# Bayesian Model Comparison

In previous chapters we encountered encountered *parameter estimation* problems where we had available to us a suitable model that described the data. There, the Bayesian inference problem to find the best possible values of the model parameters and their uncertainties.

But what if we don't know what model to use? Or, what if we have several competing models which all claim to describe the data, how do we pick the best model? These are questions of *model selection*. As we will see in this chapter, these questions can also be answered with the framework of Bayesian inference.

## 0.1 Why Not Use the Maximum Likelihood?

A reasonable first thought would be, "why not use the model that fits the data best".

Let's formalise this idea a bit. Suppose we have two models: model $A$ with some parameters $\lambda_A$, and model $B$ with parameters $\lambda_B$. Given some data $D$ we can right down the *likelihood* for both models: $\mathcal{L}(D|\lambda_A, A)$ and $\mathcal{L}(D|\lambda_B, B)$. We can numerically find the peak likelihood values of both of these distributions and consider the ratio

$$\text{Maximum Likelihood Ratio} = \frac{\max_{\lambda_A}\mathcal{L}(D|\lambda_A, A)}{\max_{\lambda_B}\mathcal{L}(D|\lambda_B, B)} = \frac{\mathcal{L}(D|\hat{\lambda}_A, A)}{\mathcal{L}(D|\hat{\lambda}_B, B)}\,. \tag{1}$$

Why is it not enough to select model $A$ is this ratio is larger than 1 and model $B$ otherwise?

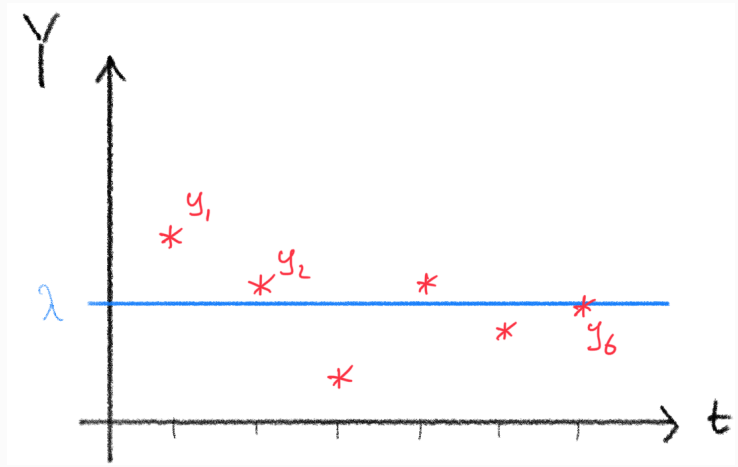
The problem with this approach is that it fails to account for the different complexity between models. Some models may have more or fewer parameters than others. Or parameters that can extend over different ranges. A model with a many parameters allowed to vary over wide ranges will generally fit the data well (i.e. it will have a high value of the

likelihood). However, this doesn't necessarily mean that this model should be chosen over a simpler model that does almost as good a job at fitting the data.

**Box 0.1: Fitting the data well isn't everything**

Suppose we have a sequence of measurements $\mathcal{D} = \{y(t_1), y(t_2), \ldots, y(t_N)\}$ of a quantity $Y(t)$ made over an interval if time; where $y$. For simplicity, assume that all measurements have identical, independent Gaussian errors $\sigma$. For example,

$$\mathcal{D} = \{1.5,\ 1.2,\ 0.4,\ 1.1,\ 0.8,\ 1.0\} \quad \text{with} \quad \sigma = 1\,. \tag{i}$$




We have two models:

- A - the quantity $Y = 0$.
- B - the quantity $Y = \lambda$, where $\lambda$ is a constant, unknown parameter.

For model B, choosing $\lambda = \text{average}(y_i)$ will fit the data best.

There could be more models: e.g. model C, with two parameters ($Y = \lambda_1 + \lambda_2 x$), model D with three ($Y = \lambda_1 + \lambda_2 x + \lambda_3 x^2$), and so on.

It is clear that in (almost[a]) any circumstances $B$ will fit the data better than A (i.e. model B will have a higher peak likelihood value). Model B cannot possibly fit the data any worse than model A because A is nested inside B and can be recovered by setting $\lambda = 0$.

Similarly, model C and will almost surely fit the data better than B, and D will fit better than C, and so on indefinitely.

Does this necessarily mean the data favours model Z? No, we have failed to account properly for the different complexity of the models.

[a]The only exception being when the average of the measurements is exactly zero, in which case the two models will fit equally well

A model that fits the data well using a small number of free parameters is sometimes described as being *parsimonious*.

## 0.2 The Odds Ratio

Considering the material covered in previous chapters, it is perhaps no surprise that instead of the likelihood we are going to consider the *posterior odds ratio*,

$$\begin{aligned} \text{Posterior Odds Ratio} \equiv \mathcal{O}_{A,B} &= \frac{P(A|D)}{P(B|D)}\,, \\ &= \frac{P(D|A)}{P(D|B)} \times \frac{P(A)}{P(B)}\,. \end{aligned} \tag{2}$$

In going from the first to the second line we have used Bayes' theorem twice, once in the numerator and once in the denominator, and the common factor of $P(D)$ has cancelled between the two.

The second term in equation 2 is called the *prior odds ratio*. As is usual in a Bayesian analysis, our final answer depends partly on what we believed before performing the experiment. In some (but not all) situations we might want to try and be fair to both models and we might take the prior odds ratio to be unity.

Notice that in our first attempt in equation 1 we evaluated the ratio of the likelihoods at the maximum points. Here, in equation 2, we do *not* wish to do this. Instead of maximising over the free parameters we will marginalise over them, thereby allowing them to take on all possible values.

$$\mathcal{O}_{A,B} = \frac{\int \mathrm{d}\lambda_A\; P(D,\lambda_A|A)}{\int \mathrm{d}\lambda_B\; P(D,\lambda_B|B)} \times \frac{P(A)}{P(B)}\,. \tag{3}$$

We will now use the *product rule* of probability to rewrite the integrands in this ratio in terms of more familiar probability distributions;

$$\mathcal{O}_{A,B} = \frac{\int \mathrm{d}\lambda_A \; P(D|\lambda_A, A) P(\lambda_A|A)}{\int \mathrm{d}\lambda_B \; P(D|\lambda_B, B) P(\lambda_B|B)} \times \frac{P(A)}{P(B)} \,. \tag{4}$$

Here we see again the likelihoods for the two models reappearing. We have also been forced to introduce *priors* for the free parameters in both models: $P(\lambda_A|A)$ and $P(\lambda_B|B)$.

But now we recognise the terms inside the integrals in both the numerator and denominator on equation 4, they are the likelihood times the prior for each model. Rewriting Eq. 4 in our earlier notation, we have

$$\mathcal{O}_{A,B} = \frac{\int \mathrm{d}\lambda_A \; \mathcal{L}(D|\lambda_A, A) \pi(\lambda_A|A)}{\int \mathrm{d}\lambda_B \; \mathcal{L}(D|\lambda_B, B) \pi(\lambda_B|B)} \times \frac{P(A)}{P(B)} \,. \tag{5}$$

But the integral of the likelihood times the prior over the full parameter space of the model is, by definition, just the *Bayesian evidence* for that model model. Using our earlier notation for the evidence ($P(D|A) = Z_A$, and $P(D|B) = Z_B$) we can write the posterior odds as

$$\mathcal{O}_{A,B} = \frac{Z_A}{Z_B} \times \frac{P(A)}{P(B)} \,. \tag{6}$$

The take home message is that when comparing two models for the same data, *compute the ratio of the model evidences*. This ratio then acts to update our relative state of belief in the two models. Before performing the experiment we have the prior odds ratio. After performing the experiment we multiply this by the ratio of evidences to get the posterior odds ratio.

You might worry about the presence of the prior odds ratio in the final answer for $\mathcal{O}_{A,B}$. It might feel unscientific in some way to allow our prior beliefs to affect the answer in this way; "would it not be better to let the data decide?" Unfortunately, this is misguided; it is important to realise that our prior beliefs are always present (and rightly so) when we interpret any data. (Although, in some situations it might be appropriate to set the prior odds ratio to unity.) In any case, in most situations the evidence ratio ends up dominating over the prior odds ratio. If one model fits the data significantly better than the other then it will have a much higher evidence and this will lead naturally lead us to favour that model regardless of the prior odds. The prior odds ratio is generally significant only in situations where both theories give similarly good fits to the data and/or the more complicated model gives a fit that is not sufficiently better to justify its extra complexity.

Of course, we could have chosen to define the odds ratio the other way around, with $B$ in the numerator and $A$ in the denominator. It makes no difference, we have

$$\mathcal{O}_{A,B} = \frac{1}{\mathcal{O}_{B,A}}\,. \tag{7}$$

**Box 0.2: An example using the odds ratio**

Consider again the data in example 0.1 and the models A and B.

Let us choose a prior odds ratio that treats both models as being equally likely;

$$\frac{P(A)}{P(B)} = 1\,. \tag{i}$$

First, consider the simpler model A. This model has no free parameters. We can write down the likelihood for this model as follows,

$$P(D|A) = \prod_{i=1}^{6} \frac{\exp\left(-\frac{1}{2}y_i^2\right)}{\sqrt{2\pi}} = 1.41 \times 10^{-4}\,. \tag{ii}$$

As there are no free parameters in this model there is no need to perform a marginalisation integral.

Second, consider the more complicated model B. This model has a single free parameter, $\lambda$. We can also write down the likelihood for this model as follows,

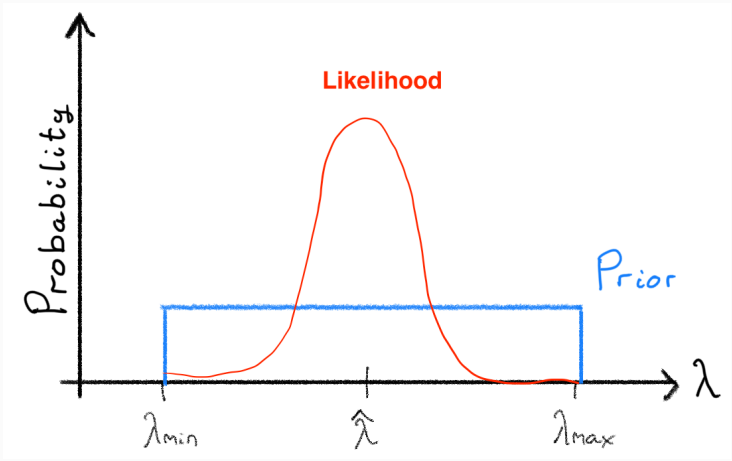
$$P(D|\lambda, B) = \prod_{i=1}^{6} \frac{\exp\left(-\frac{1}{2}(y_i - \lambda)^2\right)}{\sqrt{2\pi}}\,. \tag{iii}$$

The likelihood for model B has a maximum value of $2.84 \times 10^{-3}$ at $\lambda = \hat{\lambda} = 1$. Therefore, if we were using the maximum likelihood ratio in equation 1 to perform model selection, we would favour the more complicated model B.

However, in order to compute the posterior odds ratio in equation 2 we must first evaluate the evidence for model B. This involves first specifying a prior for the free parameter $\lambda$. Here, I will choose a wide, uniform prior symmetric about zero;

$$P(\lambda|B) = \frac{1}{2\Lambda}\mathbb{1}_{(-\Lambda,\Lambda)}(\lambda), \tag{iv}$$

where $\Lambda \gg 1$. The likelihood and prior for model B are sketched in the figure.




The evidence for model B is now given by the integral

$$P(B|D) = Z_B = \int_{-\Lambda}^{\Lambda} \mathrm{d}\lambda \, \frac{1}{2\Lambda} \prod_{i=1}^{6} \left[ \frac{\exp\left(-\frac{1}{2}(y_i - \lambda)^2\right)}{\sqrt{2\pi}} \right] . \tag{v}$$

In the limit of a wide, uninformative prior, $\Lambda \to \infty$, this integral becomes

$$P(B|D) \approx \frac{1}{2\Lambda} \int_{-\infty}^{\infty} \mathrm{d}\lambda \, \prod_{i=1}^{6} \left[ \frac{\exp\left(-\frac{1}{2}(y_i - \lambda)^2\right)}{\sqrt{2\pi}} \right] = \frac{1.45 \times 10^{-3}}{\Lambda} . \tag{vi}$$

Therefore, the posterior odds ratio is given by

$$\mathcal{O}_{A,B} = \frac{P(A|D)}{P(B|D)} \times \frac{P(A)}{P(B)} = \frac{1.41 \times 10^{-4}}{1.45 \times 10^{-3}/\Lambda} \times 1 = 0.0973\Lambda \, . \tag{vii}$$

Notice that if we choose $\Lambda$ to be larger than about 10, then the odds ratio becomes greater than unity and we favour the simpler model A. However, if we choose $\Lambda \lesssim 10$ then we favour model B.

At first, it is surprising (and a little worrying) that the answer seems to depend so much on the apparently arbitrary choice of the prior range $\Lambda$. In particular, if we were to allow the free parameter in model B to take any value (i.e. $\Lambda \to \infty$) it looks like the posterior odds would always be large and we would be forced to favour model A, regardless of what the data looks like. While this is formally

true, in practice models never have parameters that can vary over infinite ranges and after more careful consideration it is only right models with parameters that can vary over wide ranges (or multiple parameters) are properly penalised for this extra freedom.

In this example calculation, if we choose a reasonable value for $\Lambda$, we find a moderate posterior odds ratio ($\log \mathcal{O}_{A,B} \sim 0$) indicating there is no clear preference for either model. This raises an interesting question; how large (or small) does the odds ratio have to be for us to claim that there is "decisive" evidence in favour of a particular model? This is really more of a semantic question than a mathematical one; the answer depends on what you mean by "decisive". However, for Harold Jeffreys answer to this question it is interesting to consider the table shown at https://en.wikipedia.org/wiki/Bayes_factor.

**Occam's Razor:** At this point it is almost obligatory for me to mention the philosophical idea that can be stated something like "the simplest explanation is usually the right one". Some people like to see an analogy between this concept and the Bayesian model selection procedure described above. Some part of this idea seems to be captured by the fact that the evidence ratio penalises a complicated model with many parameters (or parameters that can take values spanning a wider range) and that does not fit the data sufficiently better than a simpler model. This idea is usually attributed to William of Ockham (c. 1287-1347) who was a philosopher, theologian, and Franciscan friar. Ockham did not express the idea mathematically and certainly did not use the idea to perform Bayesian model selection in the way described here.