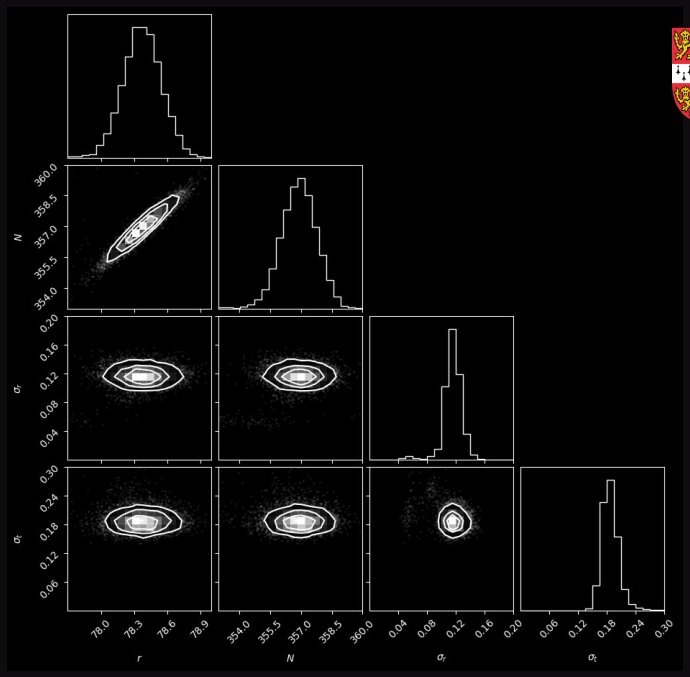


Department of Applied Mathematics
and Theoretical Physics (DAMTP)

Bayesian Statistics - Lecture 1, Tuesday

AIPHY school, Monopoli, Sep 30th - Oct 4th 2024

Christopher J. Moore

Outline

Bayes' Theorem

Bayesian versus frequentist views on probability

Stochastic Samples

Monte Carlo Methods

Bayesian Model Selection, the role of the evidence

Methods for calculating the evidence

Bayes' Theorem

$$P(A|B) = \frac{1}{P(B)} P(A) P(B|A)$$

Bayes' Theorem

$$P(\theta|d) = \frac{1}{\mathcal{Z}(d)}\pi(\theta)\mathcal{L}(d|\theta)$$

Posterior Evidence Prior Likelihood

Let the "events" be our model parameters and the observed data

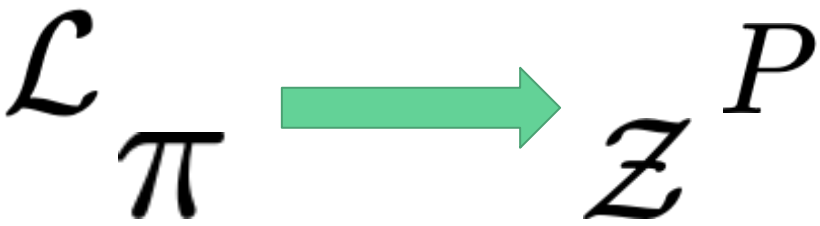
Bayes' Theorem

$$P(\theta|d) = \frac{1}{\mathcal{Z}(d)}\pi(\theta)\mathcal{L}(d|\theta)$$

Posterior Evidence Prior Likelihood

Let the "events" be our model parameters and the observed data

In a Bayesian inference, the likelihood and prior are the inputs and the posterior (and sometimes the evidence) are the outputs of the analysis



Bayesian versus frequentist views on probability

Historically, there has been a debate between Bayesian and frequentist views of probability

The debate does not concern the validity of Bayes' theorem itself, but rather the way it is applied to a model and the values of its parameters

In the frequentist view, probability is defined as a frequency of an event in many trials. There must be some random experiment that can (at least in principle) be repeated. The likelihood is a valid frequentist probability.

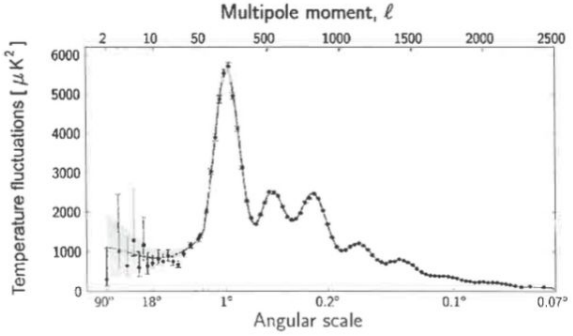
In the Bayesian view, probabilities represent our state of knowledge, or our degree of belief. The prior and posterior are valid probabilities only from the Bayesian viewpoint.

The outcome of flipping a coin; H or T ?

Will it rain tomorrow ? (Is it raining in London now ?)

What is the 10^{100}th digit of π ?

Will a particular ^{14}C atom decay in the next century ?



What's the probability of finding a particular value for a specific harmonic in this CMB data ?

Who should win this chess game ? (Given perfect play.)

Will I roll a double 6 on my next turn in Backgammon ?

The trained weights and biases for my neural network?

Bayesian versus frequentist views on probability

Historically, there has been a debate between Bayesian and frequentist views of probability

The debate does not concern the validity of Bayes' theorem itself, but rather the way it is applied to a model and the values of its parameters

In the frequentist view, probability is defined as a frequency of an event in many trials. There must be some random experiment that can (at least in principle) be repeated. The likelihood is a valid frequentist probability.

In the Bayesian view, probabilities represent our state of knowledge, or our degree of belief. The prior and posterior are valid probabilities only from the Bayesian viewpoint.

The Bayesian view allows us apply tools of probability to a wider range of problems. This has proved to be very fruitful.

The two roles of the likelihood

The likelihood $\mathcal{L}(d|\theta)$ is a PDF of the data *and* a function of the parameters

My notation is, unfortunately, not standard. Common to see it written as $L(\theta)$, emphasising its role as a function and suppressing the data dependence

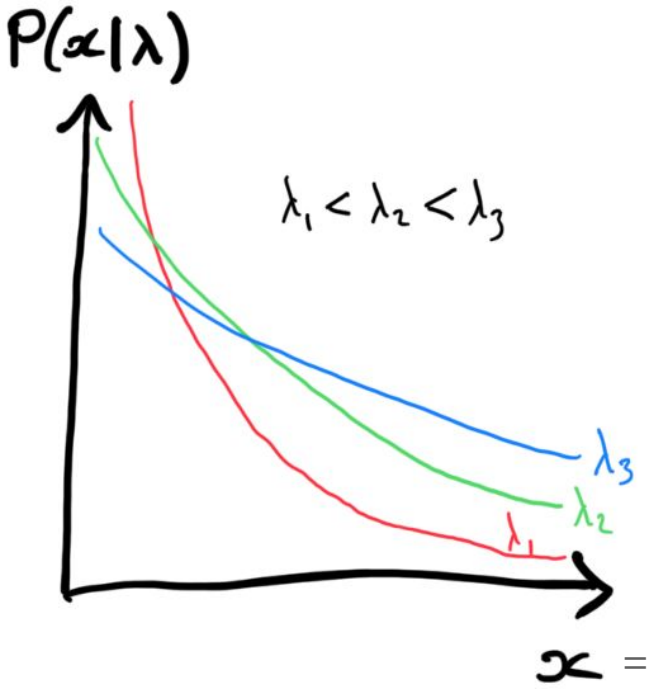
Exercise:

A source emits unstable particles that decay after travelling a distance x. A number N of decays are observed at locations $\{x_1, x_2, \ldots, x_N\}$.

For each particle, the distance travelled, $x>0$, is exponentially distributed

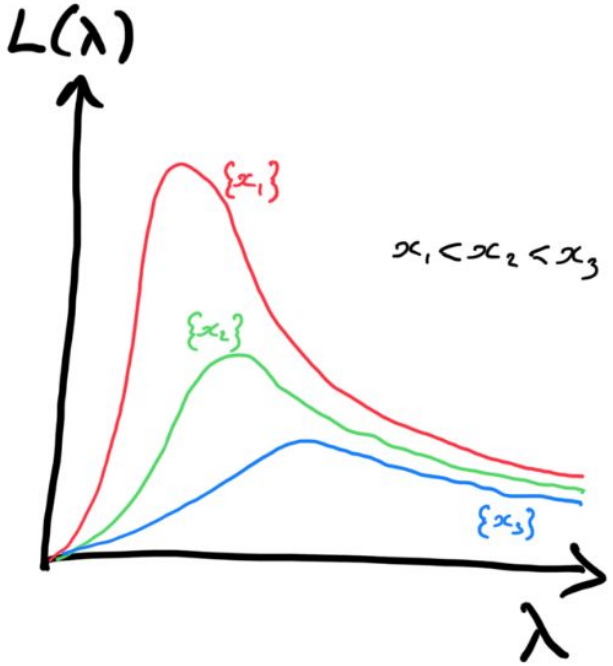
$$P(x) = \frac{\exp(-x/\lambda)}{\lambda}$$

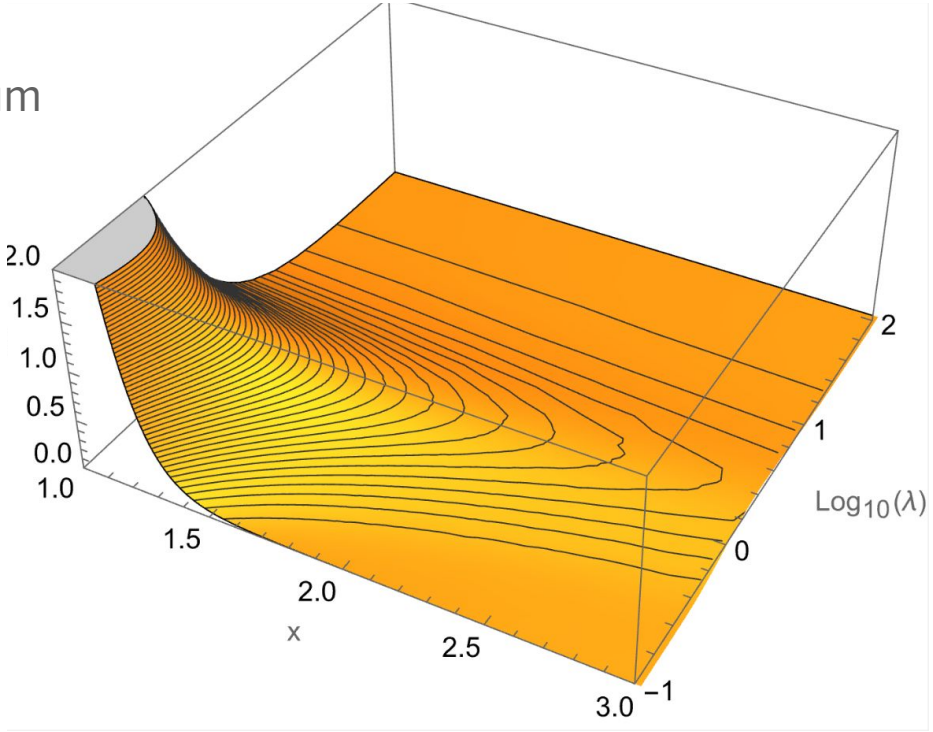
Write down the likelihood and plot/sketch it.



$$\mathcal{L}(\{x_i\}|\lambda) = \frac{\exp(-\sum_{i=1}^{N} x_i/\lambda)}{\lambda}$$

= the sum





A simple example (conjugate priors)

Exercise:

Suppose we want to measure the number density of stars, S, in a path patch of sky. A survey find $n=5$ stars in an area $A=1$ square degrees.

What is the likelihood?

We have to choose a prior - things work especially nicely for a gamma distribution

$$\pi(S) = \frac{S^{k-1}\exp(-S/\theta)}{\Gamma(k)\theta^k}$$

Gamma dist with shape and scale parameters k and θ

What is the posterior?

A simple example (conjugate priors)

Exercise:

Suppose we want to measure the number density of stars, S, in a path patch of sky. A survey find $n=5$ stars in an area $A=1$ square degrees.

What is the likelihood? Poisson: $$\mathcal{L}(n|S) = \frac{(AS)^n \exp(-AS)}{n!}$$

We have to choose a prior - things work especially nicely for a gamma distribution

$$\pi(S) = \frac{S^{k-1}\exp(-S/\theta)}{\Gamma(k)\theta^k}$$

Gamma dist with shape and scale parameters k and θ

What is the posterior?

$$P(S|n) = \frac{S^{k'-1}\exp(-S/\theta')}{\Gamma(k')\theta'^{k'}}$$

Another Gamma dist with params $k'=k+n$ and $\theta'=\theta/(A\theta+1)$

A simple example (conjugate priors)

This is an example of a conjugate prior - this is when the prior and posterior lie in the same parametric family of probability distributions

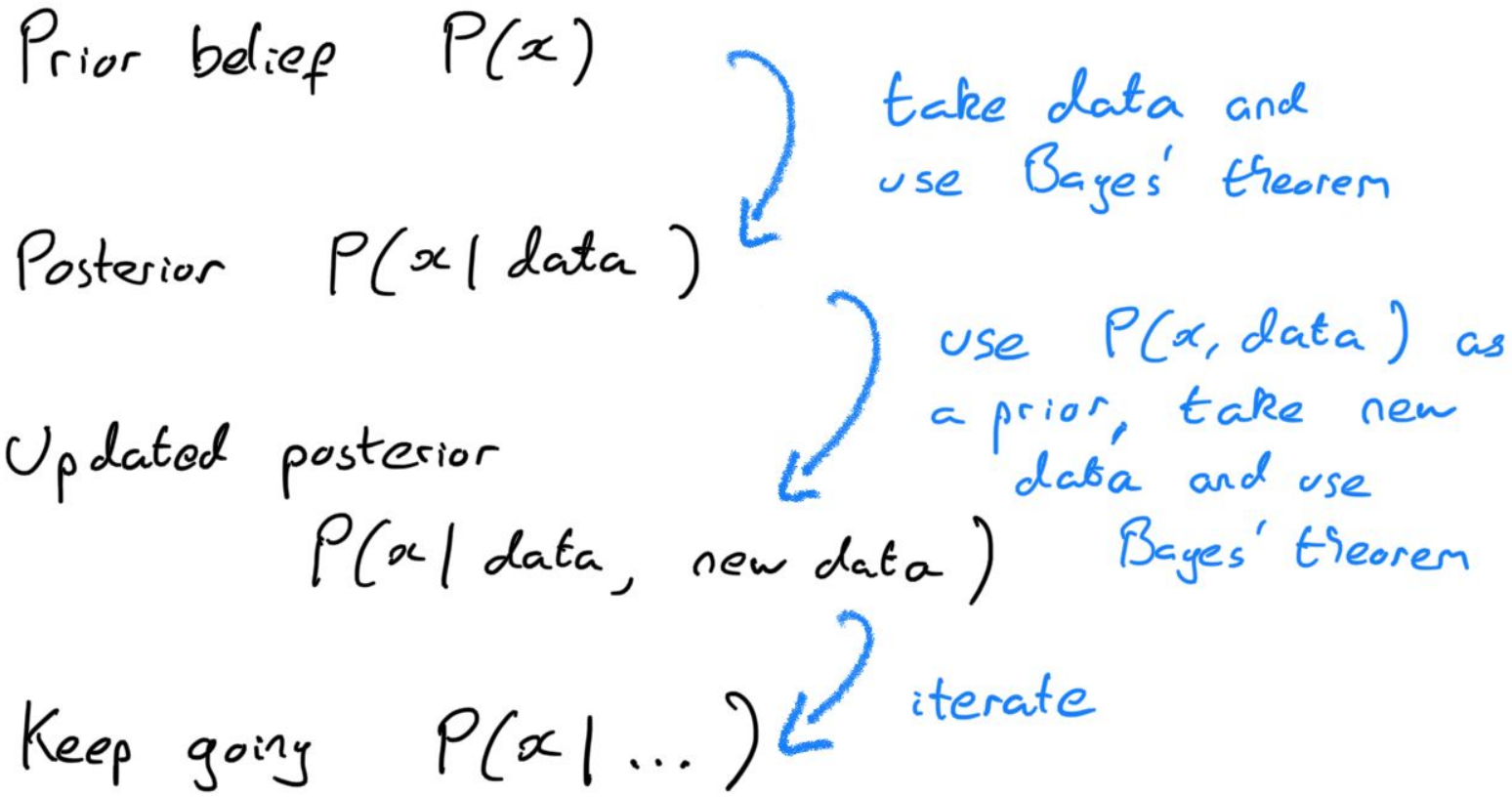
$$
\begin{aligned}
&\text{If } S \sim \text{Gamma}(k, \theta)\,, && \text{(Prior)} && \text{(i)}\\
&\text{and } n|S \sim \text{Poisson}(AS)\,, && \text{(Likelihood)} && \text{(ii)}\\
&\text{then } S|n \sim \text{Gamma}\left(k+n, \frac{\theta}{A\theta+1}\right)\,. && \text{(Posterior)} && \text{(iii)}
\end{aligned}
$$

The Gamma distribution is the conjugate prior to the Poisson likelihood function

Conjugate priors allow the process of Bayesian inference to be done analytically

Conjugate priors also illustrate the iterative nature of Bayesian inference where the likelihood acts inside Bayes' theorem to update our prior state of knowledge

Bayesian inference as an iterative process



Stochastic Samples

The main target of a Bayesian inference is usually the posterior

Models used in research in the modern physical sciences are usually complex, high dimensional and non linear

Q: How to study the resulting posteriors?

Stochastic Samples

The main target of a Bayesian inference is usually the posterior

Models used in research in the modern physical sciences are usually complex, high dimensional and non linear

Q: How to study the resulting posteriors?

A: Draw a large number of i.i.d. samples from the posterior

$$\theta_i \overset{\text{iid}}{\sim} P(\theta|d) \quad \text{for } i = 1, 2, \ldots, N$$

Such samples can be used to answer virtually any questions about the posterior

- they can approximate the full n-dimensional distribution; e.g. by a KDE
- visualising the marginal distributions; e.g. in a corner plot
- computing credible intervals (1D) or regions (2D)
- approximating integrals of the form $\int \mathrm{d}\theta \; P(\theta|d) f(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} f(\theta_i)$

Monte Carlo Methods

Two main classes of stochastic sampling algorithms used for Bayesian inference

Markov Chain Monte Carlo

Nested Sampling

MCMC

Definition. A *Markov chain* is a ordered sequence (or *chain*) of random points $x_0, x_1, x_2, \ldots$ in the sample space (i.e. $x_i \in \mathcal{X}$) that satisfies the *Markov property*,

$$P(x_{i+1}|x_0, x_2, \ldots, x_i) = P(x_{i+1}|x_i).$$

The defining feature, the next point in the chain depends only on the current point, not on any of the previous

The chain is specified by its starting point, x_0, and the transition probabilities

The chains that we will consider are *time homogenous*

$$P(x|x') \equiv \rho(x', x)$$

MCMC

We are usually interested in the long term behaviour of the chain - what will happen after a large number of iterations?

The chains that we will consider are all *irreducible* (sometimes called *ergodic*)

Definition. A Markov chain is *irreducible* if for any starting point $x_0 \in \mathcal{X}$ for any (measurable) region $A \subset \mathcal{X}$ there exists an integer $n \geq 1$ such that $\int_A \mathrm{d}x \; P(x_n|x_0) > 0$.

For time-homogeneous and irreducible chains, the long term behaviour of the chain does not depend on the starting point, x_0, it depends only on the transition probabilities $\rho(x', x)$

MCMC

We can control the long term behaviour of the chain by designing the transition probability $\rho(x', x)$

Definition. A time-homogeneous Markov chain is said to approach a *limiting distribution* λ on the space $\mathcal{X}$ if

$$\lim_{n\to\infty} P(x_n|x_0) = \lambda(x_n).$$

The limiting distribution must be *stationary* - i.e. if the chain is already in this distribution then further iterations will preserve this distribution

Definition. Given a Markov chain with transition probabilities $\rho(x', x)$, a distribution π on $\mathcal{X}$ is said to be a *stationary distribution* of the Markov chain if

$$\pi(x') = \int \mathrm{d}x\ \pi(x)\rho(x', x).$$

MCMC

In order for any of this to be useful, we now need a practical way of designing the transition probabilities $\rho(x', x)$ so that we can make the chain converge to a distribution of our choice… in a Bayesian analysis, this is the posterior

This is usually done using *detailed balance*

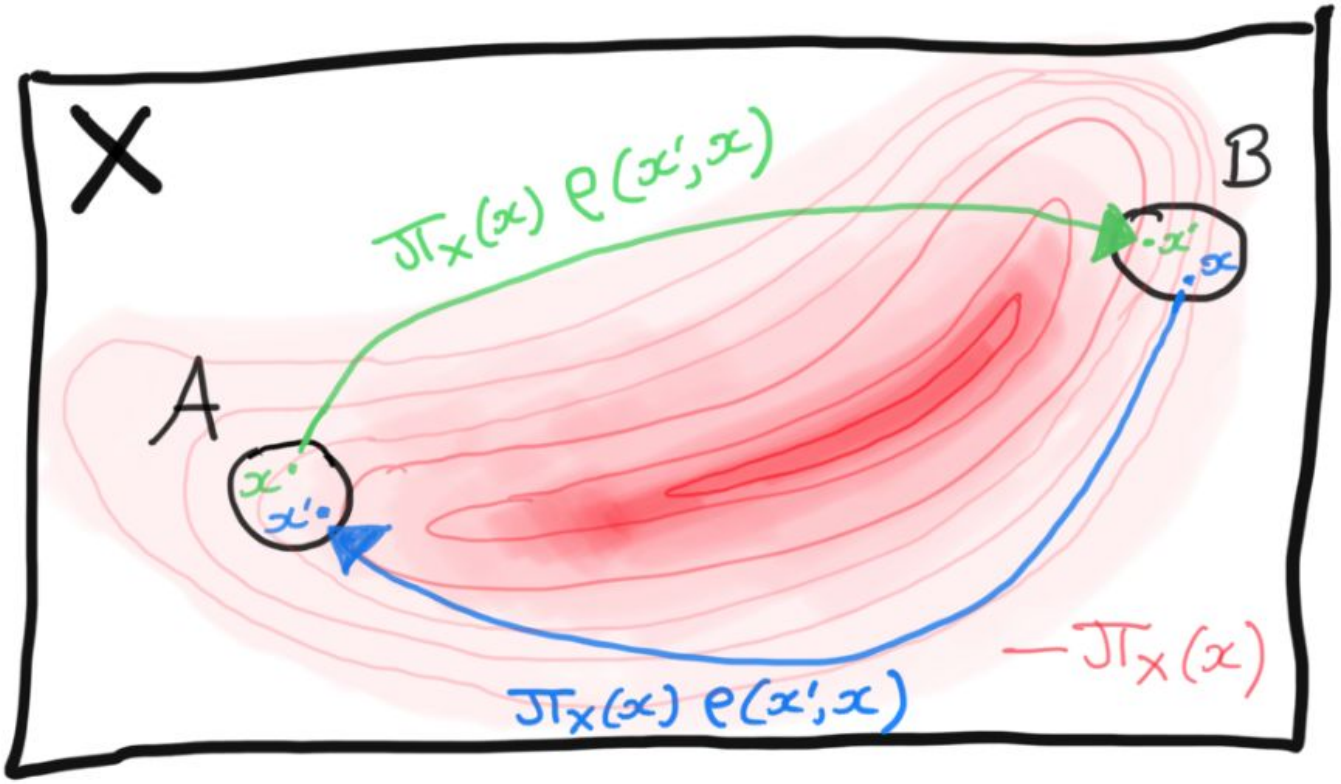
Definition. Consider a time-homogeneous Markov chain with transition probabilities $\rho(x', x)$. The chain is said to satisfy *detailed balance* with respect to π if

$$\pi(x)\rho(x', x) = \pi(x')\rho(x, x').$$

If a Markov chain satisfies detailed balance with respect to π then π is a stationary distribution of the Markov chain

MCMC

Illustration of detailed balance



MCMC

If we can design transition probabilities $\rho(x', x)$ for a time-homogeneous, irreducible Markov chain s.t. they satisfy detailed balance with $\pi = P$, then after evolving for enough iterations the distribution of the chain will approach P.

MCMC

The simplest, most widely studied and historically the most important stochastic sampling algorithm is Metropolis-Hastings

The most important ingredient is the *proposal* distribution $Q(y|x)$.

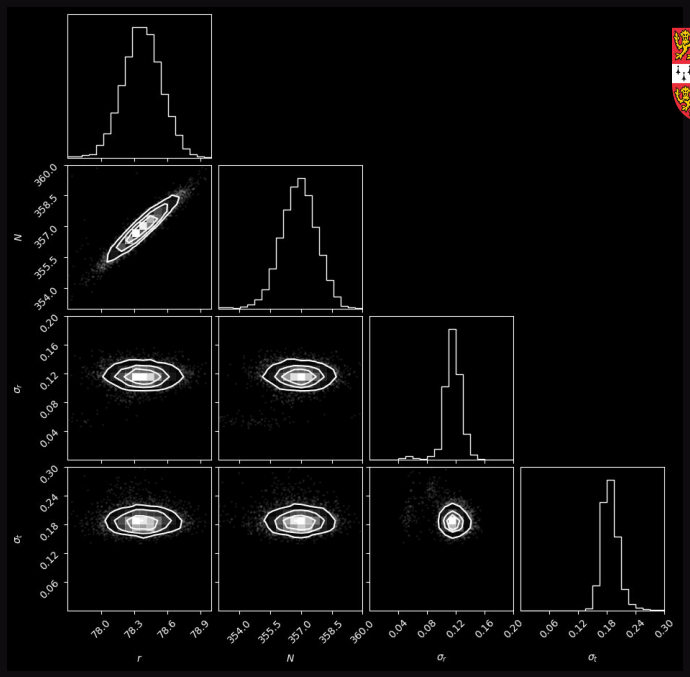
If the chain is currently at position x_i, this is used to propose possibilities for the next chain position, $y \sim Q(y|x_i)$, which is either accepted ($x_{i+1} = y$) or rejected ($x_{i+1} = y$) with probability

$$a(y, x) = \min(1, \frac{P(y)Q(x|y)}{P(x)Q(y|x)})$$

MCMC

Algorithm Metropolis Hastings

$x_0 \sim \alpha$ ▷ Initialise
$i \leftarrow 0$
while $i \geq 0$ **do** ▷ Iterate $i = 0, 1, 2, \ldots$
    $y \sim Q(y|x_i)$ ▷ Proposal
    $a \leftarrow (P(y)Q(x_i|y)) \,/\, (P(x_i)Q(y|x_i))$ ▷ MH acceptance probability
    $u \sim \mathcal{U}(0, 1)$
    if $u < a$ **then**
        $x_{i+1} \leftarrow y$ ▷ Accept
    else
        $x_{i+1} \leftarrow x_i$ ▷ Reject
    end if
    $i \leftarrow i + 1$
end while



Department of Applied Mathematics and Theoretical Physics (DAMTP)

Bayesian Statistics - Hands-on Session

AIPHY school, Monopoli, Sep 30th - Oct 4th 2024

Christopher J. Moore

The Antikythera Mechanism

Ancient Greek mechanical calculator, or *orrery*

Dates from circa 100 BC

Discovered in a shipwreck in 1901

Used for modelling the motions of the Sun, Moon and planets and for predicting other astronomical phenomena, such as eclipses

However, precise functionality unknown and must be deduced from analysis of incomplete and damaged remains

Impressively complex and well manufactured

The Antikythera Mechanism

X-ray image of the *calendar ring*

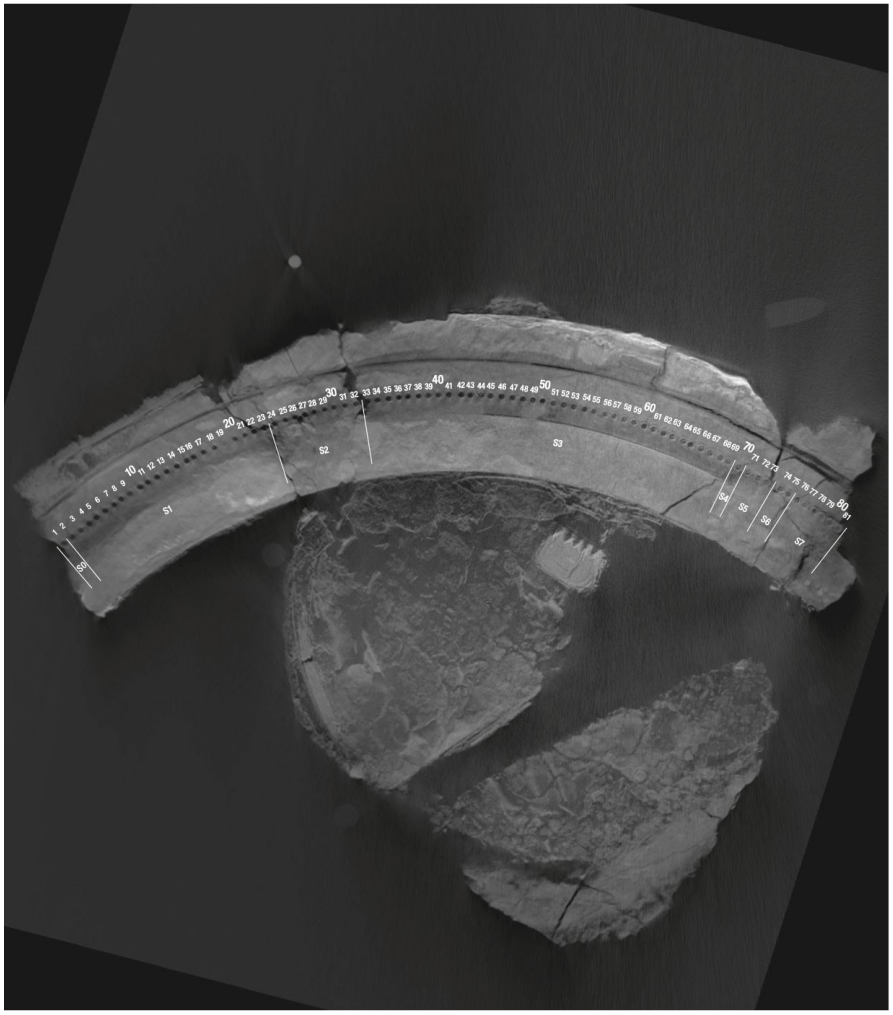
Known to be a calendar from nearby engravings of the names of the months

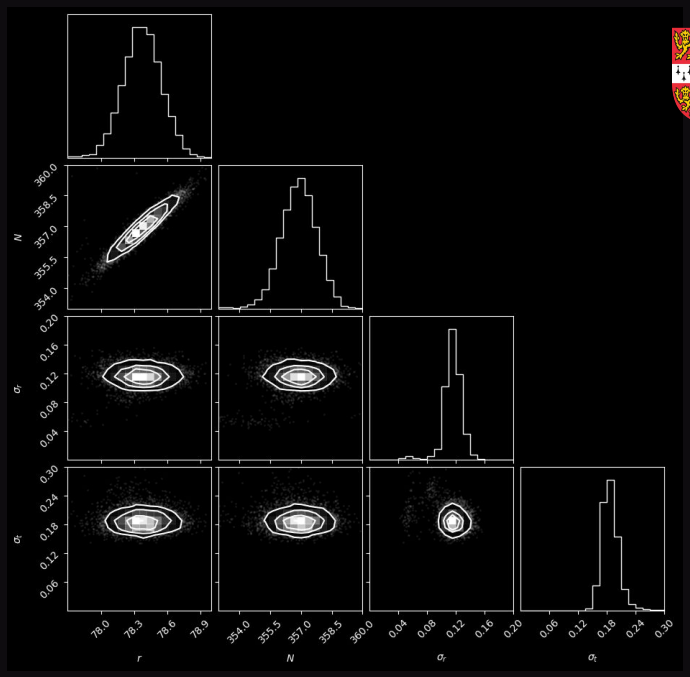
Approx. 25% of the ring survives, split into several sections

Originally thought to be a solar calendar with 365 holes around circumference

Recently suggested actually a lunar calendar with 354 holes

Can use Bayesian inference to answer the question of the number of holes





Department of Applied Mathematics and Theoretical Physics (DAMTP)

Bayesian Statistics - Lecture 2, Thursday

AIPHY school, Monopoli, Sep 30th - Oct 4th 2024

Christopher J. Moore

Outline

Bayes' Theorem

Bayesian versus frequentist views on probability

Stochastic Samples

Monte Carlo Methods

Bayesian Model Selection, the role of the evidence

Methods for calculating the evidence

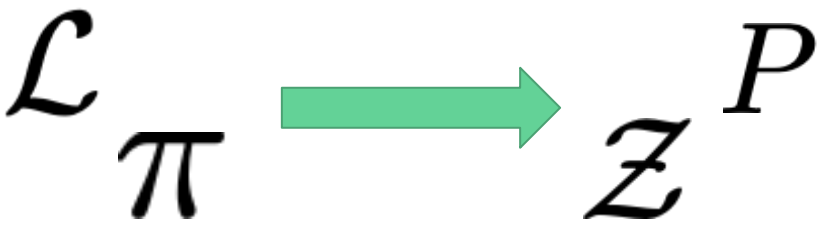
Bayes' Theorem

$$P(\theta|d) = \frac{1}{\mathcal{Z}(d)}\pi(\theta)\mathcal{L}(d|\theta)$$

Posterior Evidence Prior Likelihood

Let the "events" be our model parameters and the observed data

In a Bayesian inference, the likelihood and prior are the inputs and the posterior (and sometimes the evidence) are the outputs of the analysis



Bayes' Theorem

$$P(\theta|d,\mathrm{mod}) = \frac{1}{\mathcal{Z}(d|\mathrm{mod})}\pi(\theta|\mathrm{mod})\mathcal{L}(d|\theta,\mathrm{mod})$$

Posterior | Evidence | Prior | Likelihood

Previously, we hid the dependence of all probabilities on some important assumptions in our notations - These assumptions include the model

The evidence is just a normalising constant, but it depends on the model

The assumed model is now shown explicitly in our notation

$$\mathcal{Z}_{\mathrm{mod}} \equiv P(d|\mathrm{mod}) = \int \mathrm{d}\theta\, \pi(\theta|\mathrm{mod})\mathcal{L}(d|\mathrm{mod})$$

Bayesian Model Comparison

Why not just pick the model that gives the best fit to the data?

I.e. the maximum likelihood model.

Bayesian Model Comparison

Why not just pick the model that gives the best fit to the data?

I.e. the maximum likelihood model.

Answer: Occam's razor

The simplest explanation is usually the best one

The Bayesian solution to the model comparison naturally incorporates an appropriate penalty for more complicated model

$$\text{Posterior Odds Ratio} \equiv \mathcal{O}_{A,B} = \frac{P(A|D)}{P(B|D)}$$

The Posterior Odds

$$
\begin{aligned}
\mathcal{O}_{A,B} &= \frac{P(A|d)}{P(B|d)} && \text{(definition of odds ratio)} \\
&= \frac{P(d|A)}{P(d|B)} \times \frac{P(A)}{P(B)} && \text{(use Bayes' theorem)} \\
&= \frac{\int \mathrm{d}\theta_A \, P(d, \theta_A|A)}{\int \mathrm{d}\theta_B \, P(d, \theta_B|B)} \times \frac{P(A)}{P(B)} && \text{(use law of total probability)} \\
&= \frac{\int \mathrm{d}\theta_A \, P(d|\theta_A, A) P(\theta_A|A)}{\int \mathrm{d}\theta_B \, P(d|\theta_B, B) P(\theta_B|B)} \times \frac{P(A)}{P(B)} && \text{(use Bayes' theorem again)} \\
&= \frac{\int \mathrm{d}\theta_A \, \mathcal{L}(d|\theta_A, A) \pi(\theta_A|A)}{\int \mathrm{d}\theta_B \, \mathcal{L}(d|\theta_B, B) \pi(\theta_B|B)} \times \frac{P(A)}{P(B)} && \text{(change notation)} \\
&= \frac{\mathcal{Z}_A}{\mathcal{Z}_B} \times \frac{P(A)}{P(B)} && \text{(use definition of evidence)}
\end{aligned}
$$

Bayesian Updating

We have just shown that the posterior model odds ratio is given by

$$\mathcal{O}_{A,B} = \frac{\mathcal{Z}_A}{\mathcal{Z}_B} \times \frac{P(A)}{P(B)}$$

Again, we see that Bayesian inference can be thought of a systematic way of updating our beliefs, this time by multiplying by the evidence ratio

posterior odds ratio = evidence ratio × prior odds ratio

The Evidence

As we've seen, the evidence is the key ingredient in Bayesian model comparison

$$\mathcal{Z}_{\rm mod} \equiv P(d|{\rm mod}) = \int {\rm d}\theta \; \pi(\theta|{\rm mod}) \mathcal{L}(d|{\rm mod})$$

Unfortunately, this is usually hard to calculate, especially for high-dimensional, models with complicated, multimodal posteriors

Methods for Calculating the Evidence

We need a method for evaluating integrals of this kind

$$\mathcal{Z}_{\mathrm{mod}} \equiv P(d|\mathrm{mod}) = \int \mathrm{d}\theta \, \pi(\theta|\mathrm{mod}) \mathcal{L}(d|\mathrm{mod})$$

We will briefly discuss a few options:

- Analytically, e.g. using a conjugate prior
- The Laplace approximation
- “Normal” numerical integration, e.g. quadrature methods, or trapezium rule
- What about MCMC?
- Nested sampling

Laplace Approximation

If the data is highly informative, i.e. in the limit of large signal-to-noise ratio, the posterior may be approximately Gaussian, at least around the peak

$$\log \mathcal{L}(d|\theta) = \log \mathcal{L}(d|\hat{\theta}) - \frac{1}{2} \sum_{\mu\nu} C_{\mu\nu} (\theta - \hat{\theta})^{\mu} (\theta - \hat{\theta})^{\nu}$$

where

$$C_{\mu\nu} = -\frac{\partial^2}{\partial\theta^{\mu}\partial\theta^{\nu}}\bigg|_{\theta=\hat{\theta}} \log \mathcal{L}(d|\theta)$$

$$\log \pi(\theta) = \text{const}$$

Finding the evidence is then just a simple Gaussian integral

$$\mathcal{Z} \approx \text{const} \times \mathcal{L}(d|\hat{\theta}) \sqrt{\frac{(2\pi)^{\dim}}{\det \boldsymbol{C}}}$$

“Normal” Numerical integration

We need a method for evaluating integrals of this kind

$$\mathcal{Z}_{\rm mod} \equiv P(d|{\rm mod}) = \int {\rm d}\theta\, \pi(\theta|{\rm mod})\mathcal{L}(d|{\rm mod})$$

“This is just a numerical integral, what’s the big deal? Use the trapezium rule!”

Standard numerical integration routines (such as trapezium or quadrature rules) scale very poorly with the dimensionality of the space of θ

If we use 100 regularly spaced points along dimension, total points = $100^{\dim(\theta)}$

Things are even worse when the *typical set* of the posterior is small compared to the prior, we may need much more that 100 points along each dimension

What about MCMC?

Recall, MCMC methods are able to generate stochastic samples

$$\theta_i \overset{\text{iid}}{\sim} P(\theta|d) \quad \text{for } i = 1, 2, \ldots, N$$

These are great for visualising the shape of the posterior, but cannot be used to the normalising evidence

(There are variations of MCMC that do compute the evidence using *thermodynamic integration*)

Methods for Calculating the Evidence

We need a method for evaluating integrals of this kind

$$\mathcal{Z}_{\mathrm{mod}} \equiv P(d|\mathrm{mod}) = \int \mathrm{d}\theta \, \pi(\theta|\mathrm{mod}) \mathcal{L}(d|\mathrm{mod})$$

We will briefly discuss a few options:

- Analytically, e.g. using a conjugate prior — **Great, but hardly ever possible**
- The Laplace approximation — **OK when posterior ≃ normal**
- “Normal” numerical integration — **Only in low dimensions**
- What about MCMC? — **Thermodynamic integration**
- **Nested sampling**

Nested Sampling

Nested sampling is an algorithm to calculate the *evidence*, integrals of the form

$$Z = \int \mathrm{d}x \; \mathcal{L}(d|x)\pi(x)$$

We are especially interested in high-dimensional problems where this integral cannot be evaluated (or suitably approximated) analytically

Let $\xi(L)$ be the prior probability associated with likelihoods greater than L

$$\xi(L) = \int_{\mathcal{L}>L} \mathrm{d}x \; \pi(x)$$

This is a monotonically decreasing function satisfying

$$\xi(0) = 1 \quad \text{and} \quad \xi(\mathcal{L}_{\max}) = 0$$

Nested Sampling

Illustration of $\xi(L) = \int_{\mathcal{L}>L} \mathrm{d}x\ \pi(x)$

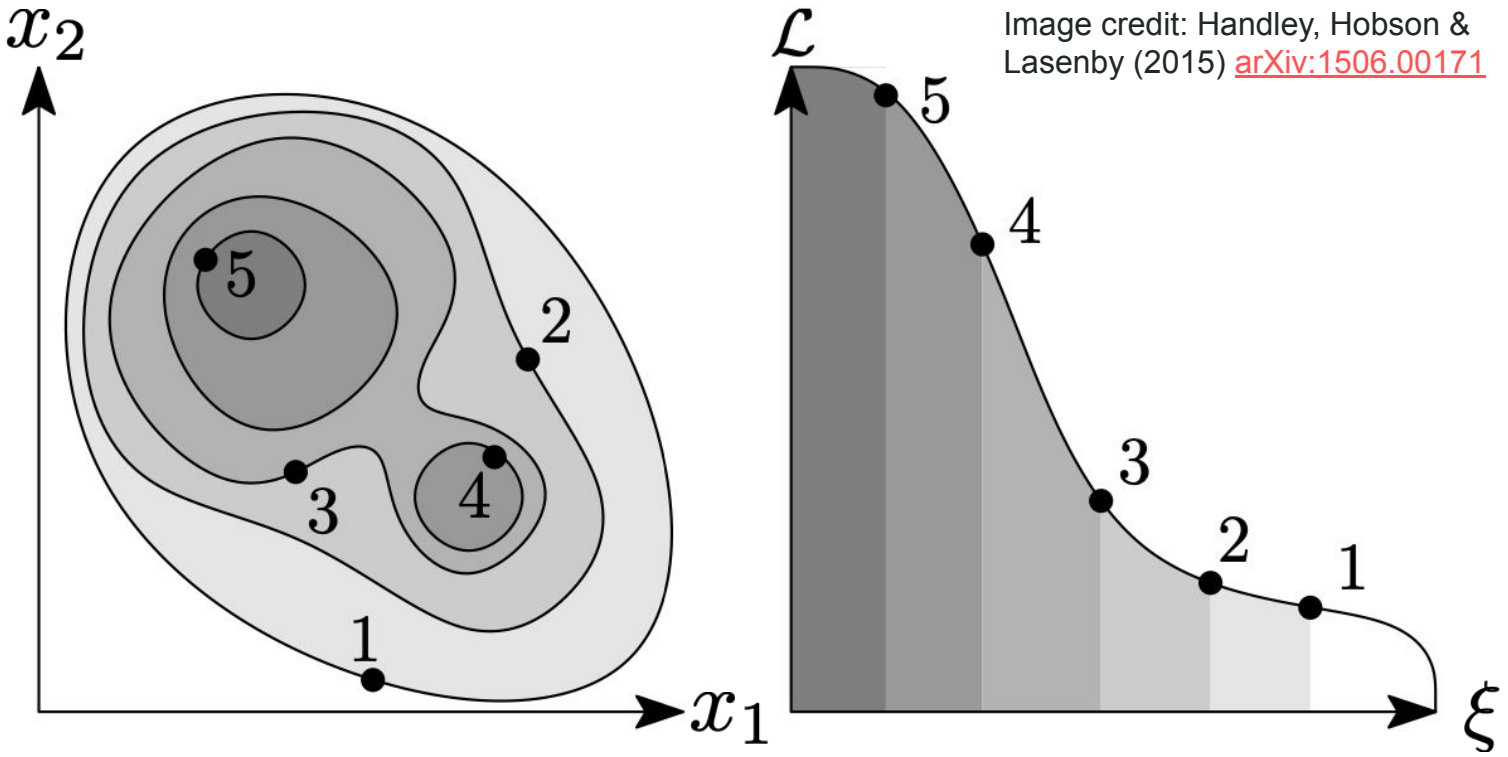


Image credit: Handley, Hobson & Lasenby (2015) arXiv:1506.00171

Nested Sampling

We can now write the evidence as $$\mathcal{Z} = \int_0^1 \mathrm{d}\xi \; L(\xi)$$

At first, this appears somewhat miraculous - we have changed a high-dimensional integral into a 1-dimensional integral

Of course, we have just moved the difficulty into finding and inverting the function $\xi(L)$

Given a sequence of points in the prior volume, $0 < \xi_M < \xi_{M-1} < \ldots < \xi_1 < 1$, and the likelihoods $L_i = L(\xi_i)$, the evidence can be approximated using the *trapezium rule*

$$Z \approx \frac{1}{2} \sum_{i=1}^{M} (\xi_{i-1} - \xi_{i+1}) L_i$$

Nested Sampling

Algorithm The nested sampling algorithm

for $j = 0, 1, \ldots, N_{\text{live}} - 1$ **do**
    $x_j \sim \pi$ ▷ Initialise live points from prior
    $L_j \leftarrow \mathcal{L}(d|x_j)$
end for
samples $\leftarrow$ [] ▷ Empty list to store weighted samples
$i \leftarrow 0$
while $i \geq 0$ **do** ▷ Iterate $i = 0, 1, 2, \ldots$
    $L_i \leftarrow \min(L_0, L_1, \ldots, L_{N_{\text{live}}-1})$ ▷ Current worst live point
    samples $\leftarrow$ Concatenate(samples, $[w, x]$) ▷ Append weighted sample to list
    $i \leftarrow i + 1$
end while

Nested Sampling

The main challenge when designing a nested sampling algorithm is drawing samples from the constrained prior

There are many ways this can be done:

- Rejection sampling
- MCMC — CPnest
- Slice sampling — PolyChord
- Bounding distribution — e.g. ellipsoidal sampling, Multinest
- Galilean Monte Carlo — Feroz+Skilling (2013)
- Normalising flows — Nessai

Nested Sampling Demo

For simplicity, we will demonstrate nested sampling in a 2-dimension unit square, (x_0, x_1), with a uniform prior

$$\pi(\mathbf{x}) = \begin{cases} 1 & \text{if } 0 \leq x_0, x_1 \leq 1 \\ 0 & \text{else} \end{cases}$$

We will explore the following simple likelihood function with ϵ=2

$$\mathcal{L}(d|\mathbf{x}) = ((x_0 x_1 (1 - x_0)(1 - x_1))^{\epsilon}$$

For this toy problem, the evidence can be calculated analytically in terms of the *beta function* - We can use this as a check on our nested sampling

$$Z = \int_0^1 \mathrm{d}x_0 \int_0^1 \mathrm{d}x_1 \, \mathcal{L}(d|\mathbf{x})\pi(\mathbf{x}) = B(\epsilon + 1, \epsilon + 1)^2$$