Advanced Statistical Methods

Lecture 5

Content:

- Bayesian inference: Model comparison, Occam's razor, Priors and the Maximum Entropy Principle
- Monte Carlo Methods: MCMC, Detailed Balance, Metropolis-Hastings

Model comparison & Jeffrey's scale

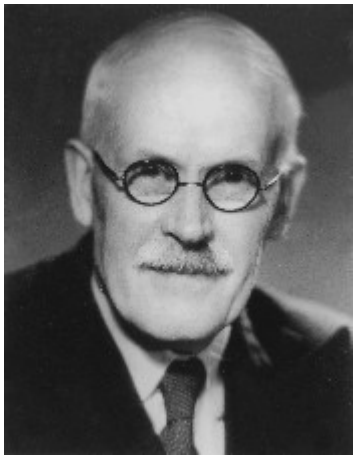
In Bayesian statistical inference, we perform ***Model comparison*** (this replaces *hypothesis testing* in the Frequentist approach). The central quantity is here the **odds ratio**

$$O_{ij} = \frac{P(M_i|\vec{x}, I)}{P(M_j|\vec{x}, I)}$$

which just depends on the model posteriors of a model *i* and *j*. Using Bayes' theorem, this can be written as

$$O_{ij} = \frac{P(M_i|I)}{P(M_j|I)} \underbrace{\frac{P(\vec{x}|M_i, I)}{P(\vec{x}|M_j, I)}}_{= B_{ij}}$$

Priors *Likelihoods*



H. Jeffreys
1891-1989

Bayes factor

- The Bayes factor describes how much *additional* credibility a model gains due to the available data.
- It is independent of priors on the model, but in general depends on priors on the model parameters.
- A wildly used interpretation of the Bayes factor is given by the "Jeffrey's scale"

Jeffrey's scale

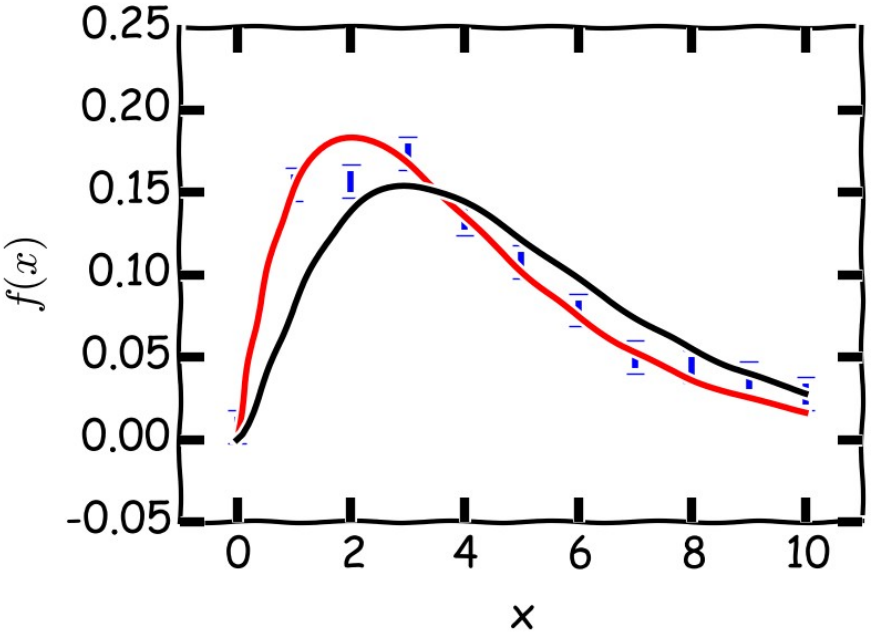
B	"Strength of evidence"
< 1	Negative
1 – 3	Barely worth mentioning
3 – 10	Substantial
10 – 30	Strong
30 – 100	Very strong
> 100	Decisive

Model comparison vs Hypothesis testing

Example:
We consider the two (simple) hypotheses that a given data set is described by a chi-squared distribution with either four or five degrees of freedom:

$$M_0 : \ \chi^2_{k=4}$$
$$M_1 : \ \chi^2_{k=5}$$

We use here mock data in eleven bins, with a standard deviation of σ=0.015.



Hypothesis testing:
We consider some test statistic, here the chi-squared goodness-of-fit

$$\chi^2 = \sum_{i=1}^{n_{\rm bins}} \frac{(E_i - O_i)^2}{\Delta E_i^2}$$

for which we obtain

$$\chi^2(M_0) \approx 10 \quad \chi^2(M_1) \approx 40$$

For eleven degrees of freedom, we obtain the p-values

$M_0 : \ p \sim 0.5$ (accepted)
$M_1 : \ p \sim 3.6 \cdot 10^{-5}$ (rejected at 4.0σ)

Model comparison:
The Bayes factor is given by the likelihood ratio

$$B_{01} = \frac{P(\vec{x}|M_0, I)}{P(\vec{x}|M_1, I)}$$

which in the present case (we assume normal distributed errors) is approximately given by

$$\sim \frac{\exp(-\frac{1}{2}(\chi^2(M_0))}{\exp(-\frac{1}{2}(\chi^2(M_1))} \sim 10^6$$

According to Jeffrey's scale, this is “decisive”.

Sigmas in different fields of science

A grumpy colleague might say:

$$3.0\sigma_{\text{sociology}} > 3.0\sigma_{\text{HEP}} > 3.0\sigma_{\text{astronomy}}$$

Comments:

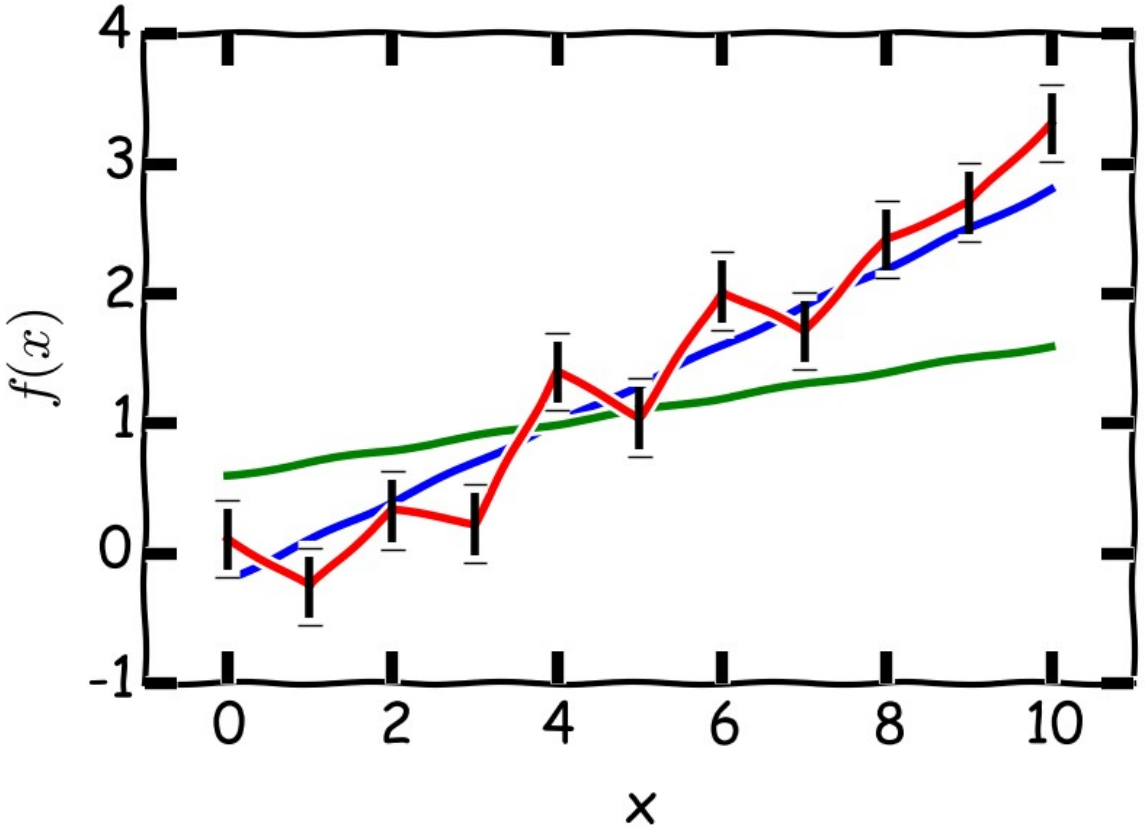
- In *high energy physics*, the standard threshold to talk about a discovery is 5.0σ. However, the high energy physics folks start getting excited at excesses around 3.0σ (which in most cases turn out to be wrong).
- In *astroparticle physics*, backgrounds are usually much more uncertain than in HEP, which is compensated for by higher statistical thresholds.
- In *sociology and psychology*, 2.0σ is enough to talk about evidence. Since their hypotheses are quite frequently true (compared to "This is evidence for new physics"), the rather high rate of type I errors does not dominate the field (hopefully!).
- If we want to prove that *astrology* is true, we probably have to require >20σ or more, given the practically infinite number of ways "miracles" can occur.

Having said that, the "Jeffery's scale" is apparently not adjusted for use in HEP. If you have to use it, *please use it with care.*

B	"Strength of evidence"	Corresponding one-sided p-value:
< 1	Negative	
1 – 3	Barely worth mentioning	
3 – 10	Substantial	<0.5σ
10 – 30	Strong	>1.3σ
30 – 100	Very strong	>1.8σ
> 100	Decisive	>2.3σ

Going too far

Three theorists propose three models (model "red", "green" and "blue") for some piece of observational data. Their best fits look like this:



Who got it wrong the most?

Occam's razor

"Occam's razor" denotes a principle that was formulated by many, and that became such a common-sense requirement for good science that is is rarely explicitly mentioned.

"Entities must not be multiplied beyond necessity."
(attributed to **William of Occam**)

"We are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances. Therefore, to the same natural effects we must, so far as possible, assign the same causes." (**Isaac Newton**)

"Whenever possible, substitute constructions out of known entities for inferences to unknown entities." (**Bertrand Russell**)

"Everything should be made a simple as possible, but not simpler." (**Albert Einstein**)

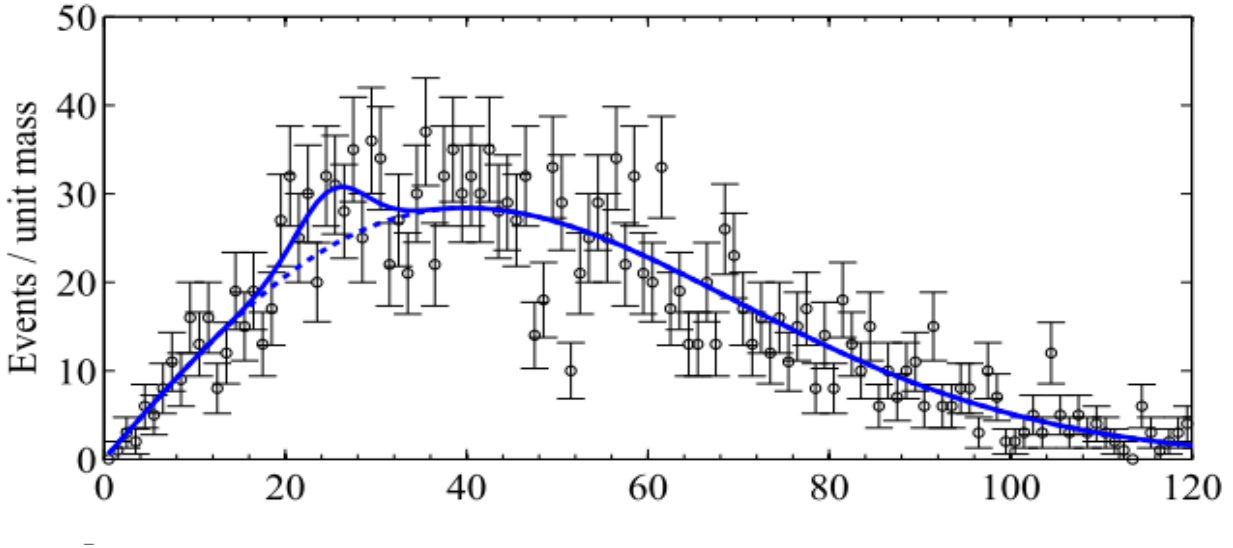
Occam
c. 1287-1347

Adding new entities (e.g. new particles) to your description of some phenomena is only justified if they are *required*. This requirement can be either observational (e.g. a detection at a particle collider) or theoretical (e.g. symmetries that reduce the number of free parameters). The overall goal is to:

Explain the data well while keeping the number of free parameters low.

Fortunately, this concept is automatically build into Bayesian inference!

Model comparison & Occam's razor



Scenario

- We consider two models: One (M_1) with a free parameter θ, and another (M_0) where this parameter is fixed to some value θ_0.
- The likelihood functions are connected via

$$P(\vec{x}|M_0, I) = P(\vec{x}|\theta_0, M_1, I)$$

- The model likelihood of M_1 is obtained after marginalization over the model parameter θ

$$P(\vec{x}|M_1, I) = \int d\theta P(\theta|M_1, I) P(\vec{x}|\theta, M_1, I)$$

Model comparison & Occam's razor

Marginalization, priors etc

- The two relevant factors in the marginalization

$$P(\vec{x}|M_1, I) = \int d\theta P(\theta|M_1, I) P(\vec{x}|\theta, M_1, I)$$

are the likelihood and the prior, both functions of θ.

- The prior is normalized to one, and hence depend on the *allowed parameter range*

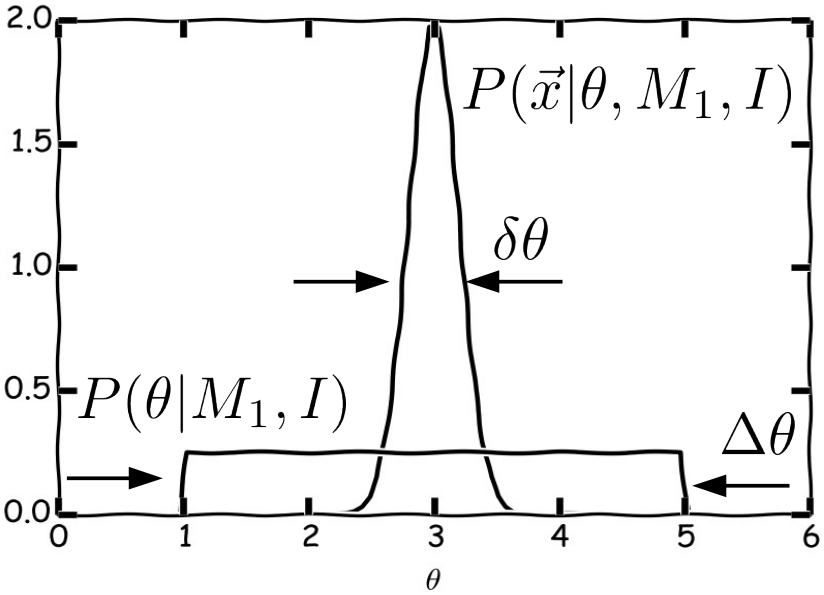
$$\int_{\Delta\theta} d\theta \underbrace{P(\theta|M_1, I)}_{\sim 1/\Delta\theta} = 1$$

- The likelihood function usually peaks around some value, which means that roughly

$$\int_{\Delta\theta} d\theta \; P(\vec{x}|\theta, M_1, I) \simeq P(\vec{x}|\hat{\theta}, M_1, I)\delta\theta$$

and hence the model likelihood for M_1 is

$$P(\vec{x}|M_1, I) \simeq P(\vec{x}|\hat{\theta}, M_1, I)\frac{\delta\theta}{\Delta\theta}$$



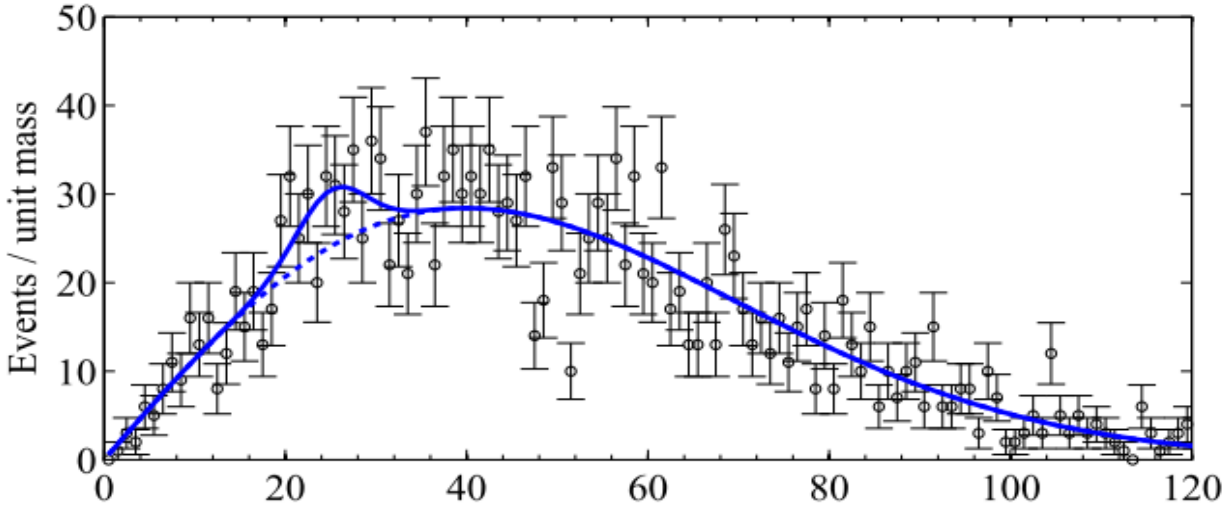
Bayes factor

- We find that

$$B_{10} \simeq \frac{P(\vec{x}|\hat{\theta}, M_1, I)}{P(\vec{x}|\theta_0, M_1, I)} \frac{\delta\theta}{\Delta\theta}$$

- The last factor acts as a *panelization* for any additional parameter that has a large prior uncertainty. This is *Occam's razor* in the Bayesian sense.

Example



Hypothesis testing

- We can test the existence of a significant line signal by considering the usual test statistic

$$TS = -2 \ln \frac{\mathcal{L}(0|\vec{x})}{\mathcal{L}(\hat{A}_s|\vec{x})}$$

- TS > 25 would correspond to a 5.0σ discovery of a signal.

Model comparison

- The marginal likelihood of the signal hypothesis (after integrating over signal strengths) is given by

$$P(\vec{x}|M_1, I) \simeq P(\vec{x}|\hat{A}_s, M_1, I) \frac{\delta A_s}{\Delta A_s}$$

- The Bayes factor is then

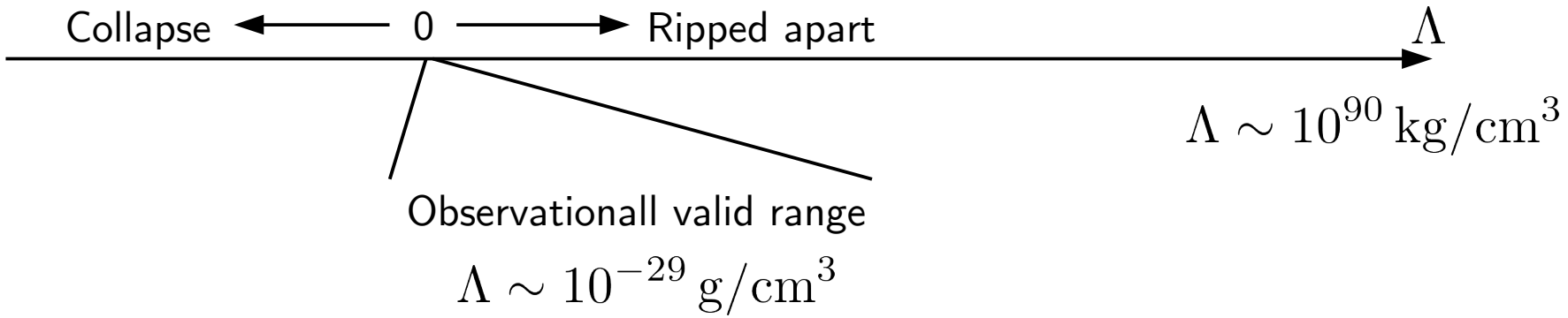
$$B_{10} = \frac{P(\vec{x}|M_1, I)}{P(\vec{x}|M_0, I)} \approx \frac{\mathcal{L}(\hat{A}_s|\vec{x})}{\mathcal{L}(0|\vec{x})} \frac{\delta A_s}{\Delta A_s}$$

which is up to the "Occam panelty" the same quantity relevant for hypothesis testing.

Connection with "fine-tuning"

In theoretical physics, one tries in general to avoid theories or scenarios where the correct description of an observed phenomenon requires a ridiculous amount of parameter adjustment. One famous example where this goes completely wrong is general relativity and the cosmological constant, which is 122 orders of magnitude away from its naively expected scale.

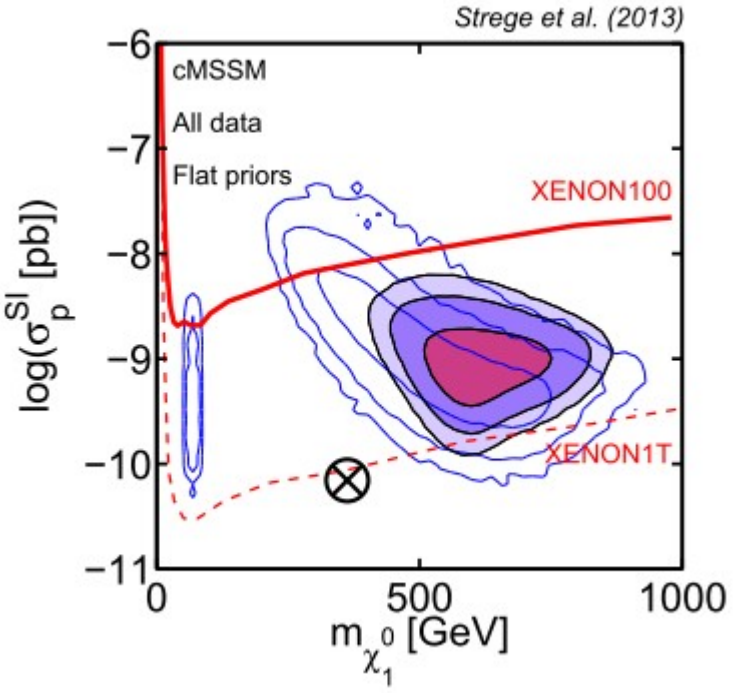
Occam's razor automatically penalizes for that, and prefer any theory that *predicts* this value from other principles.

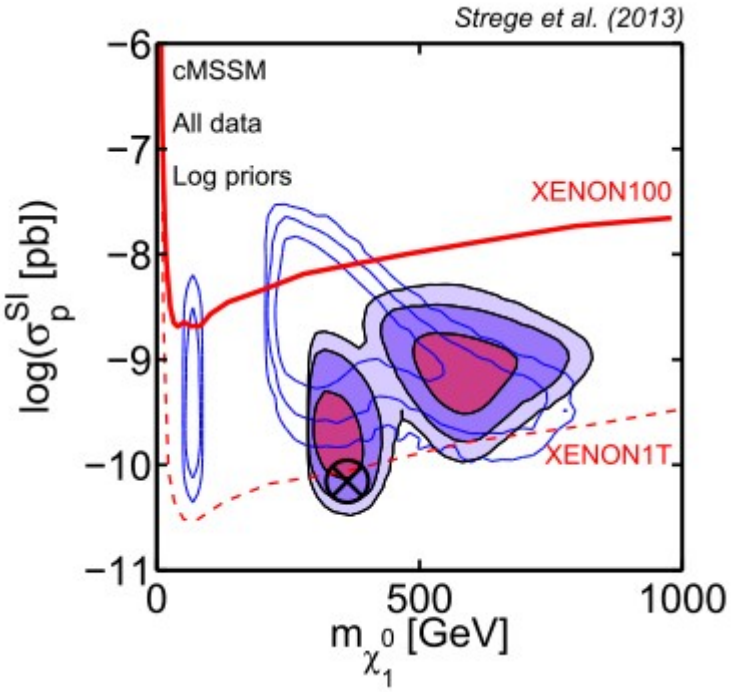


For a flat prior, we would obtain as panelization $\dfrac{\delta\Lambda}{\Delta\Lambda} \sim 10^{-122}$

The prior matters I

Combined limits on WIMP-nucleon and WIMP mass. Results depend on whether logarithmic or linear priors are adopted.





Whether or not the best-fit value (maximum likelihood) is actually included in the credible interval can depend on the prior! This is especially the case if the data is not strongly constraining the model.

The prior matters II

Upper limits on a Poisson variable

- We derive upper limits on the mean of a Poisson process with no background.
- The posterior is given by

$$P(\mu|c,I) = \frac{P(c|\mu,I)\cdot P(\mu|I)}{P(c|I)} = \frac{\mu^c e^{-\mu}\cdot P(\mu|I)}{\int_0^\mu \mu^c e^{-\mu}\cdot P(\mu|I)}$$

- In the case of a *flat prior*, this simply becomes

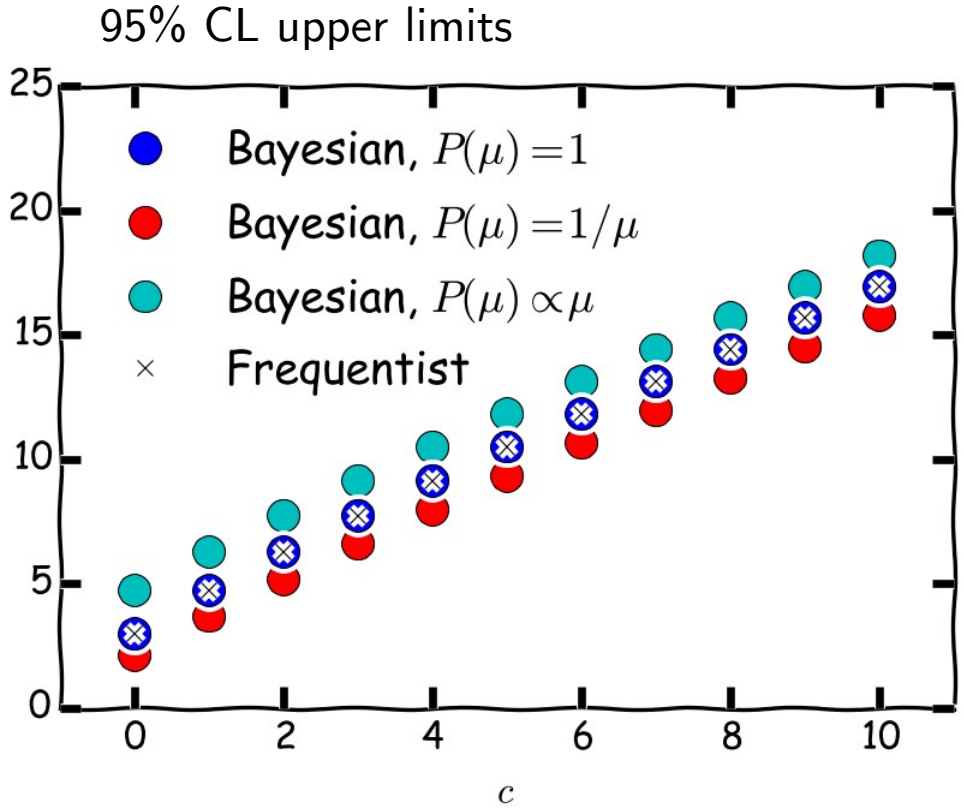
$$P(\mu|c,I) = \frac{\mu^c e^{-\mu}}{c!}$$

- The credible interval is obtained from a integration over the posterior

$$\int_{\mu_{\rm UL}}^{\infty} d\mu\, P(\mu|c,I) = 0.95$$

- The confidence interval is obtained from a sum over the likelihood function

$$\sum_{k=0}^{c} P(k|\mu_{\rm UL},I) \simeq 0.05$$



Common Priors

In the fortunate case of very abundant and precise data, the posterior distribution not strongly depend on the prior. However, often this is not the case, and the choice of the prior affects the result. A selection of most typical priors is listed below.

Uniform prior (for known scales)

- Appropriate if scale of parameter is roughly know.

$$P(\theta) = \frac{1}{\beta - \alpha} \qquad \theta \in [\alpha, \beta]$$

Jeffrey's prior (for scale free)

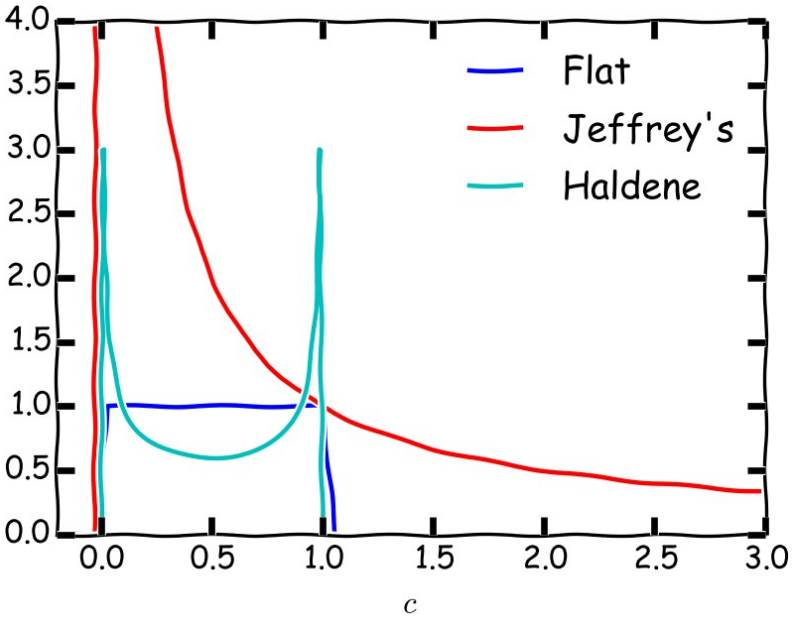
- Appropriate if scale of parameter is unknown.
- Uniform on log scale: the same probability in [1, 10] and [10, 100].

$$P(\theta) = \frac{\ln(\beta/\alpha)}{\theta} \qquad \theta \in [\alpha, \beta]$$

Haldane prior (for yes/no)

- Special prior for cases where we expect one of two models being true (corresponding to zero and one), but we also want to allow for mixed cases.

$$P(\theta) = \frac{1}{\pi} \frac{1}{\sqrt{\theta(1-\theta)}} \qquad \theta \in [0, 1]$$



The maximum entropy principle

Principle:
Out of all the possible probability distributions which agree with the given constraint information, select the one that is maximally non-committal with regard to missing information.

Question:
How do we accomplish the goal of being maximally non-committal about missing information?

Answer:
The greater the missing information, the more uncertain the estimate. Therefore, make estimates hat maximize the uncertainty in the probability distribution, while still being maximally constrained by the given information.

What is uncertainty and how do we measure it?

Example: Which one is the most uncertain?

$p_1 = p_2 = \frac{1}{2}$
$p_1 = \frac{1}{4},\ p_2 = \frac{3}{4}$
$p_1 = \frac{1}{100},\ p_2 = \frac{99}{100}$

$p_1 = p_2 = \frac{1}{2}$
$p_1 = p_2 = p_3 = p_4 = \frac{1}{4}$
$p_1 = p_2 = \cdots = p_8 = \frac{1}{8}$

[Gregory p185]

Shannon entropy

A rigorous way of defining priors, which takes ignorance as guiding principle, is the definition that is based on information (or Shannon) entropy.

C. Shannon
1916-2001

Shannon entropy:

- For a discrete probability mass function $p_i = P(x_i)$, the Shannon entropy is defined as

$$S(p_1, p_2, \ldots, p_n) = -\sum_{i=1}^{n} p_i \ln p_i$$

- It quantifies the amount of information that we would obtain when measuring the state of the system.

- Example: "No choice"

$$p_1 = 1, \;\; p_{i\geq 2} = 0 \quad \Rightarrow \quad S = 0$$

- Example: "Flat prior"

$$p_1, \ldots, p_n = \frac{1}{n} \quad \Rightarrow \quad S = \ln n$$

Connection to "information": With k bits, one can encode $n = 2^k$ different states. If the state is initially unknown (flat prior), the Shannon entropy is proportional to the number of bits.

$$S = k \ln 2$$

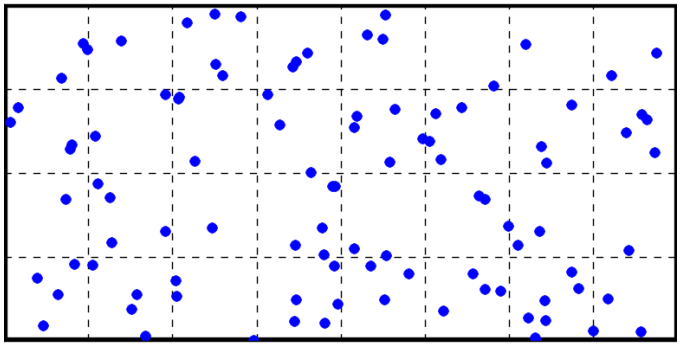
The "Monkey argument"

Ape

Imagine the following situation: A group of monkeys is throwing a very large number of balls in a grid of small boxes. The number of balls N is much larger than the number of boxes M, and we can infer a PDF via

$$p_i = \frac{n_i}{N}\,, \quad \text{where} \quad N = \sum_{i=1}^{N} n_i\,,$$

$$N \gg M\,, \text{ and it follows that } \sum_{i=1}^{n} p_i = 1\,.$$



The monkeys repeat this process a large number of times, and we discard all proposal PDFs that do not fulfill the additional constraint information (e.g. that the mean and/or the variance are fixed to a certain value).

We are now interested in the *frequencies* of different non-discarded PDFs. They will depend on

$$F(\{p_i\}) = \frac{\text{number of ways obtaining } \{n_j\}}{M^N} \quad \sim \frac{N!}{n_1! n_2! \ldots n_M!}$$

where the last step relies on combinatorial arguments that are used when deriving the Binomial distribution. Taking the log (and using Sterling's approximation), we find that

$$\ln(F) = \text{const} - N \sum_{i=1}^{M} p_i \ln p_i$$

PDFs from the MaxEnt principle

Idea

- Search probability mass function that *maximizes the Shannen entropy* under some external constraints.

$$\frac{\partial S}{\partial p_i} = 0 \qquad\qquad S(p_1, p_2, \ldots, p_n) = -\sum_{i=1}^{n} p_i \ln p_i$$

- Depending on the external constraints, this procedure leads to different well known probability mass functions.
- One can extend this approach also to continuous functions (with some additional input).

Examples:

- The minimum constraint: probabilities sum up to one

$$\sum_{i=1}^{n} p_i = 1 \quad \Rightarrow \quad p_1, p_2, \ldots, p_n = \frac{1}{n} \qquad \text{(Uniform prior)}$$

- Another constraint: mean value is fixed

$$\sum_{i=1}^{n} x_i p_i = \langle x \rangle \quad \Rightarrow \quad p_i = A e^{-\lambda x_i} \qquad \text{(Exponential distribution)}$$

- Another constraint: mean value and variance fixed

$$\sum_{i=1}^{n} (x_i - \mu)^2 p_i = \sigma^2 \qquad \Rightarrow \qquad \text{(Normal distribution)}$$

Monte Carlo

Main **computational challenges** for statistical inference in large systems:

- Multi-dimensional integrals over PDFs
  - analytically challenging (and in most cases impossible)
  - numerically expensive
  - evaluation of likelihood function often very time consuming
- Non-trivial boundary conditions
- Complicated estimators (estimators of estimators...)

Solution: We perform a large number of pseudo-experiments, drawing randomly from the PDFs, and study the statistical properties of the outcome.

Stanislaw Ulam
1909 – 1984

Example: estimating pi

- Draw large number of random number pairs (*x, y*) with

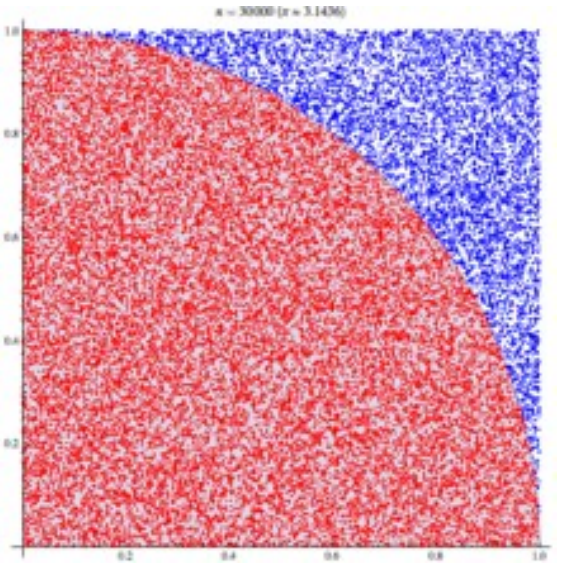
$$0 \leq x, y \leq 1$$

- Count points which fulfill (here red)

$$\sqrt{x^2 + y^2} \leq 1$$

- Pi is then approximated by the ratio

$$\pi \approx \frac{4\, n_{\mathrm{red}}}{n_{\mathrm{total}}} + \mathcal{O}\left(\frac{1}{\sqrt{n_{\mathrm{red}}}}\right)$$



Monte Carlo Integration

The prototypical problem in Bayesian analyses is to calculate expectation values from posterior distribution functions:

$$\langle f(\theta)\rangle = \int f(\theta)\, p(\theta|\vec{x}, I)\, d\theta = \int g(\theta)\, d\theta$$

This is here equivalent to evaluate the integral over function *g(θ)*.

This integral can be approximated in terms of the *mean* and *variance* of *g(θ)* over a finite volume *V* (where the integral contributes non-negligibly).

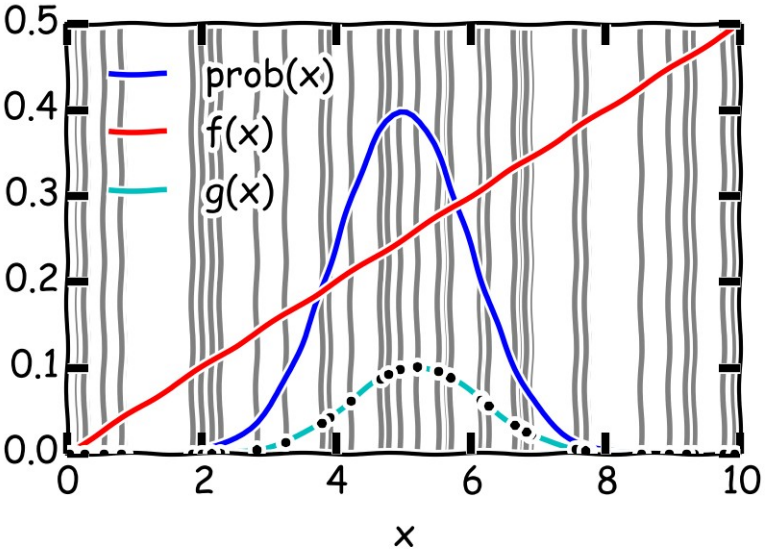
$$\langle f(\vec{\theta})\rangle = \int_V g(\vec{\theta}) d^n\theta \approx V \times \langle g(\vec{\theta})\rangle \pm V \times \sqrt{\frac{\langle g^2(\vec{\theta})\rangle - \langle g(\vec{\theta})\rangle^2}{n}}$$

Here, we used the definitions

$$\langle g(\vec{\theta})\rangle = \frac{1}{n}\sum_{i=1}^{n} g(\vec{\theta}_i) \qquad \langle g^2(\vec{\theta})\rangle = \frac{1}{n}\sum_{i=1}^{n} g^2(\vec{\theta}_i)$$

V: Volume in which *g(θ)* significantly contributes to the integral

$\vec{\theta}_i$: Large number of parameters randomly drawn from *V*



Note:

- The above error estimate is only an *estimate* and can be too small if most of the variance of *g* comes from a small region.

Importance sampling

In traditional MC integration, we evaluate f(theta) typically at a large number of points where the PDF is very small. We can instead base the estimate on a sample from the PDF.

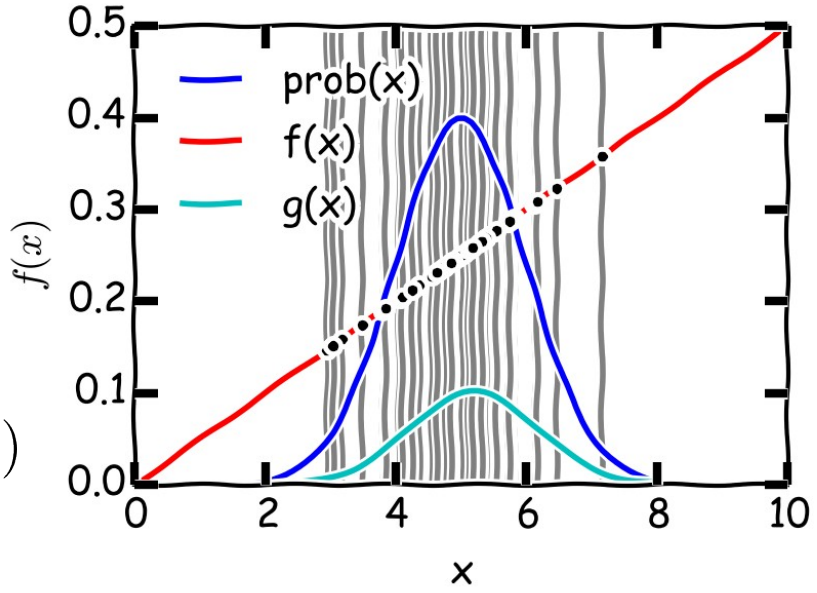
$$\langle f(\theta)\rangle = \int f(\theta)\, p(\theta|\vec{x}, I)\, d\theta$$

In order to sample a point from

$$p(\theta|\vec{x}, I)$$

we select a random number *r* in the range

$$[0, p_{\max}] \quad \text{with} \quad p_{\max} = \max_{\theta} p(\theta|\vec{x}, I)$$



If we find that

$$r \leq p(\theta_i|\vec{x}, I)$$

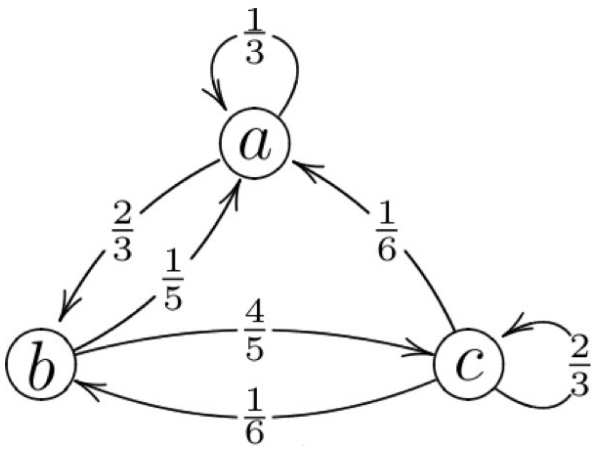
we accept the point and add it to the list

$$\{\theta_i\}_{i=1,\ldots,n}$$ (this is the usual "acceptance-rejection method")

Expectation values are then given by

$$\langle f(\theta)\rangle \approx \frac{1}{n}\sum_{i=1}^{n} f(\theta_i) \pm \sqrt{\frac{1}{n}}\sqrt{\frac{1}{n}\sum_i f(\theta_i)^2 - \left(\frac{1}{n}\sum_i f(\theta_i)\right)^2}$$

Markov Chain Monte Carlo



Example for (irreversible) Marov Chain, with states {*a*, *b*, *c*}.

Andrey Markov
1856 – 1922

Markov chain:

- A Markov chain describes a system that undergoes random transitions from one state to another in some state space.
- The random transitions have to fulfill the *Markov property:*

$$P(X_{n+1} = x | X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$
$$= P(X_{n+1} = x | X_n = x_n)$$

Each new state depends only on the previous one, not on the entire history (Markov chains are "memoryless").

The prototypical MCMC algorithm

Metropolis-Hastings Algorithm

1) Generate initial state x (this can be a vector, of course).
2) Randomly pick new state according to proposal distribution

$$g(x \to x')$$

3) Accept state as new state with acceptance probability

$$A(x \to x') = \begin{cases} 1 & \text{if } P(x') > P(x) \\ P(x')/P(x) & \text{if } P(x') \leq P(x) \end{cases}$$

4) If the step is accepted, set $x = x'$ and save the new state in a list.
5) Go back to step 2.

N. Metropolis
1915 – 1999

Notes:

- In many cases, the adopted proposal distributions are symmetric

$$g(x \to x') = g(x' \to x)$$

- The transition probability (or transition matrix) in the Markov chain sense is given by

$$\pi(x \to x') = g(x \to x')A(x \to x')$$

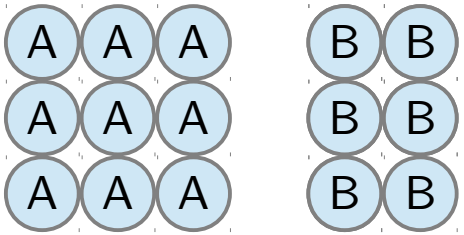
- Breaking of loop can be based on number of accepted points.
- If the chain converges, $\{x_i\}$ is a sample drawn from $P(x)$ (thanks to detailed balance).

Detailed balance

Concept of detailed balance: *"Each process is balanced by its reverse process."*

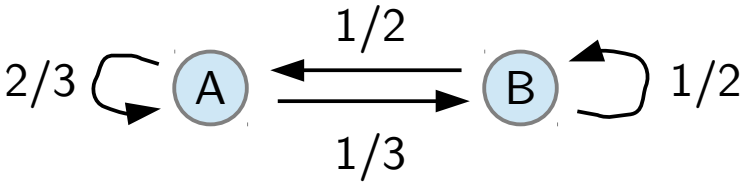
Example:

System (with 15 elements):



$$P(A) = \frac{3}{2}P(B)$$

Rates:



$$\pi(A \to B) = 1/3$$
$$\pi(B \to A) = 1/2$$

Same number of transitions in both direction:

$$\pi(A \to B)P(A) = \pi(B \to A)P(B)$$

A Markov Process satisfies detailed balances if it is ***reversible***, i.e. if the product of transition rates over any closed loop is the same in both directions (Kolmogorov's criterion).

for example:

$$\pi_{D\to B}\pi_{B\to C}\pi_{C\to D} = \pi_{D\to C}\pi_{C\to B}\pi_{B\to D}$$