PART 3: AN INTRODUCTION TO BAYESIAN INFERENCE

Matthew Pitkin

GraWIToN School

25 March 2015

University of Glasgow

OVERVIEW

In this part of the course we will discuss

- parameter estimation
- hypothesis testing

from a Bayesian perspective.

FREQUENTIST VS. BAYESIAN APPROACH

Recall that:

- Frequentist
  - data are a **repeatable** random sample
  - parameters **remain fixed** during this repeatable process
- Bayesian
  - data are an **observation from the realised sample**
  - parameters are unknown and described **probabilistically**

THE BAYESIAN APPROACH

In the Bayesian approach, we can test our model given our data (e.g. rolls of a die) and see how our knowledge of its parameters evolve, for any sample size, considering only the data that we *did* actually observe.

All we need is Bayes theorem:

$$\overbrace{p(\text{model}|\text{data}, I)}^{\text{Posterior}} \propto \overbrace{p(\text{data}|\text{model}, I)}^{\text{Likelihood}} \times \overbrace{p(\text{model}|I)}^{\text{Prior}}$$

- **prior**: *what we knew before*
- **likelihood**: *the influence of our observations*
- **posterior**: *what we know now*

A BAYESIAN EXAMPLE: IS A COIN FAIR?

How can we determine if a coin is fair[1]? We can consider a large number of contiguous propositions over the range in which the bias weighting H of the coin might lie:

- $H = 0$ coin produces a tail every time
- $H = 1$ coin produces a head every time
- $H = 0.5$ is a 'fair' coin with 50:50 chance of head or tail
- continuum of probabilities $0 \leq H \leq 1$

Given some data (an observed number of coin tosses) we can assess how much we believe each of these propositions (e.g. $0 \leq H < 0.01$, $0.01 \leq H < 0.02$, and so on) to be true, e.g.

$$\text{Prob}(0 \leq H < 0.01|d)$$

[1] See e.g. Chap. 2 of Sivia [1].

A BAYESIAN EXAMPLE: IS A COIN FAIR?

In the limiting case where our propositions each lie in the infinitesimal range $\mathrm{d}H$ our inference about the bias weighting is summarised by the pdf for the conditional probability $p(H|d, I)$, i.e. the *posterior*. We can use Bayes' theorem to calculate it.

For coin flips, assuming that they are independent events, the probability of obtaining 'r heads in n tosses' is given by the binomial distribution, so our *likelihood* is:

$$p(d|H, I) \propto H^r(1 - H)^{n-r}$$

But, what should we use as our *prior*?

A BAYESIAN EXAMPLE: IS A COIN FAIR?

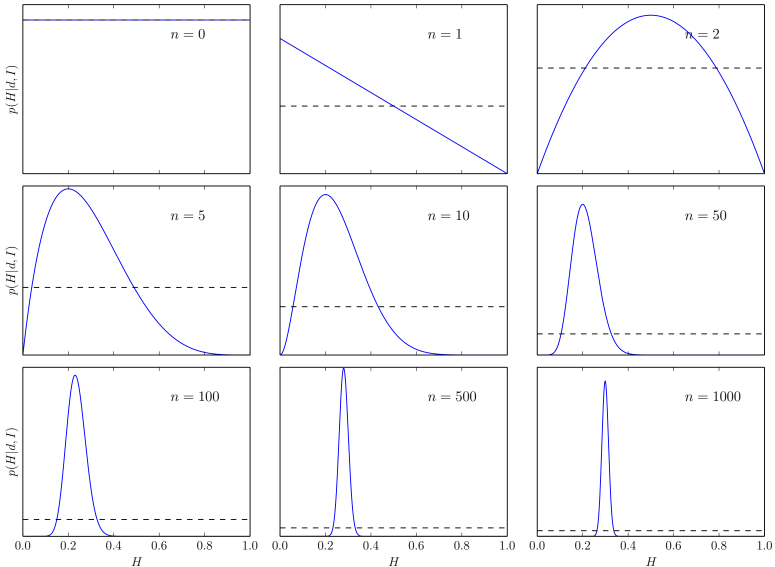
But, what should we use as our *prior*?

Assuming we have no knowledge about the provenance of the coin, or the person tossing it, and want to reflect total ignorance of the possible bias, then a simple probability reflecting this is a **uniform**, or **flat**, pdf:

$$p(H|I) = \begin{cases} 1, \text{if } 0 \leq H \leq 1, \\ 0, \text{otherwise.} \end{cases}$$

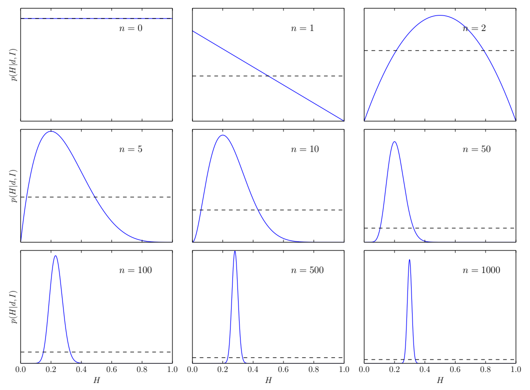
Using these we can calculate our posterior, $p(H|d, I)$, as we obtain more data (counting r as the number of coin tosses, n, increases).

A BAYESIAN EXAMPLE: IS A COIN FAIR?



A BAYESIAN EXAMPLE: IS A COIN FAIR?

As the number of coin tosses increases the posterior evolves from the uniform prior to a tight range in H with the most probable value being $H = 0.3$.



A BAYESIAN EXAMPLE: IS A COIN FAIR?

What about a *different* prior?

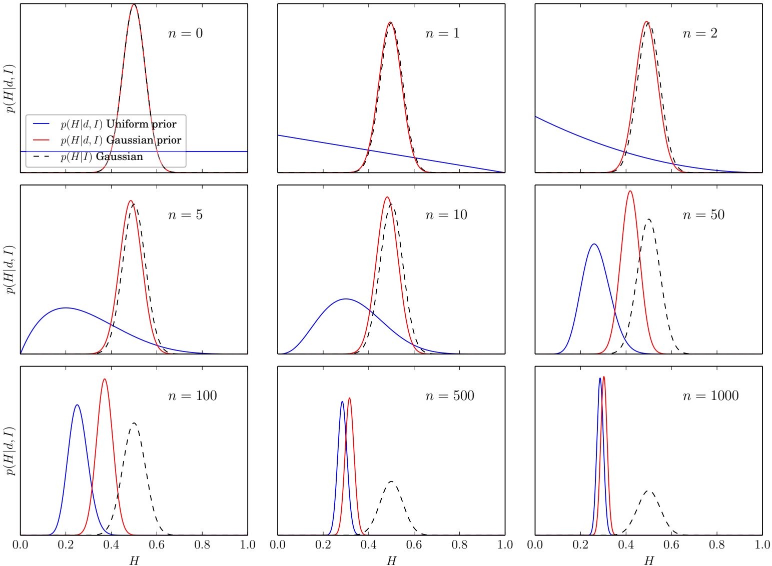
We know that coins are generally fair, so what if we assume this one is too?

We can assign a Gaussian prior distribution that focusses the probability around the expected ‘fair coin’ value

$$p(H|I) \propto \exp\left(-\frac{1}{2}\frac{(H-\mu_H)^2}{\sigma_H^2}\right),$$

with $\sigma_H = 0.05$ and $\mu_H = 0.5$.

A BAYESIAN EXAMPLE: IS A COIN FAIR?



A BAYESIAN EXAMPLE

What do we learn from this?

- As our data improve (i.e. we gather more samples), the posterior pdf narrows and becomes less sensitive to our choice of prior (i.e. the likelihood starts to dominate)
- The posterior conveys our (evolving) degree of belief in different values of H given our data
- If we want to express our belief as a **single number** we can adopt e.g. the mean, median or mode
- We can use the **variance** of the posterior to assign and *error* for H
- It is very straighforward to define *Bayesian confidence intervals* (more correctly termed **credible intervals**)

BAYESIAN CREDIBLE INTERVAL

We define a **credible interval** $[\theta_a, \theta_b]$ as a (*non-unique*) range that a contains a certain amount of posterior probability, X,

$$X = \int_{\theta_a}^{\theta_b} p(\theta|d, I)\mathrm{d}\theta.$$

If $X = 0.95$ then we can find $[\theta_a, \theta_b]$ that e.g. gives the minimum range containing 95% of the probability.

The meaning of this is simple: *we are 95% sure that θ lies between θ_a and θ_b.*

This is just based on the data at hand and requires no assumptions about a frequency of measuring a statistic over multiple trials.

EXAMPLE: FITTING A LINE

Returning to the problem of fitting a line $y = mx + c$ to data, d, we can write the posterior for the parameters

$$p(m, c|d, I) \propto \underbrace{p(d|m, c, I)}_{\text{Likelihood}} \times \underbrace{p(m, c|I)}_{\text{Prior}} .$$

If the prior on the parameters is uniform and independent, so

$$p(m, c|I) = p(m|I)p(c|I) = \text{constant},$$

then the posterior is

$$p(m, c|d, I) \propto p(d|m, c, I)$$

and we can use the machinery of maximum likelihood to estimate the parameters (i.e. maximum likelihood can be derived from Bayesian reasoning with certain priors).

EXAMPLE: FITTING A LINE

However, we will use this to show to general concept of fitting any (even *non-linear*) model. If the likelihood is Gaussian, with known values of σ_i, then

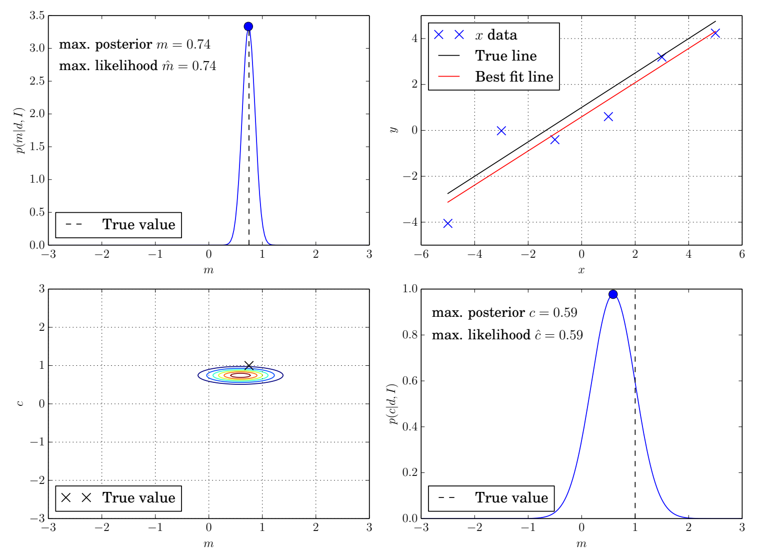
$$p(m, c|d, I) = p(m, c|I) \left(\frac{1}{2\pi\sigma_i^2}\right)^{n/2} \exp\left(-\sum_{i=1}^{n} \frac{[d_i - (mx_i + c)]^2}{2\sigma_i^2}\right)$$

and we can evaluate the posterior over a grid in the parameters m and c.

We can also compute the marginal posteriors on m and c as, e.g.

$$p(m|d, I) = \int_{-\infty}^{\infty} p(m, c|d, I) \mathrm{d}c.$$

EXAMPLE: FITTING A LINE



In the above example, in practice, when $p(m, c|I) =$ constant and $\sigma_i = \sigma$ are constant, we can just calculate the posterior[2] over a grid in m and c

$$\ell(m_{i_m}, c_{i_c}) = \ln p(m_{i_m}, c_{i_c}|d, I) = -\sum_{i=1}^{n} \frac{[d_i - (m_{i_m} x_i + c_{i_c})]^2}{2\sigma^2}$$

and get the marginal posteriors through numerical integration, e.g.

$$p(m_{i_m}|d, I) \propto \sum_{i_c}^{n_c} \exp\left(\ell(m_{i_m}, c_{i_c}) - \text{max}\ell(m, c)\right)\Delta c$$

where Δc are the grid step sizes in c (or you could use the trapezium rule for more acurracy).

[2] We generally work in natural logarithm space due to numerical precision issues.

EXAMPLE: FITTING A LINE

What if we don't know σ?

In this case we can treat σ as another unknown variable and marginalise over it, e.g.

$$p(m, c|d, I) = p(m, c|I) \int_0^\infty p(d|m, c, \sigma, I) p(\sigma|I) \mathrm{d}\sigma$$

If the likelihood is Gaussian and we assume a flat prior on all parameters, e.g.

$$p(\sigma|I) = \begin{cases} C, \sigma > 0 \\ 0, \sigma \leq 0 \end{cases}$$

Then we have

$$p(m, c|d, I) \propto \int_0^\infty \sigma^{-n} \exp\left(- \sum_{i=1}^{n} \frac{[d_i - (mx_i + c)]^2}{2\sigma^2} \right) \mathrm{d}\sigma$$

EXAMPLE: FITTING A LINE

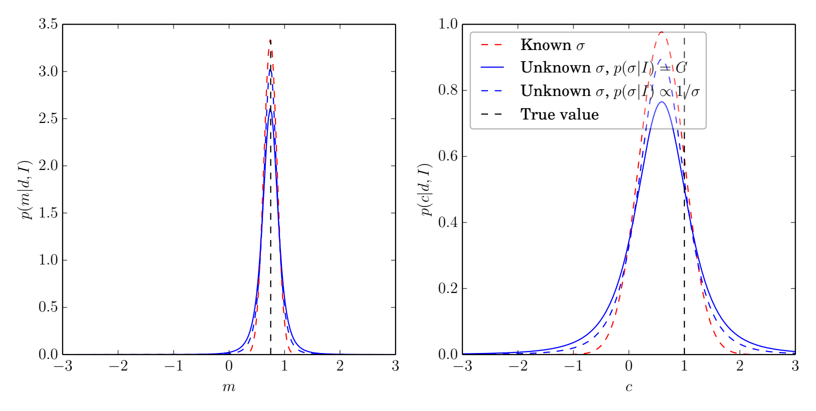
This integral is analytic, and through some substitution (see e.g. Chap. 3 of Sivia [1]), becomes

$$p(m, c|d, I) \propto \left(\sum_{i=1}^{n} [d_i - (mx_i + c)]^2 \right)^{-(n-1)/2}$$

This is essentially a *Student's t-distribution* with $\nu = (n-2)$ degrees of freedom.

Note: if we were instead to use a prior on σ of $p(\sigma|I) \propto 1/\sigma$ it would lead to a Student's t-distribution with $\nu = n - 1$ degrees of freedom.

EXAMPLE: FITTING A LINE



EXAMPLE: SPECTRAL LINE ESTIMATION

We take a simple example: *finding the frequency of a sinusoid in noisy data*[3]. We can model the signal as

$$y(t_i) = C\cos\left(\omega t_i + \phi\right) = \underbrace{A}_{C\cos\phi}\cos\left(\omega t_i\right) + \underbrace{B}_{-C\sin\phi}\sin\left(\omega t_i\right)$$

where $\omega = 2\pi f$ is the angular frequency and A and B are two unknown amplitudes (accounting for an unknown amplitude and initial phase of the signal).

A standard way to do this is performing a Fourier transform, and creating e.g. a power spectrum.

[3] Performing spectral estimation in a Bayesian sense is discussed in great detail in Bretthorst [2]

EXAMPLE: SPECTRAL LINE ESTIMATION

In this case we have four unknowns: A, B, ω, and σ (we assume unknown noise) but we are only interested in ω, so given some data d, we want to calculate

$$p(\omega|d, I) \propto \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{0}^{\infty} p(d|\omega, A, B, \sigma, I) p(\omega, A, B, \sigma|I) \mathrm{d}A \mathrm{d}B \mathrm{d}\sigma$$

where, assuming Gaussian noise on the data, the likelihood is given by

$$\begin{aligned} p(d|\omega, A, B, \sigma, I) &\propto \sigma^{-n} \exp\left(-\sum_{i=1}^{n} \frac{[d_i - y(t_i)]^2}{2\sigma^2}\right) \\ &= \sigma^{-n} \exp\left(-\frac{nQ}{2\sigma^2}\right). \end{aligned}$$

EXAMPLE: SPECTRAL LINE ESTIMATION

The quadratic form is given by

$$Q = \underbrace{\overline{d^2}}_{\frac{1}{n}\sum_{i=1}^{n} d_i^2} - \frac{2}{n}[A \overbrace{R(\omega)}^{\sum_{i=1}^{n} d_i \cos \omega t_i} + B \underbrace{I(\omega)}_{\sum_{i=1}^{n} d_i \sin \omega t_i}] + \frac{1}{2}(A^2 + B^2)$$

where the following approximations have been made

$$\sum_{i=1}^{n} \cos^2 \omega t_i \approx \sum_{i=1}^{n} \sin^2 \omega t_i \approx \frac{n}{2} \text{ and } \sum_{i=1}^{n} \sin \omega t_i \cos \omega t_i \approx 0,$$

which are valid if $n \gg 1$ and the data doesn’t contain very low frequencies.

EXAMPLE: SPECTRAL LINE ESTIMATION

Assuming flat priors on A and B we can analytically integrate them (e.g. with `Mathematica`), giving

$$p(\omega|d, I) \propto \int_0^\infty \sigma^{-n+2} \exp\left[-\frac{n}{2\sigma^2}\{\bar{d}^2 - 2(R^2 + I^2)/n\}\right] p(\sigma|I).$$

If we take $p(\sigma|I) \propto 1/\sigma$, the the intergal over σ is again analytic, again leading to the Student's t-distribution, with $\nu = (n-3)$ degrees of freedom

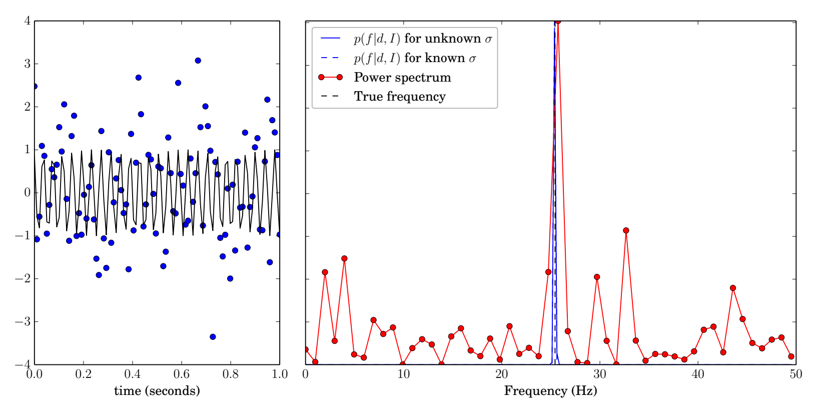
$$p(\omega|d, I) \propto \left[1 - \frac{2(R^2 + I^2)}{n\bar{d}^2}\right]^{-(n-2)/2}$$

Or, if we know σ we would just have

$$p(\omega|d, I) \propto \exp\left(\frac{R^2 + I^2}{\sigma^2}\right)$$

EXAMPLE: SPECTRAL LINE ESTIMATION

We can compare this with a standard power spectrum:



BAYESIAN PARAMETER ESTIMATION

From these examples we see that we could apply it to any model with any number of parameters. However, for these examples we have either had an intrinsically small number of parameters, or been able to analytically marginalise over them. This makes it easy to compute posteriors on a grid.

As the number of parameters increases the number of grid points would have to also increase to keep a decent resolution on each parameter, e.g. a posterior evaluated on a grid of 10 points for each of M parameters, would required 10^M evaluations. This can be prohibitive and we may need to use methods like *Markov chain Monte Carlo*.

DEFINING PROBABILITIES: CHOICE OF PRIOR

Probability (even with frequentist methods) *is* subjective and depends on the available information.

But: Subjective $\neq$ Arbitrary

Bernoulli (1713) set out the ‘Principle of insufficient reason’ and Keynes (1921) the ‘Principle of indifference’:

Given the ***same*** *background information, two observers should assign the* ***same*** *probabilities*

E.g. for a die we should all agree that given it has 6 faces (and no contrary information) we should assign the proposition $X_i \equiv$ “the face on top has i dots” $p(X_i|I) = \frac{1}{6}$ for all i

DEFINING PROBABILITIES: CHOICE OF PRIOR

Can we justify this more fundamentally?

In the case of the die, we could have assigned the proposition $X_i \equiv$ “the face on top has $7-i$ dots”, but we *should still have* $p(X_i|I) = \frac{1}{6}$ for all i

If we extend this to the continuum case, setting x to be a **location parameter**[4], the ‘principle of indifference’ means we should have

$$p(x|I)\mathrm{d}x = p(x+\Delta|I)\mathrm{d}(x+\Delta)$$

where Δ is a constant. Since $\mathrm{d}(x+\Delta)/\mathrm{d}x = 1$, we must have

$$\boxed{p(x|I) = \text{constant}}$$

[4] E.g. a parameter that can be defined over positive and negative values.

DEFINING PROBABILITIES: CHOICE OF PRIOR

Similarly, if we let s be a **scale parameter**[5], the ‘principle of indifference’ means we should have

$$p(s|I)\mathrm{d}s = p(\beta s|I)\mathrm{d}(\beta s)$$

where β is a positive constant (e.g. a conversion factor between two sets of units). Since $\mathrm{d}(\beta s)/\mathrm{d}s = \beta$ we must have

$$\frac{p(s|I)}{p(\beta s|I)} = \beta,$$

so,

$$\underbrace{\boxed{p(s|I) \propto \frac{1}{s}}}_{\text{Jeffreys' prior}}$$

[5] E.g. a parameter defined as only having positive values and indifferent to a change in units.

DEFINING PROBABILITIES: JEFFREYS' PRIOR

A Jeffreys' prior represents complete ignorance about the value of a scale parameter (we used this for the prior in σ in the spectral estimation case).

It is equivalent to a uniform pdf for the logarithm of s

$$p(\log s|I)\mathrm{d}s = \text{constant.}$$

If an upper and lower range on s were known then we have

$$p(s|I) = \frac{1}{s\ln\left(s_{\mathrm{max}}/s_{\mathrm{min}}\right)}$$

DEFINING PROBABILITIES: JEFFREYS' PRIOR

This form of the Jeffreys' prior, $p(L|I) \propto 1/L$, is just the special case of a more general result. If we had a likelihood with parameters $\vec{\theta}$ then the **Jeffreys' prior** is a non-informative prior defined as

$$p(\vec{\theta}) \propto [|I(\vec{\theta}|]^{1/2}.$$

In this $I(\vec{\theta})$ is the **Fisher Information** defined as

$$I(\vec{\theta})_{i,j} = E\left[\frac{\partial \ln L(\vec{\theta})}{\partial \theta_i} \frac{\partial \ln L(\vec{\theta})}{\partial \theta_j}\right]$$

This prior is **invariant** under **any** re-parameterisation of $\vec{\theta}$ (i.e. scalings or translations).

DEFINING PROBABILITIES: MAXIMUM ENTROPY

How do we deal with situation where we know some constraints?

E.g. suppose a six-sided die was rolled many times and the average results was 4.5 then what probability should be assign the various outcomes $p(X_i|I)$? This constraint information

$$\sum_{i=1}^{6} ip(X_i|I) = 4.5$$

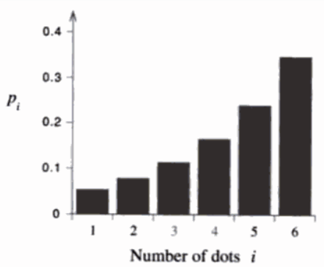
is **testable information**. A way to assign $p(X_i|I)$ is to **maximise the entropy**

$$S = -\sum_{i=1}^{6} p(X_i|I) \ln p(X_i|I)$$

given the above constraint and the condition $\sum_i p(X_i|I) = 1$.

DEFINING PROBABILITIES: MAXIMUM ENTROPY

$p(X_i|I)$ can be solved for in $\frac{\mathrm{d}S}{\mathrm{d}X} = 0$, using e.g. *Lagrange Multipliers*, and yields:



The MaxEnt approach can be justified using a varirty of approaches, e.g.

- Independence arguments (see e.g. *the kangeroo problem*[6])

[6] http://cmm.cit.nih.gov/maxent/kangaroo.html

DEFINING PROBABILITIES: COMMON PDFS

MaxEnt can be used to yield many common pdfs.

If we know the expected value, μ, of a continuous physical quantity, then using MaxEnt we find

$$\underbrace{p(x|\mu, I) = \frac{1}{\mu}\exp\left(-\frac{x}{\mu}\right)}_{\text{Exponential distribution}}$$

or, for a *discrete* physical quantity N we find

$$\underbrace{p(N|\mu, I) = -\frac{\mu^N e^{-\mu}}{N!}}_{\text{Poisson distribution}}$$

DEFINING PROBABILITIES: COMMON PDFS

If we know the expected value, μ, and variance, σ^2, of a continuous physical quantity, then MaxEnt shows

$$\underbrace{p(x|\mu, \sigma, I) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}_{\text{Normal distribution}}$$

This justifies the relevance of common pdfs. E.g. if we know μ and σ^2 then our least informative (so generally most conservative) pdf is the normal, which is often why it is so ubiquitous as a likelihood function.

However, if we have other information then we could improve our posteriors by using that information to assign a more relevant pdf.

BAYESIAN HYPOTHESIS TESTING

Unlike frequentist methods that require assessing some statistic to compare two hypotheses in Bayesian hypothesis testing (or **model selection**) we just directly calculate the probability of each hypothesis and compare them. E.g., if we have two hypotheses H_1 and H_2 (these could be any propositions), then

$$\text{if } \frac{p(H_1|I)}{p(H_2|I)} > 1 \text{ then } H_1 \text{ is favoured,}$$
$$\text{if } \frac{p(H_1|I)}{p(H_2|I)} < 1 \text{ then } H_2 \text{ is favoured,}$$

This is a **Bayesian odds ratio**, but how do we compute it?

BAYESIAN HYPOTHESIS TESTING

Suppose our two hypothesis represent two models defined by parameters $\vec{\theta}_1$ and $\vec{\theta}_2$, and we want to test which model if favoured given a set of data d. We can just apply Bayes' theorem

$$\mathcal{O}_{12} = \frac{p(H_1|d,I)}{p(H_2|d,I)} = \frac{\left(\frac{p(d|H_1,I)p(H_1|I)}{\cancel{p(d|I)}}\right)}{\left(\frac{p(d|H_2,I)p(H_2|I)}{\cancel{p(d|I)}}\right)} = \underbrace{\frac{p(d|H_1,I)}{p(d|H_2,I)}}_{\text{Bayes factor}} \underbrace{\frac{p(H_1|I)}{p(H_2|I)}}_{\text{Prior odds}}$$

- **Prior odds** - our prior knowledge about each model, often set to 1 (i.e. no preference for either model)
- **Bayes factor** - ratio of **evidences**, or **marginal likelihoods**, for each model.

BAYESIAN HYPOTHESIS TESTING

Note $p(H|d,I)$ is a *probability density function* not a probability, so we can't interpret particular values of it (e.g. $p(H_1|d,I)$) on their own.

Unless we can define the entire hypothesis space, so probabilities are given by e.g.

$$\text{prob}(H_1 \leq H \leq H_2|d,I) = \int_{H_1}^{H_2} p(H|d,I)\mathrm{d}H$$

we can only compare $p(H|d,I)$ for distinct different hypotheses

$$\frac{p(H_1|d,I)}{p(H_2|d,I)}.$$

But, in general, this is all we want to do.

BAYESIAN HYPOTHESIS TESTING

How do we calculate the **evidence**? For model 1 we have a posterior on $\vec{\theta}_1$ defined as

$$p(\vec{\theta}_1|d, H_1, I) = \frac{p(\vec{\theta}_1|d, H_1, I)p(\vec{\theta}_1|H_1, I)}{\underbrace{p(d|H_1, I)}_{\text{evidence}}}$$

The *evidence*, Z, is the normalisation constant that makes $\int p(\vec{\theta}_1|d, H_1, I)\mathrm{d}\vec{\theta}_1 = 1$, so

$$Z_1 = p(d|H_1, I) = \int_{\vec{\theta}_1} p(\vec{\theta}_1|d, H_1, I)p(\vec{\theta}_1|H_1, I)\mathrm{d}\vec{\theta}_1$$

BAYESIAN HYPOTHESIS TESTING

If this integral is analytic then the *evidence* is easy to calculate, or *if* $\vec{\theta}$ contains only a few (say less than 6) parameters, then it is relatively easy to calculate numerically over a parameter space grid. But, as with parameter estimation, over high dimensional models it becomes a computational challenge to compute.

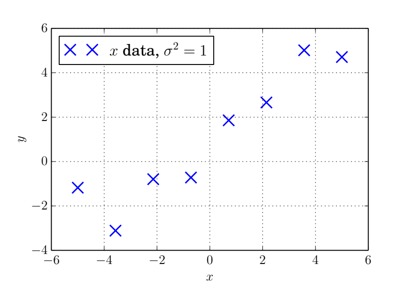
We may need to use integral approximations, or e.g. *nested sampling* [3].

BAYESIAN HYPOTHESIS TESTING: EXAMPLE

Let's take two hypotheses:

- H_1 - the data contains a line defined by an unknown gradient m and y-intercept c
- H_2 - the data is consistent with Gaussian random noise with zero mean and unit variance

We will assume that the variance, $\sigma^2 = 1$, of the noise in the data is known. So, we need to find the evidence for each hypothesis.



BAYESIAN HYPOTHESIS TESTING: EXAMPLE

We first need to assign priors on m and c. Just looking at the data we can see that reasonable ranges are:

- $0 \leq m \leq 5$ (i.e. the slope isn't negative, or very steep)
- $-2 \leq c \leq 4$

so,

$$p(m|H_1, I) = \begin{cases} 1/(5-0) = 0.2, \text{ if } 0 \leq m \leq 5 \\ 0, \text{ otherwise} \end{cases} \quad \text{and}$$

$$p(c|H_1, I) = \begin{cases} 1/(4 - -2) = 0.167, \text{ if } -2 \leq m \leq 4 \\ 0, \text{ otherwise} \end{cases}$$

Later we will see how changes in these prior ranges effect things.

BAYESIAN HYPOTHESIS TESTING: EXAMPLE

So, we now need to compute

$$
\begin{aligned}
Z_1 &= \int_{m=0}^{5}\int_{c=-2}^{4} p(m|H_1, I)p(c|H_1, I)p(d|m, c, H_1, I)\mathrm{d}m\mathrm{d}c,\\
&= \frac{0.033}{(2\pi\sigma^2)^{n/2}}\int_{m=0}^{5}\int_{c=-2}^{4}\exp\left(-\sum_{i=1}^{n}\frac{(d_i-(mx_i+c))^2}{2\sigma^2}\right)\mathrm{d}m\mathrm{d}c
\end{aligned}
$$

This integral is analytical (or at least involves erf) so can be computed easily. For our given data, with $n = 8$ and $\sigma^2 = 1$, we find (working in natural logarithms)

$$\ln Z_1 = -16.0.$$

BAYESIAN HYPOTHESIS TESTING: EXAMPLE

We now need the evidence that the data just consists of Gaussian noise with zero mean and variance $\sigma^2 = 1$, so we have

$$\begin{aligned} Z_2 &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\sum_{i=1}^{n} \frac{d_i^2}{2\sigma^2}\right), \\ \ln Z_2 &= -42.4. \end{aligned}$$

So, in this case the odds ratio for the two models is:

$$\mathcal{O}_{12} = \exp\left(\ln Z_1 - \ln Z_2\right) = \exp\left(-16.0 + 42.4\right) = 2.8 \times 10^{11}!$$

The data is definitely better explained by a line than Gaussian noise.

BAYESIAN HYPOTHESIS TESTING: EXAMPLE

What happens in the above case if we say that we already know that $c = 1$ (call this hypothesis H_3)?

We can again calculate the evidence in this case as

$$\begin{aligned} Z_3 &= \int_{m=0}^{5} p(m|H_3, I) p(d|m, c, H_3, I) \mathrm{d}m, \\ &= \frac{0.2}{(2\pi\sigma^2)^{n/2}} \int_{m=0}^{5} \exp\left(-\sum_{i=1}^{n} \frac{(d_i - (mx_i + 1))^2}{2\sigma^2} \right) \mathrm{d}m \\ \ln Z_3 &= -14.1. \end{aligned}$$

We see H_3 is favoured over H_1 by a factor of $e^{(-14.1+16.0)} = 6.7$. But, H_3 is just a subhypothesis included within H_1, so why is it favoured?

BAYESIAN HYPOTHESIS TESTING: OCCAM FACTOR

This is an example of the **‘Occam factor’** that is automatically incorporated in Bayesian model selection.

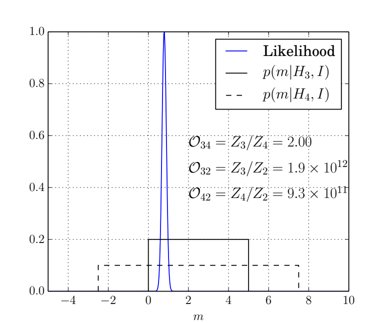
Occam’s razor: “**Plurality should not be posited without necessity**”

This principle means that more complicated models (i.e. with more parameters or larger parameter ranges) should be penalised *if* they do not contribute enough additional evidence.

BAYESIAN HYPOTHESIS TESTING: OCCAM FACTOR

We can also see Occam's razor in action if we again take hypothesis H_3, but also have H_4 with double the prior range to $p(-2.5 \leq m \leq 7.5|H_4, I) = 0.1$. Here H_3 is twice as probable as H_4: the likelihood is entirely within $p(m|H_3, I)$, so expanding the range adds no extra information.

If the prior were hugely expanded the model containing a line would still highly favoured over Gaussian noise.



BAYESIAN HYPOTHESIS TESTING

We also can make statements about compound hypotheses e.g.

$$\mathcal{O}_{H_1 \text{ and } H_2 \text{ vs. } H_3} = \frac{p(H_1|d,I)p(H_2|d,I)}{p(H_3|d,I)}$$

or,

$$\mathcal{O}_{H_1 \text{ or } H_2 \text{ vs. } H_3} = \frac{p(H_1|d,I) + p(H_2|d,I)}{p(H_3|d,I)}.$$

Or, we could compare multiple hypotheses, such as data being described by a polynomial with different numbers of coefficients, N, to see which provides the best fit.

BAYESIAN HYPOTHESIS TESTING: APPROXIMATIONS

Calculating the evidence can be computationally costly, so an approximate way to calculate the Bayes factor is the **Bayesian Information Criterion** (BIC) [4]

$$\text{BIC} = 2\ln L_{\max} + k\ln n$$

where $L_{\max}$ is the maximum likelihood, k is the number of parameters in model, and n is the number of data points.

We can (roughly) say:

- If $\text{BIC}_1 - \text{BIC}_2 > 2$ there is positive evidence for model 1
- If $\text{BIC}_1 - \text{BIC}_2 > 6$ there is *strong* evidence for model 1

A better approximation to the Bayes factor is the **Laplace approximation**, which assumes a multivariate Gaussian posterior (see e.g. [5]).

PARAMETER ESTIMATION: THE GAUSSIAN APPROXIMATION

Posteriors can sometimes be approximated by a Gaussian distribution. This can be justified through a maximum likelihood type approach (which in turn provides a Bayesian foundation to frequentist maximum likelihood estimators).

The maximum posterior is found when

$$\left.\frac{\partial p(\theta|d,I)}{\partial\theta}\right|_{\theta=\theta_0} = 0$$

Equivalently we can define $\ell = \ln p(\theta|d,I)$ and compute $\left.\frac{\partial\ell}{\partial\theta}\right|_{\theta=\theta_0} = 0$. If we Taylor expand $\ell(\theta)$ around $\theta = \theta_0$ then

$$\ell(\theta) = \ell(\theta_0) + \underbrace{\cancel{\left.\frac{\partial\ell}{\partial\theta}\right|_{\theta=\theta_0}}}_{=0}(\theta-\theta_0) + \frac{1}{2}\left.\frac{\partial^2\ell}{\partial\theta^2}\right|_{\theta=\theta_0}(\theta-\theta_0)^2 + \ldots$$

PARAMETER ESTIMATION: THE GAUSSIAN APPROXIMATION

So, the posterior is given by

$$p(\theta|d,I) = \exp\left[\ell(\theta)\right]$$

and neglecting higher order terms in $\ell(\theta)$ (the **Gaussian approximation**) gives

$$p(\theta|d,I) \propto \exp\left(-\frac{A}{2}(\theta-\theta_0)^2\right),$$

where

$$A = -\left.\frac{\partial^2\ell}{\partial\theta^2}\right|_{\theta=\theta_0}$$

This is equivalent to a **normal** distribution with $\sigma^{-2} = A$.

In these cases we can summarise inference from the posterior with: $\theta = \theta_0 \pm \sigma$.

THE GAUSSIAN APPROXIMATION (2D)

For two parameters we have a posterior

$$p(\theta_1, \theta_2 | d, I) \propto p(d | \theta_1, \theta_2, I) \times p(\theta_1, \theta_2 | I)$$

and the ‘best’ estimator at the posterior maximum is

$$\left. \frac{\partial p(\theta_1, \theta_2 | d, I)}{\partial \theta_j} \right|_{\theta_j = \theta_{0,j}} = 0$$

Compute $\left. \frac{\partial \ell}{\partial \theta_j} \right|_{\theta_j = \theta_{0,j}}$ where $\ell = \ln p(\theta_1, \theta_2 | d, I)$.

THE GAUSSIAN APPROXIMATION (2D)

We again Taylor expand $\ell(\theta_1, \theta_2)$ about $\theta_{0,j}$ giving

$$\begin{aligned}
&\ell(\theta_1, \theta_2) = \ell(\theta_{0,1}, \theta_{0,2}) + \cancel{\tfrac{\partial \ell}{\partial \theta_1}\Big|_{\theta_1 = \theta_{0,1}}}(\theta_1 - \theta_{0,1}) + \\
&\cancel{\tfrac{\partial \ell}{\partial \theta_2}\Big|_{\theta_2 = \theta_{0,2}}}(\theta_2 - \theta_{0,2}) + \tfrac{1}{2}\Bigg[\tfrac{\partial^2 \ell}{\partial \theta_1^2}\Big|_{\theta_1 = \theta_{0,1}}(\theta_1 - \theta_{0,1})^2 + \\
&\tfrac{\partial^2 \ell}{\partial \theta_2^2}\Big|_{\theta_2 = \theta_{0,2}}(\theta_2 - \theta_{0,2})^2 + 2\tfrac{\partial^2 \ell}{\partial \theta_1 \partial \theta_2}\Big|_{\theta_j = \theta_{0,j}}(\theta_1 - \theta_{0,1})(\theta_2 - \theta_{0,2})\Bigg]
\end{aligned}$$

so, the Gaussian approximation, is

$$p(\theta_1, \theta_2 | d, I) \propto \exp\left[\ell(\theta_1, \theta_2)\right] \propto \exp\left[-\frac{1}{2}Q\right]$$

THE GAUSSIAN APPROXIMATION (2D)

We have put this in matrix form to have the quadratic form

$$Q = (\theta_1 - \theta_{0,1}, \theta_2 - \theta_{0,2}) \left[\begin{array}{cc} A & C \\ C & B \end{array} \right] \left(\begin{array}{c} \theta_1 - \theta_{0,1} \\ \theta_2 - \theta_{0,2} \end{array} \right),$$

where

$$A = \frac{\partial^2 \ell}{\partial \theta_1^2}\bigg|_{\theta_1 = \theta_{0,1}}, B = \frac{\partial^2 \ell}{\partial \theta_2^2}\bigg|_{\theta_2 = \theta_{0,2}}, \text{ and } C = \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_2}\bigg|_{\theta_j = \theta_{0,j}}$$

This is the **bivariate normal distribution** with covariance matrix defined using C.

This can be generalised to any number of parameters

$$\sigma_{i,j}^2 = \langle (\theta_i - \theta_{0,i})(\theta_j - \theta_{0,j}) \rangle = \left[-\frac{\partial^2 \ell}{\underbrace{\partial \theta_i \partial \theta_j}_{\text{Fisher information matrix}}} \right]^{-1}.$$

The **Fisher Information Matrix** (FIM)

$$\mathbf{F} \equiv F_{i,j} = \frac{\partial^2 \ell}{\partial\theta_i \partial\theta_j}$$

provides a measure of how much **information** a given dataset can yield about the parameters of a model.

It tells us which combinations of parameters can be well constrained by the data. E.g. if $F_{i\neq j} = 0$ (the matrix is diagonal) then if the $i^{\rm th}$ element of the FIM is a large negative number then the **variance** of the parameter θ_i is small (i.e. it is well constrained).

THE GAUSSIAN APPROXIMATION

So if, for our model:

- the likelihood is Gaussian in shape, or we can approximate it as Gaussian (i.e. if the higher order terms in the Taylor expansion of ℓ can be neglected),
- and, the parameters have broad, uniform priors,

then the posterior will also be Gaussian.

If we can evaluate the first and second derivatives of ℓ (*and* find the maximum of the posterior) we can:

- compute the Fisher Information Matrix and covariance matrix.

THE GAUSSIAN APPROXIMATION

Is the Gaussian approximation a good idea?

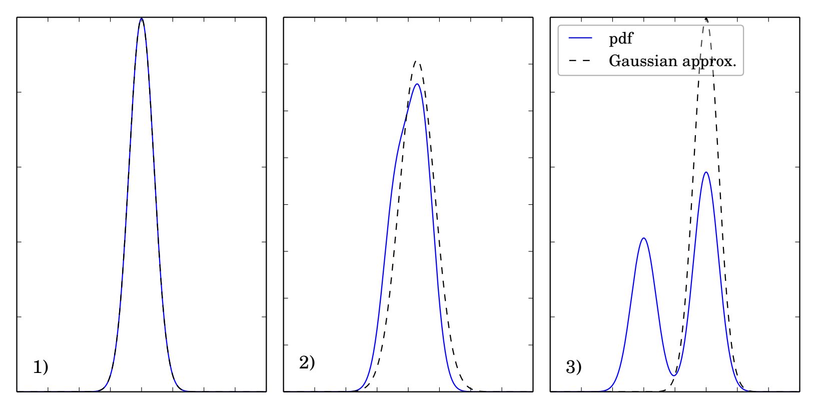
- It greatly simplifies calculations as we only need to calculate elements of the Fisher Information Matrix
- However, with modern computers and algorithms it is now often fairly simple to just calculate the full posterior pdf

It is often a good approximation in case of high signal-to-noise ratio (i.e. parameters are well constrained), but can be poor at low signal-to-noise ratio. It should not be used if the posterior is multi-modal.

THE GAUSSIAN APPROXIMATION

How good is the Gaussian approximation?

Case 1) very good, Case 2) OK, Case 3) very poor.



BIBLIOGRAPHY

[1] D. S. Sivia. *Data analysis: A Bayian Tutorial*. Oxford University Press, 2006.

[2] G. L. Bretthorst. *Bayesian Spectrum Analysis and Parameter Estimation*. Springer-Verlag, 1988.

[3] J. Skilling. Nested Sampling for General Bayesian Computation. *Bayesian Analysis*, 1(4):833–860, 2006.

[4] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461, 1978.

[5] R. Trotta. Applications of Bayesian model selection to cosmological parameters. *MNRAS*, 378:72–82, June 2007.