Bayesian Parameter Estimation

In this section we will see a number of examples of the application of Bayes' theorem to simple parameter estimation problems where the model contains just a single parameter. The computational tools needed for applying the same methodology to more interesting (and realistic) multi-parameter problems will be described later in the course. This section will also highlight the importance of the prior in Bayesian analysis.

In order to perform a Bayesian analysis we first have to choose a prior for the model parameters. How should the prior be chosen? Usually, the model parameters are continuous, so the prior is a probability density function. So the question becomes how should we choose this PDF? There is no unique answer. However, as a general principle, the prior should be chosen to describe our state of knowledge about the model parameters *before* performing the experiment. The following subsections contain several examples of how this rather vague statement can be applied to the example problem of estimating the number desnity of stars.

0.0.1 Ignorance Priors (uniform)

We often pretend that we don't know much (or anything) about the value of model parameters, except perhaps that is must lie in a certain range of values. (Is this every really true?) But if we really don't have any information to guide us then we should try to choose a prior that reflects our ignorance.

For some types of parameters, a uniform distribution is the obvious way to try and achieve this. For location parameters such as the (x, y) coordinates of a point on an image we would naturally want our prior to be invariant under a translation of the origin by a constant Δx;

$$\pi(x)\mathrm{d}x = \pi(x + \Delta x)\mathrm{d}(x + \Delta x) \quad \Rightarrow \quad \pi(x) \propto \text{const.} \tag{1}$$

The uniform distribution (within certain limits) is the only distribution that has this property.

Notation: it will sometimes be convenient to use the following notation for the *indicator function* for distributions with compact support;

$$\mathbb{1}_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{else} \end{cases} \quad . \tag{2}$$

So in 1D, $x \in \mathbb{R}$, we can write the Heaviside step function as $\Theta(x) = \mathbb{1}_{[0,\infty)}(x)$.

Strictly speaking, the prior is a probability distribution and must therefore be properly normalised (or at least normalisable; meaning its integral is finite). When using a uniform distribution as a prior this means we must specify lower/upper limits for distribution corresponding to the minimum/maximum allowed values of the parameter. However, in practice these are often omitted. Omitting the lower/upper limits means the prior is unnormalisable; a prior that is not normalisable is referred to as an *improper prior*.

Box 0.1: Number density of stars. . . uniform prior

Here, we choose to use a uniform prior on the parameter S. This requires that we choose a prior range; we choose $0 < S < S_{\max} = 20\,\mathrm{deg}^{-2}$. The justification for this choice of upper limit comes at the end of the analysis when we will find that the posterior has negligible support in the region $S \sim 20\,\mathrm{deg}^{-2}$ (see plot of the posterior Fig. 1).

$$\pi(S) = \frac{\mathbb{1}_{(0,S_{\max})}(S)}{S_{\max}} \tag{i}$$

Using the Poisson likelihood, Bayes' theorem gives the posterior as

$$P(S|n) = \frac{(AS)^n \exp(-AS)\mathbb{1}_{(0,S_{\max})}(S)}{S_{\max} n! Z}, \tag{ii}$$

where the evidence (integral evaluated numerically) is

$$Z = \int_0^{S_{\max}} \mathrm{d}S\, \frac{(AS)^n \exp(-AS)}{S_{\max} n!} \approx 0.050. \tag{iii}$$

0.0.2 Ignorance Priors (log uniform)

A uniform distribution is not the right choice for all parameters. For *scale parameters* associated with the magnitude a particular quantity (such as length, mass, lifetime, etc) we would naturally want our prior to be invariant under a change of the units (by a constant factor α) used to measure this quantity;

$$\pi(x)\mathrm{d}x = \pi(\alpha x)\mathrm{d}(\alpha x) \quad \Rightarrow \quad \pi(x) \propto \frac{1}{x}. \tag{3}$$

This is equivalent to placing a uniform prior on the logarithm of the scale parameter; $\pi(\log x) \propto \text{const}$. This can be easily checked by a change of variables.

The log-uniform distribution (again, within certain limits) is the only distribution that has the required invariance property.

This is sometimes called *Jeffreys prior*.

Box 0.2: Number density of stars. . . log-uniform prior

Here, we choose to use a log-uniform prior on the density parameter, $\pi(S) \propto S^{-1}$. This requires that we choose a non-zero value for the lower limit of the prior range; we choose $S_{\text{min}} = 1\,\text{deg}^{-2}$. (Actually, this choice turns out to be a little bit too large; see posterior in Fig. 1).

$$\pi(S) = \frac{\mathbb{1}_{(S_{\text{min}},S_{\text{max}})}(S)}{S\log(S_{\text{max}}/S_{\text{min}})} \tag{i}$$

Using the Poisson likelihood, Bayes' theorem gives the posterior as

$$P(S|n) = \frac{A(AS)^{n-1}\exp(-AS)\mathbb{1}_{(S_{\text{min}},S_{\text{max}})}(S)}{n!\log\left(\frac{S_{\text{max}}}{S_{\text{min}}}\right)Z}, \tag{ii}$$

where the evidence (integral evaluated numerically) is

$$Z = \int_{S_{\text{min}}}^{S_{\text{max}}} \mathrm{d}S\;\frac{A(AS)^{n-1}\exp(-AS)}{n!\log\left(\frac{S_{\text{max}}}{S_{\text{min}}}\right)} \approx 0.067. \tag{iii}$$

0.0.3 Conjugate Priors

In some special cases it is possible to solve to whole inference problem (including finding the evidence) analytically.

Suppose we have some simple analytic expression for the likelihood in a particular problem;

$$\text{data}|\text{parameters} \sim \mathcal{L}. \tag{4}$$

It is sometimes possible to find a known family of probability distributions (perhaps parameterised by one or more free parameters α) specifically taylored to this likelihood which have the property that if we use it as a prior when applying Bayes' theorem the posterior is is the same family of distributions (usually with some different parameters α').

$$\begin{aligned} &\text{parameters} \sim \text{NiceDistribution}(\alpha) && \text{(Prior)} && (5)\\ &\text{parameters}|\text{data} \sim \text{SameNiceDistribution}(\alpha') && \text{(Posterior)} && (6) \end{aligned}$$

The role of the likelihood is to update our state of knowledge in going from the prior to the posterior. In the special case of a conjugate prior this corresponds to updating the parameters of our distribution according to $\alpha \to \alpha'$.

Box 0.3: Number density of stars. . . gamma prior

In this problem our likelihood is the Possion distribution. The gamma distribution is the conjugate prior to the Poisson likelihood in the following sense;

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

We will now prove this relationship using Bayes' theorem.

We will use gamma distribution (with *shape parameter* $k=2$ and *scale parameter*

$\theta = 4\,\mathrm{deg}^{-2}$) as a prior on S. We can denote this is in a couple of different ways:

$$S \sim \mathrm{Gamma}(k, \theta), \tag{iv}$$

$$\pi(S) = \begin{cases} \frac{S^{k-1}\exp(-S/\theta)}{\Gamma(k)\theta^k} & \text{if } S > 0 \\ 0 & \text{else} \end{cases}. \tag{v}$$

Using the Poisson likelihood, Bayes' theorem gives the posterior as

$$P(S|n) = \frac{A^n S^{k+n-1}\exp(-[A\theta+1]S/\theta)}{\Gamma(k)\theta^k n! Z} \quad \text{for } S > 0. \tag{vi}$$

However, we recognise the S dependence Eq.vi as that another gamma distribution for the parameter S with updated shape and scale parameters $k' = k + n$ and $\theta' = \theta/(A\theta + 1)$ respectively. Therefore, we must have

$$P(S|n) = \frac{S^{k+n-1}\exp(-[A\theta+1]S/\theta)}{\Gamma(k+n)\left(\frac{\theta}{A\theta+1}\right)^{k+n}} \quad \text{for } S > 0, \tag{vii}$$

$$\text{Or } S|n \sim \mathrm{Gamma}(k'|\theta'). \tag{viii}$$

As a bonus the conjugate prior also gives us an exact expression for the evidence. We have two expressions for the posterior in eqs. vi and vii, comparing these gives

$$Z = \frac{\Gamma(k+n)(A\theta)^n}{\Gamma(k)(A\theta+1)^{k+n} n!} \approx 0.079. \tag{ix}$$

For most realistic problems it is not possible to find a conjugate prior and it is necessary to resort to the numerical techniques described in the next chapter.

A conjugate prior is mainly an algebraic convenience; it allows us to find a closed-form expression for the posterior (and the evidence). Additionally, because they lead to a such a simple rule for updating parameters, conjugate priors also help to build intuition for the iterative process of Bayesian inference where the likelihood acts to update our prior state of knowledge.

0.0.4 Different Priors, Different Posteriors

In the above examples, we have now obtained posteriors from three different priors for the number density of stars from our survey: in Boxes 0.1 to 0.3 we have obtained different expressions for the posterior $P(S|n)$ (and the evidence $Z = P(n)$) using a uniform, a log-uniform, and a gamma prior.

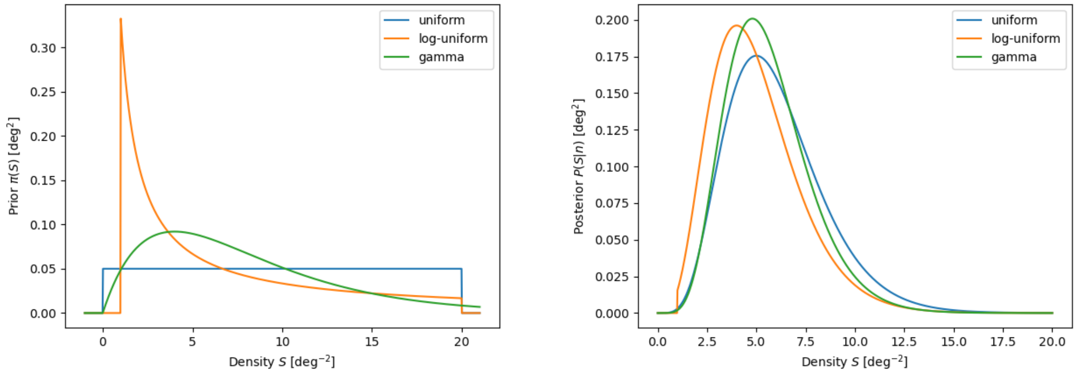


Figure 1: Comparison of the three different priors (left) and posteriors (right) on the density parameter S used in the Poisson process examples in the above Boxes.

Which is right? In fact, can't we make the posterior anything we like simply by a suitable choice of prior?

Firstly - It's not really a matter of right or wrong; these are three reasonable posteriors that follow from 3 reasonable choices of the prior. Secondly - Yes, but this misses the point; we are supposed to choose a reasonable prior ahead of time and then work out the consequences.

Figure 1 shows the 3 different posterior distributions we have obtained so far. It takes some getting used to that we can have such different answers to a seemingly simple question as "what is the density?" that are all correct. In practice however, this is usually not an issue for the following reason. If our data is good then our freedom in the choice of a (reasonable) prior becomes unimportant. For example, suppose that instead of measuring $n = 5$ stars in $A = 1$ square degree, we had instead surveyed a larger area and measured $n' = 50$ stars in $A' = 10$ square degrees. In this case the posteriors would look like Fig 2. The different priors have a much smaller effect in this case.

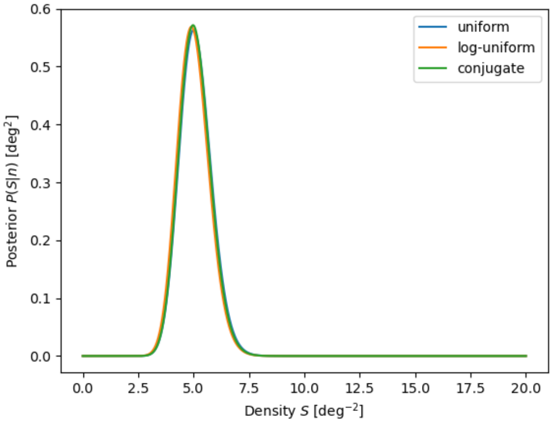


Figure 2: Comparison of the three different posteriors on the density parameter S for a survey with $n' = 50$ and $A' = 10\,\mathrm{deg}^2$. With this more informative likelihood function, the effect of the different priors is greatly reduced compared to that seen in Fig. 1.

The better our data the more constraining our likelihood and hence the smaller the effect on the posterior of a small change in the prior.

This argument can be turned around. If we ever find ourselves in a situation where making different (reasonable) choices of the prior has a significant effect on our conclusions then this suggests that likelihood doesn't contain much information about what we are trying to measure.

0.1 Iterative Bayesian Inference

Remember, the general principle is that for each experiment the prior should be chosen to describe our state of knowledge before performing the experiemnt. The likelihood (used in Bayes' theorem) then updates this knowledge and the posterior describes our new state of knowledge after performing the experiment.

If we now perform another experiment aimed at learning more about the same quantity, it is natural to use the posterior from the previous experiment as a prior. It's the same distribution, we're only changing what we call it; the person doing the inference for the first experiment called it the posterior, but the person doing the inference for the second experiment calls it the prior.

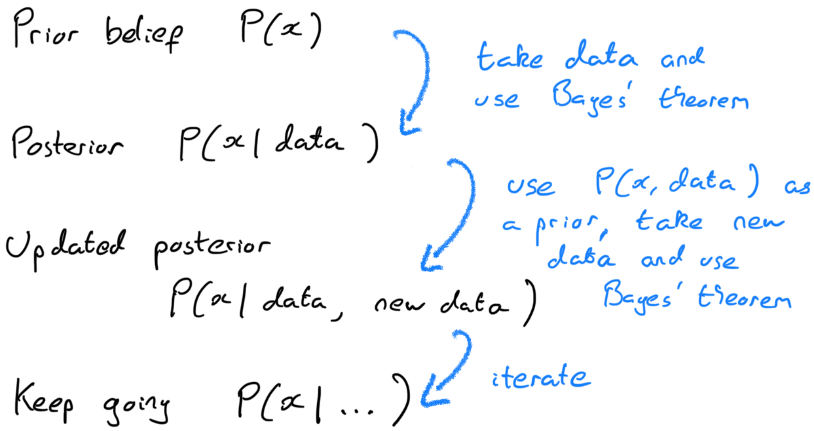


Figure 3: A metaphor for the scientific method?

Box 0.4: Number density of stars... multiple surveys

In our first survey we found $n_1 = 5$ stars in $A_1 = 1$ square degree. Using the gamma prior ($S \sim \text{Gamma}(k, \theta)$) we made a measurement of the density and obtained the posterior, $S|n_1 \sim \text{Gamma}(k + n_1, \theta/[A_1\theta + 1])$.

Suppose we now decide that this measurement isn't precise enough and we decide to perform a second survey to improve our measurement; we find $n_2 = 6$ stars in a (different) area $A_2 = 1$ square degree.

We can use the posterior distribution from the first survey as a prior for the second survey. After all, this is the distribution that properly reflects our state of knowledge about S *before* the second survey.

$$S|n_1 \sim \text{Gamma}\left(k + n_1, \frac{\theta}{A_1\theta + 1}\right) \qquad \text{(New Prior)} \qquad \text{(i)}$$

Our likelihood hasn't changed, it is still the Poisson distribution.

$$n_2|S \sim \text{Poisson}(A_2 S) \qquad \text{(Likelihood for Survey 2)} \qquad \text{(ii)}$$

We now use Bayes' theorem to find the new posterior. But we have already solved this problem in Box 0.3 just with some quantities relabelled.

$$S|n_1, n_2 \sim \text{Gamma}\left(k + n_1 + n_2, \frac{\theta}{(A_1+A_2)\theta + 1}\right) \qquad \text{(New Posterior)} \qquad \text{(iii)}$$

We are using the fact that the results of our two surveys are independent; i.e. that n_2 doesn't depend on n_1 in Eq. ii.

At each stage the likelihood acts to update our prior. This is reflected by updating the parameters of the gamma distribution according the rule found in Box 0.3:

$$(k, \theta) \to \left(k + n_1, \theta' = \frac{\theta}{A_1\theta+1}\right) \to \left(k + n_1 + n_2, \theta'' = \frac{\theta'}{A_1\theta'+1} = \frac{\theta}{(A_1 + A_2)\theta+1}\right).$$

Reassuringly, the new posterior in Eq. iii is the same one that we would have obtained if we had performed the Bayesian inference in a single step for a single survey of both areas ($A = A_1 + A_2$) that found the same $n = n_1 + n_2$ stars.

Box 0.5: Number density of stars. . . selection effects (question)

Consider again our survey of an area A_1 that found n_1 stars above a certain threshold magnitude.

Suppose we now learn of a second survey of a larger area $A_2 > A_1$ of sky that found n_2 stars. However, the second survey uses a less sensitive instrument. A star that would have been found in the first survey is only found by the second survey with probability p.

If our goal is to learn as much as we can about S, which survey is better?

One way of answering the question posed in Box. 0.5 would be to perform a Bayesian inference for both surveys (separately) and compare the widths of the posteriors on S (e.g. quantified using the standard deviation). The survey with the narrower posterior is best. However, realistic Bayesian inferences are often expensive and time consuming. It would be nice to be able to answer questions of this type without having to explicitly obtain the posteriors.

The Fisher information is measurement of the amount of information that some data carries about an unknown parameter. It can be used for answering questions of the type posed in Box. 0.5. This is demonstrated in Box. 0.6

Box 0.6: Number density of stars. . . selection effects (answer)

Continuing from Box. 0.5.

For the first survey, the likelihood is a Poisson distribution with rate $A_1 S$;

$$n_1|S \sim \text{Poisson}(A_1 S) \quad \Rightarrow \quad P(n_1|S) = \frac{(A_1 S)^{n_1} \exp(-A_1 S)}{n_1!} \tag{i}$$

For this likelihood, the score is $\mathcal{S}_1(S) = (n_1/S) - A_1$. The definition of the Fisher information (either version) involves an integral with respect to a continuous data variable, x. In this problem the data is the discrete number of stars, n_1. Therefore, the Fisher information will involve a sum over n_1. We can compute the Fisher information from either of the definitions given, Working from the definition involving the score, we have

$$\begin{aligned}
\mathcal{I}_1(S) &= \sum_{n_1=0}^{\infty} \mathcal{L}(n_1|S) \left(\frac{n_1}{S} - A_1 \right)^2 , && \text{(ii)} \\
&= \sum_{n_1=0}^{\infty} \mathcal{L}(n_1|S) \left(\frac{n_1^2}{S^2} - \frac{2n_1 A_1}{S} + A_1^2 \right) , && \text{(iii)} \\
&= \left(\left[\frac{A_1}{S} + A_1^2 \right] - 2A_1^2 + A_1^2 \right) , && \text{(iv)} \\
&= \frac{A_1}{S} . && \text{(v)}
\end{aligned}$$

In going from the second to the third line we have used the fact that both the mean and the variance of the Poisson distribution are equal to the rate parameter.

Hopefully, it makes sense that the Fisher information is proportional to A_1; the bigger the area we survey the more information we get.

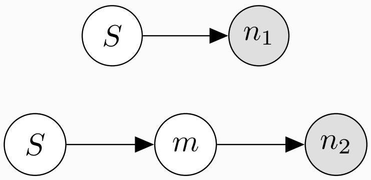
For the second survey, the likelihood is more complicated. The actual number of stars m in the surveyed area is a Poisson random variable with rate parameter $A_2 S$. But the number of stars we actually detect is a Binomial random variable with m trials and probability p of success for each trial.

$$Degreem|S \sim \text{Poisson}(mS) \tag{vi}$$

$$n_2|m \sim \text{Binomial}(m, p) \tag{vii}$$

The variable m is an example of what is sometime called a *latent variable*.

Probabilistic Graphical Models (PGMs) can be a helpful way of visualising the conditional dependencies of observations (i.e. data) and model variables in more complicated models. PGMs will be discussed in more detail when we discuss hierarchical bayesian models later in the course. However, here are the PGMs for the first and second surveys.



Using the law of total probability, the likelihood for the second survey is

$$\begin{aligned}
\mathcal{L}(n_2|S) &= \sum_{m=n_2}^{\infty} P(n_2|m)P(m|S), & \text{(viii)}\\
&= \sum_{m=n_2}^{\infty} {}^{m}C_{n_2} p^{n_2}(1-p)^{m-n_2} \cdot \frac{(A_2S)^m \exp(-A_2S)}{m!}, & \text{(ix)}\\
&= \frac{(A_2pS)^{n_2}\exp(-A_2S)}{n_2!}\sum_{\mu=0}^{\infty}\frac{\big((1-p)A_2S\big)^{\mu}}{\mu!}, & \text{(x)}\\
&= \frac{(A_2p_2S)^{n_2}\exp(-A_2pS)}{n_2!}. & \text{(xi)}
\end{aligned}$$

We recognise this as another Poisson distribution with rate parameter A_2pS. Therefore, reusing the results from above, the score for the second survey is $\mathcal{S}_2(S) = (n_2/S) - A_2$ and the Fisher information is

$$\mathcal{I}_1(S) = \frac{A_2p}{S}. \tag{xii}$$

The second survey contains more information than the first survey about the density parameter if $A_2p > A_1$.