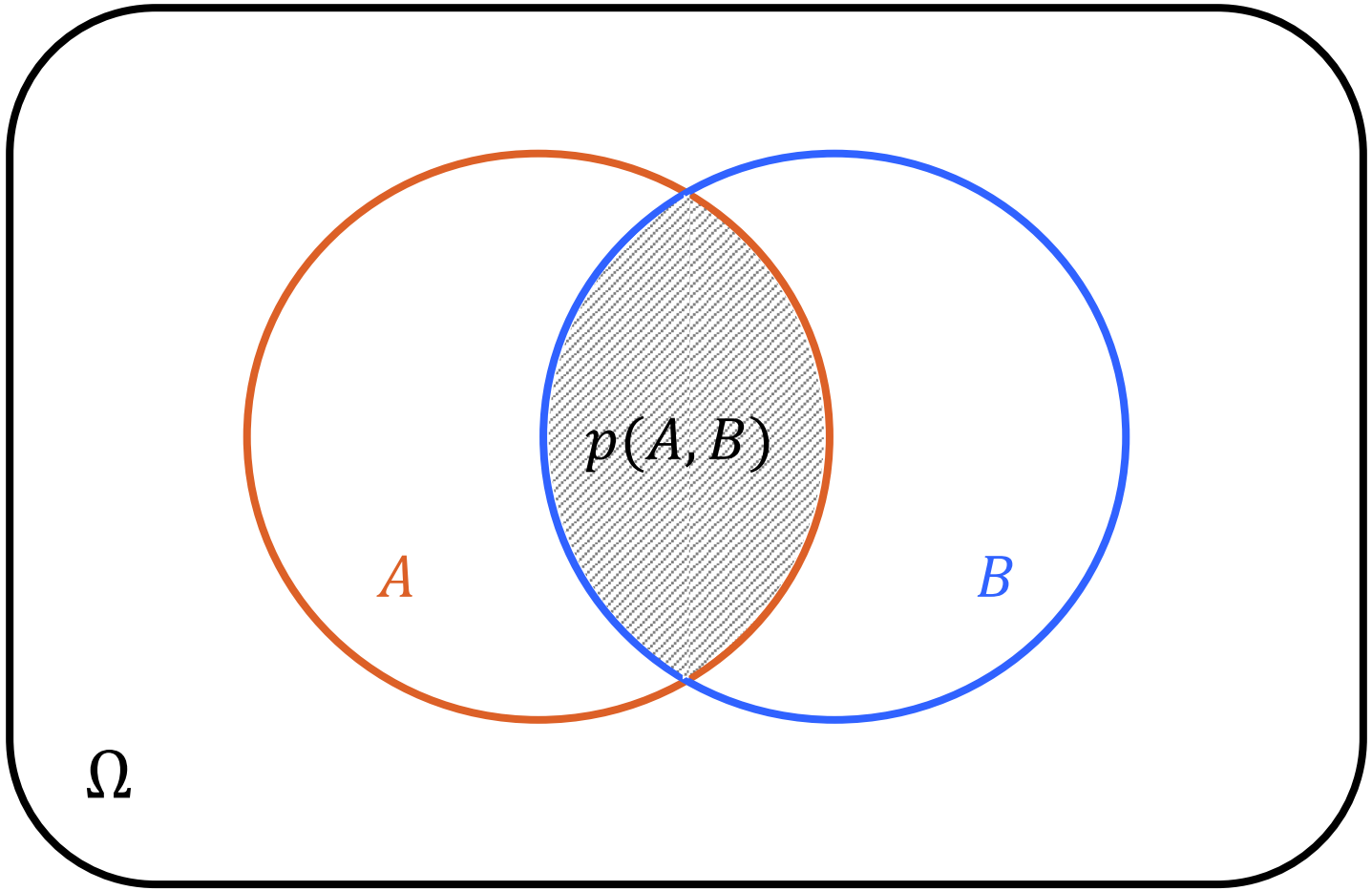
In order to understand *the rules of probability*, we need to understand **three kinds of probabilities**

- *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.

The rules of probability

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

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 rules of probability.

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

— Pierre-Simon Laplace, 1819

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?

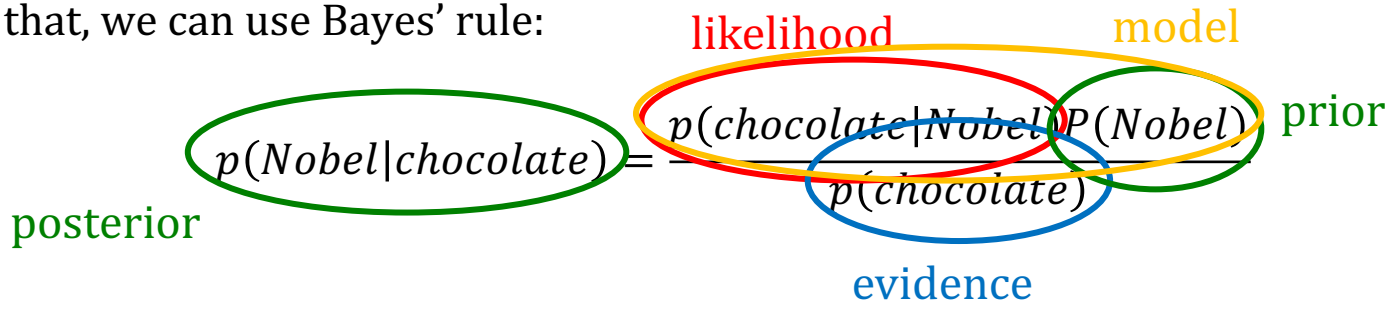
- Why is Bayes' rule important?
- It allows us to invert conditional probabilities, ie to pass from $p(B|A)$ to $p(A|B)$:

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

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

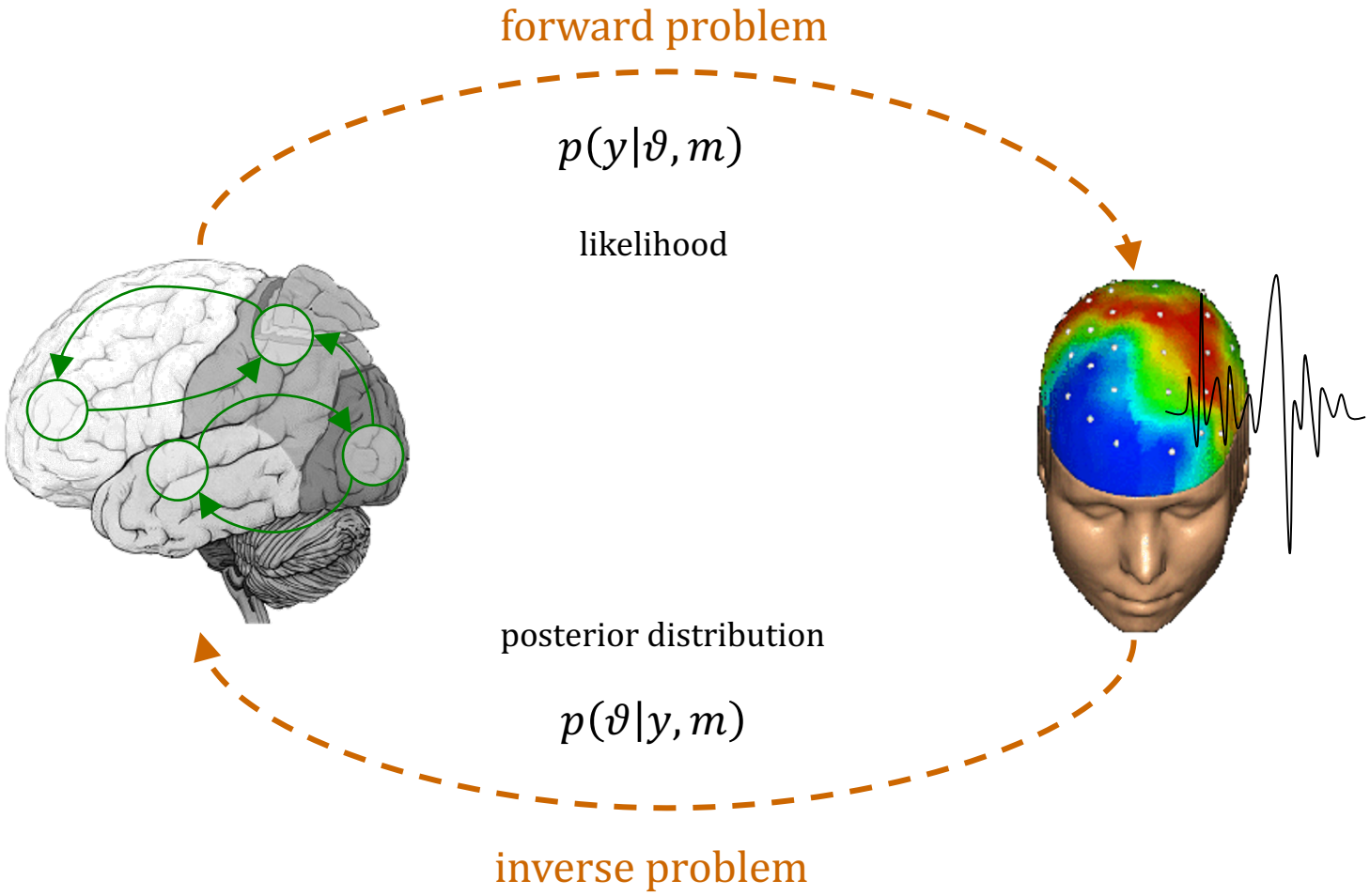
Bayes' rule: the chocolate example

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:

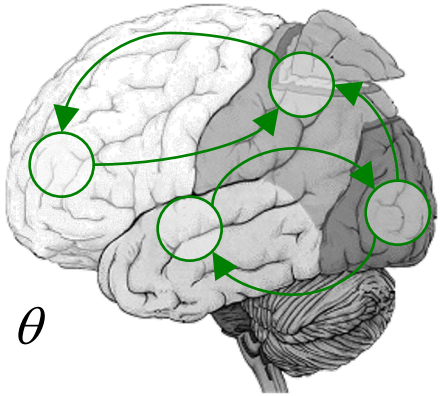


Inference on the quantities of interest in neuroimaging studies has exactly the same general structure.

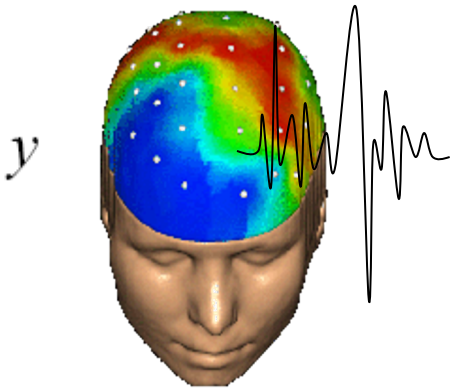
Inference in SPM



Inference in SPM



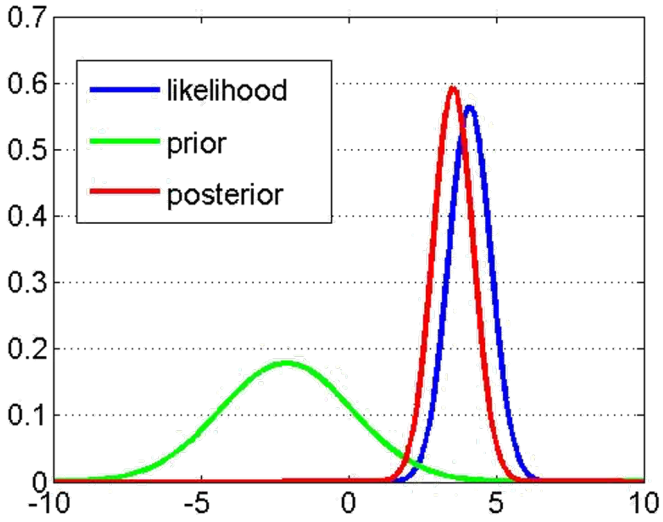
generative model m



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?

- By comparing their arithmetic means

Why do we take means?

- If we take the mean as our estimate, the error in our estimate is the mean of the errors in the individual measurements
- Taking the mean as maximum-likelihood estimate implies a **Gaussian error distribution**
- A Gaussian error distribution appropriately reflects our **prior** knowledge about the errors whenever we know nothing about them except perhaps their variance

A simple example of Bayesian inference

What next?

- 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

The procedure in brief:

- Determine your question of interest («What is the probability that...?»)
- Specify your model (likelihood and prior)
- Calculate the posterior using Bayes' theorem
- Ask your question of interest of the posterior

All you need is the rules of probability theory.

(Sometimes you'll encounter a nasty integral. But that's only a technical difficulty, not a conceptual one, and software packages like SPM will solve it for you – normally).

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:

Likelihood (Gaussian):

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

This is the probability of making observations $\{y_i\}_{i=1,\dots,n}$ if the **mean** of the sampling distribution is μ and its **precision** is λ.

Prior (Gaussian-gamma):

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

This is our assumption about the realistic range in which we expect to find μ and λ, determined by the **hyperparameters** μ_0, κ_0, a_0, and b_0.

A simple example of Bayesian inference

- Applying Bayes' rule gives us the **posterior hyperparameters** μ_n, κ_n, a_n and b_n

- If we choose prior hyperparameters $\kappa_0 = 0$, $a_0 = 0$, $b_0 = 0$, the posterior hyperparameters are:

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

- This means that all we need is n, the number of data points; $\bar{y}$, their mean; and s^2, their variance.

- If we choose different prior hyperparameters, the equations for the posterior hyperparameters look a bit more complicated, but in any case they can easily be calculated for our example model.

- In many applications of Bayesian inference, the posterior cannot be calculated analytically and written in terms of a function determined by hyperparameters. In these cases, **approximate Bayesian inference** has to be used, using for example *Monte Carlo sampling* or *variational calculus*.

A simple example of Bayesian inference

The joint posterior distributions of lifetimes μ_A of products from manufacturer A and μ_B are $p(\mu_A|\{y_i\}_A)$ and $p(\mu_B|\{y_k\}_B)$, respectively.

We can now use them to answer our question: what is the probability that parts from B live at least 3 hours longer than parts from A?

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

Note that the t-test told us that there was «no significant difference» even though according to our Bayesian calculation 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:

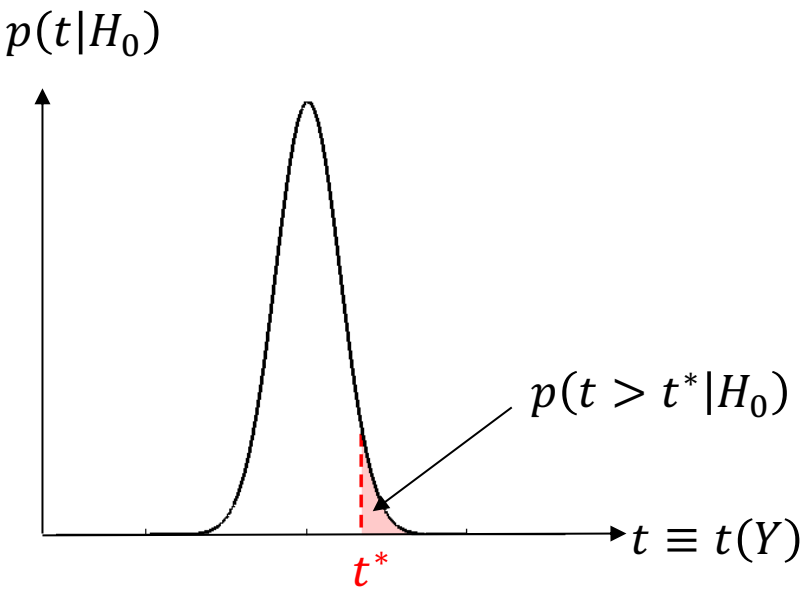
- Determine your question of interest («What is the probability that...?»)
- Specify your model (likelihood and prior)
- Ask your question of interest of the posterior

All you need is the rules of probability theory.

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

Classical

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

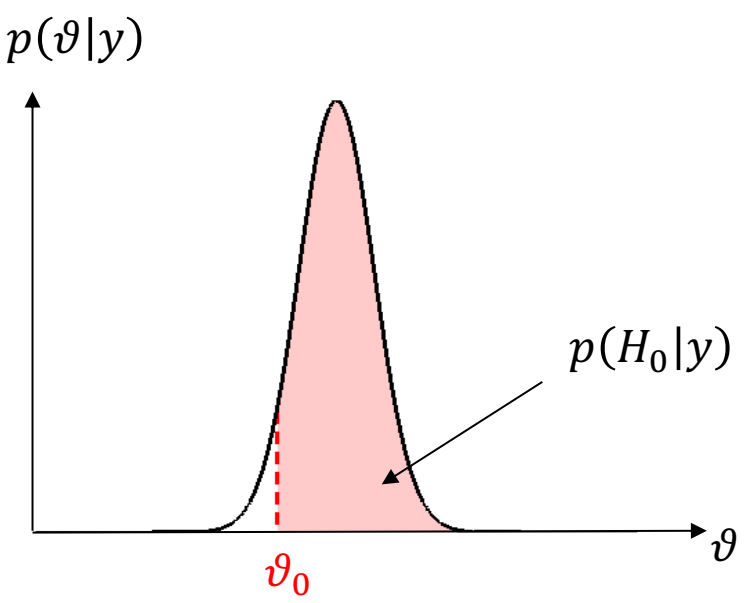


- estimate parameters (obtain test stat. t^*)
- apply decision rule, i.e.:

`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

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*

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

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

$$= \log \int q(\vartheta)\frac{p(y,\vartheta|H)}{q(\vartheta)}\,\mathrm{d}\vartheta \qquad \text{multiply by } 1 = \frac{q(\vartheta)}{q(\vartheta)}$$

$$\geq \int q(\vartheta)\log\frac{p(y,\vartheta|H)}{q(\vartheta)}\,\mathrm{d}\vartheta \qquad \text{Jensen's inequality}$$

a lower bound on the log-model evidence

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

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

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

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

Kullback-Leibler divergence

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 give a higher model evidence (which is a limitation of the concept of model evidence).

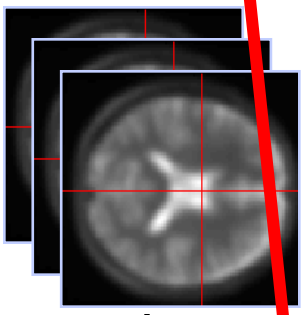
Applications of Bayesian inference

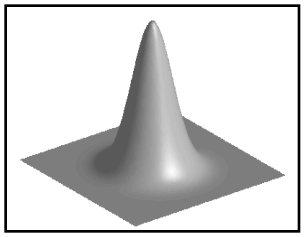
segmentation and normalisation

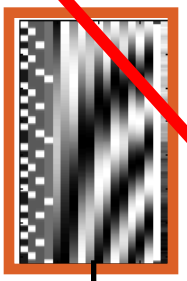
posterior probability maps (PPMs)

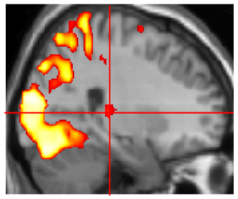
dynamic causal modelling

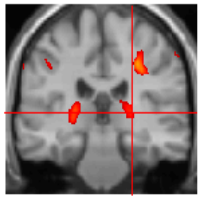
multivariate decoding

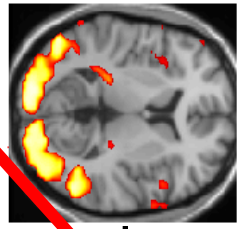












realignment

smoothing

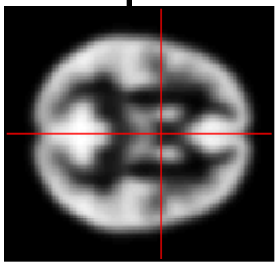
general linear model

normalisation

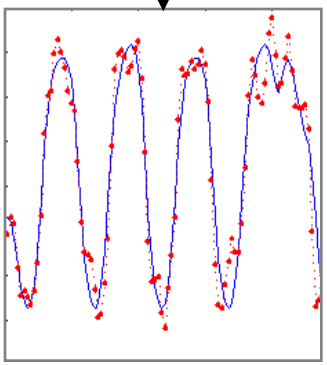
statistical inference

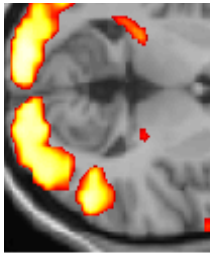
Gaussian field theory

p <0.05

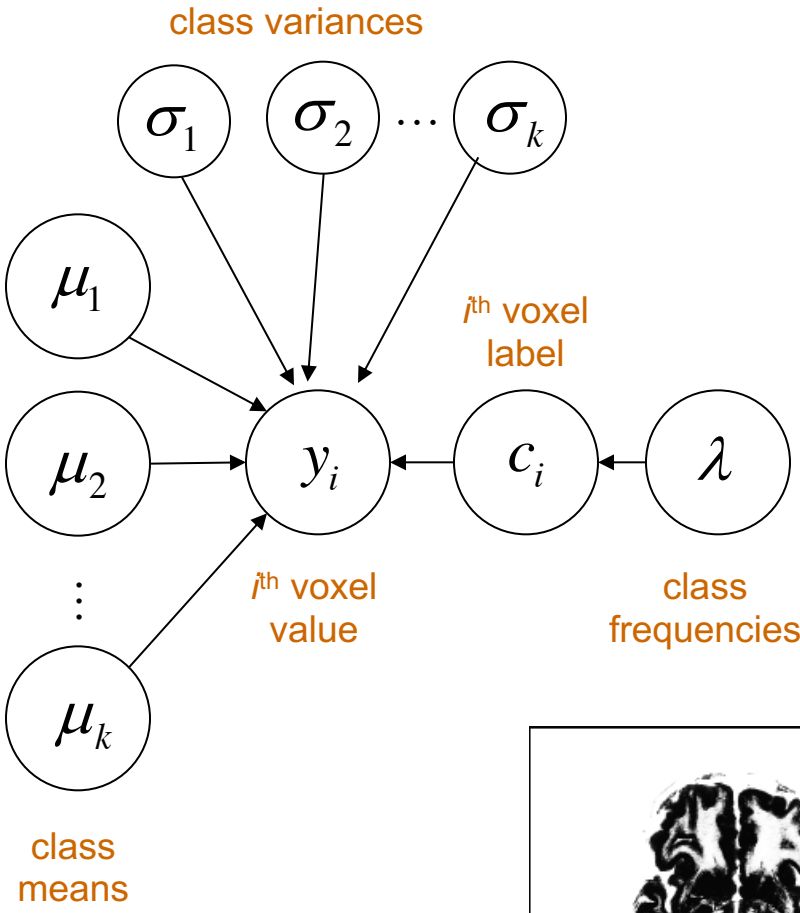


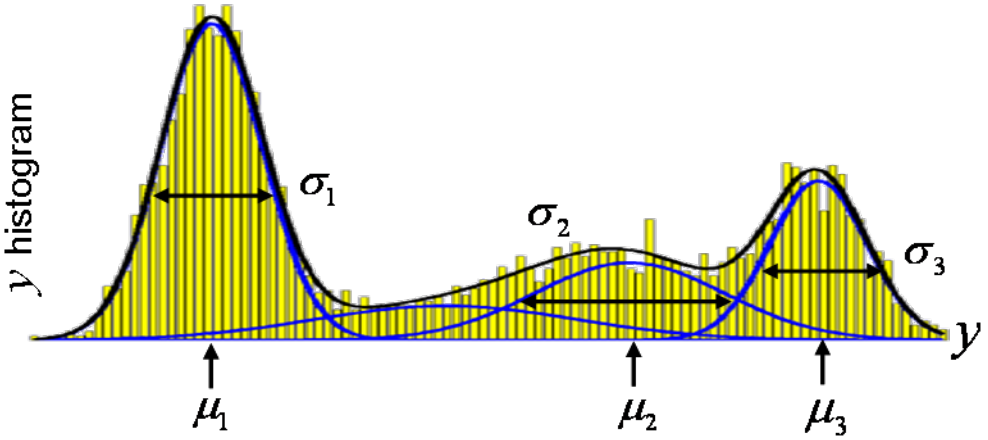
template

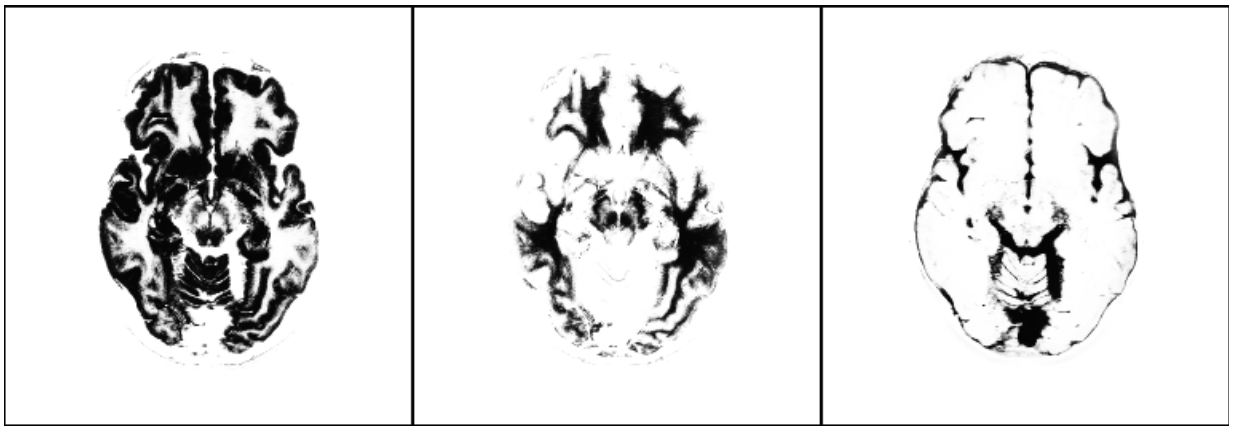




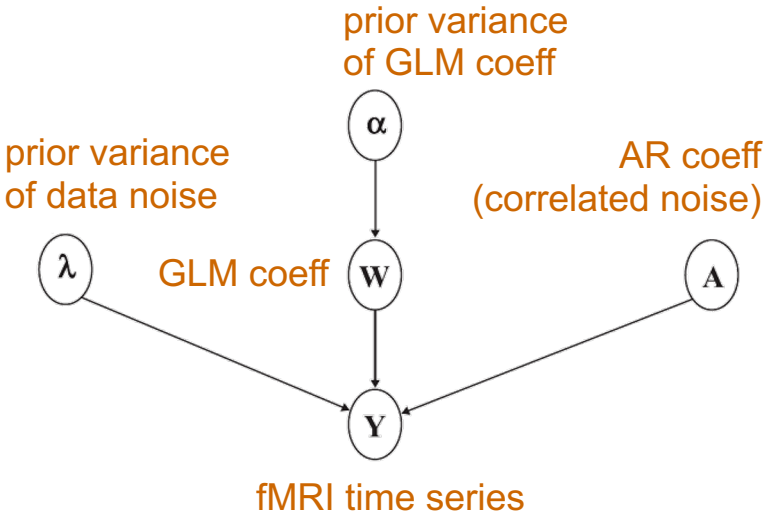
Segmentation (mixture of Gaussians-model)



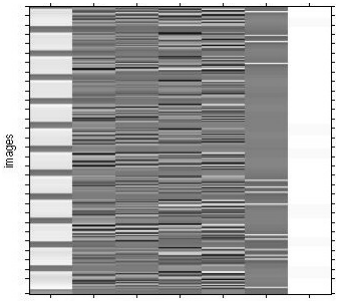




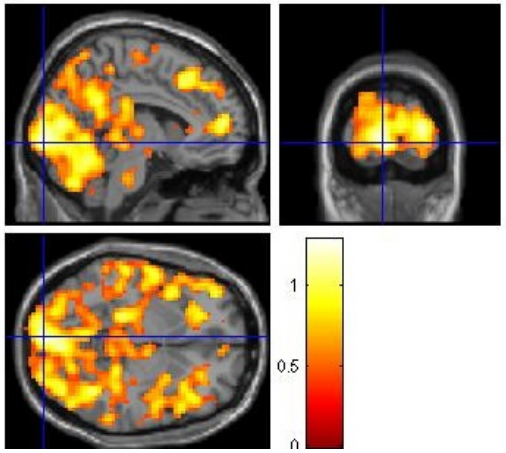
fMRI time series analysis



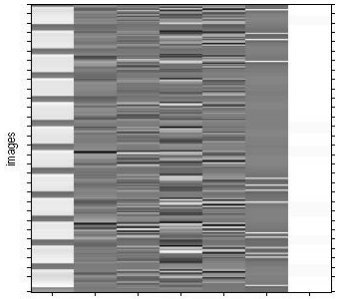
short-term memory
design matrix (X)



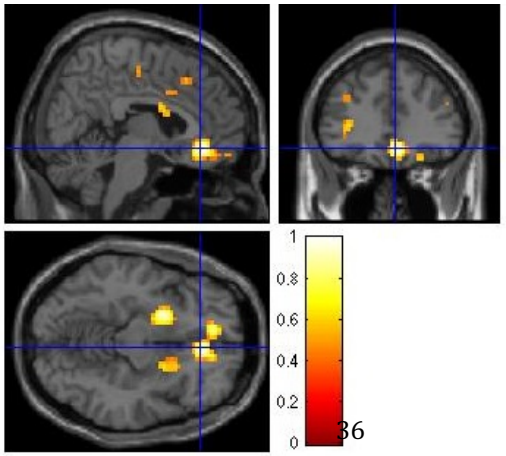
PPM: regions best explained
by short-term memory model



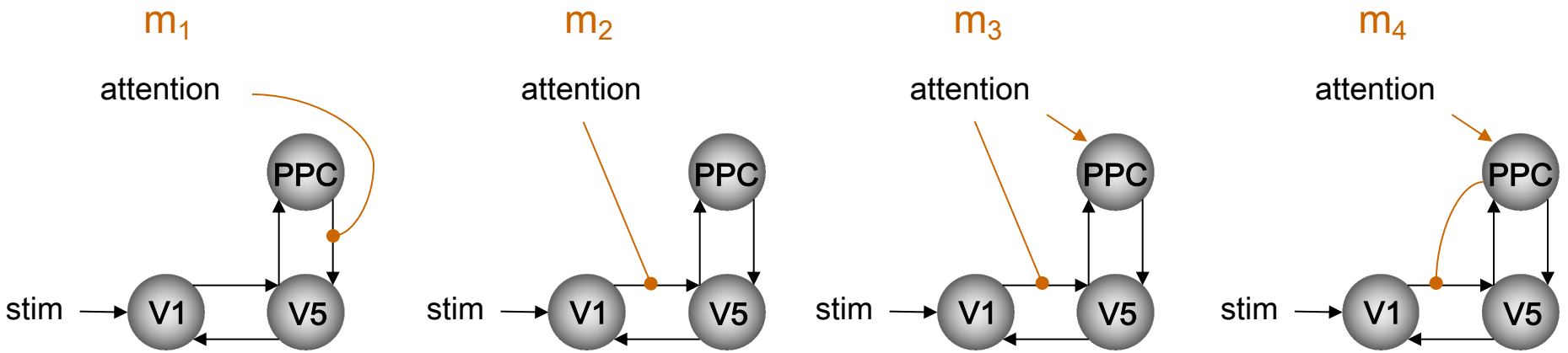
long-term memory
design matrix (X)

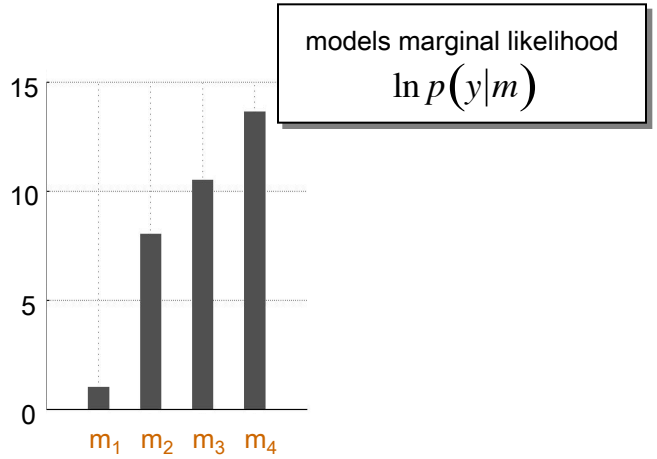


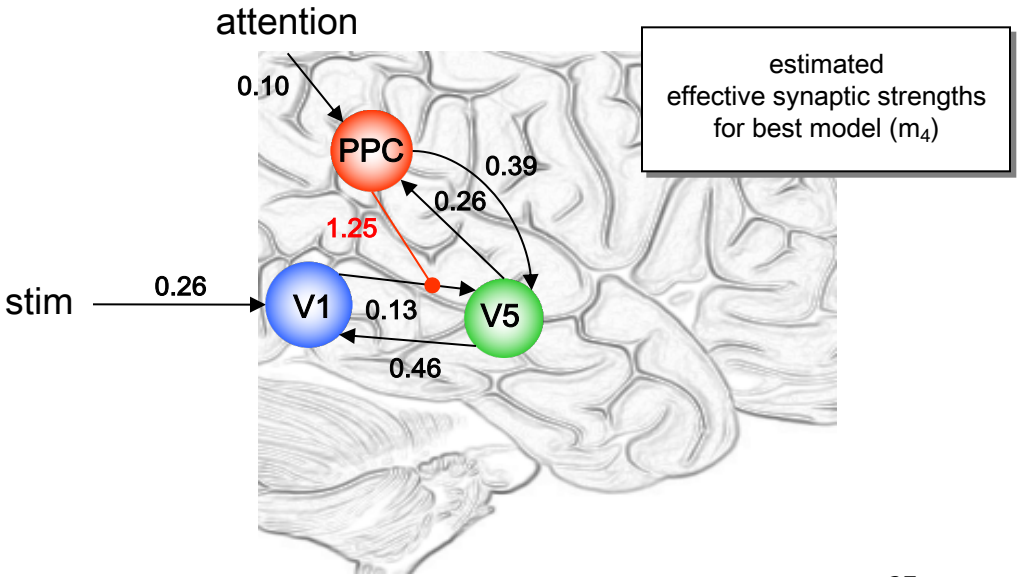
PPM: regions best explained
by long-term memory model



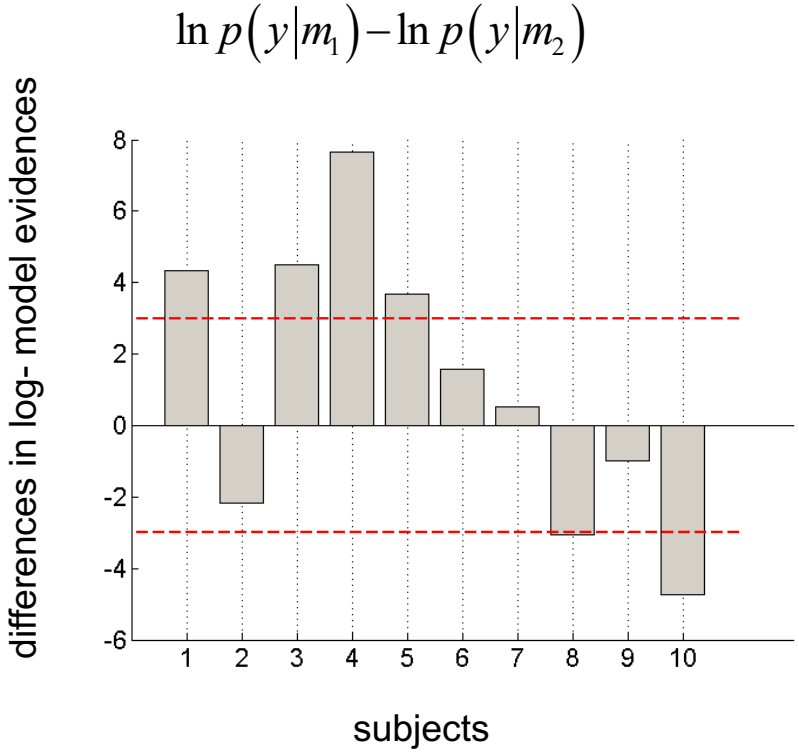
Dynamic causal modeling (DCM)

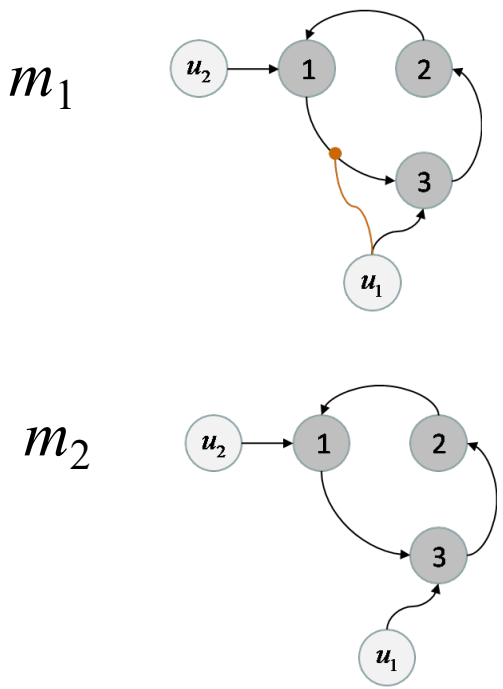






Model comparison for group studies





Fixed effect — Assume all subjects correspond to the same model

Random effect — Assume different subjects might correspond to different models

Thanks