Bayesian Techniques

David Parkinson

June 20, 2008

Acknowledgements: I would like to thank all my collaborators for related discussions over the last seven years, in particular: Bruce A. Bassett, Rob Crittenden, Martin Kunz, Andrew Liddle, Pia Mukherjee and Roberto Trotta.

This is a course that demonstrates how the manipulation of conditional probabilities using Bayes' theorem allows us to make inferences about the universe in science.

The main reference for this course is **"Data Analysis: A Bayesian Tutorial"** by Devenderjit Sivia (Oxford University Press, 1996). There is also a second edition, Sivia and Skilling (Oxford University Press, 2006). For more technical details I would recommend **"Information Theory, Inference and Learning Algorithms"** by David MacKay (Cambridge University Press, 2003)

1 Introduction

To many statisticians, probability is considered as the frequency outcome of a long-run of experiments. For example, the probability of a coin landing either heads or tails face-up can be found by many flipping it many times. In *frequentist* statistics, it only makes sense to assign a probability to data. A parameter value of some object or theory is a constant, and a hypothesis is either true or false (though these statements are still made quoting degree of confidence).

In Bayesian statistics however, probability represents *degree-of-belief* or plausibility, that is, how likely it is that something is true given the data. Thus parameters can be treated as random variables and have probabilities distributions. Even models can have probabilities associated with them, allowing them to be ranked in order of likelihood. In this way probability

represents states of knowledge, allowing us to make statements both about the universe as we understand it now, but also predictions about what we will see in the future.

2 Basic definitions

In these lectures, we use the symbol P to refer to both probability (such as that for discrete events or ranges of continuous variables) and probability density function (for a continuous variable).

2.1 Sum Rule

For a single discrete event or proposition A, it should be obvious that the probability of it being true or false are related, such that

$$P(A) + P(\text{not}A) = 1\,. \tag{1}$$

Likewise for a set of discrete outcomes $\{A_i\}$,

$$\sum_i P(A_i) = 1 \tag{2}$$

Generalising this to a continuous varaible X, we introduce the probability density function (pdf),

$$P(X_1 < X < X_2) = \int_{X_1}^{X_2} P(X)dX\,. \tag{3}$$

Here the sum rule means that, integrating over all possible outcomes for X,

$$P(-\infty < X < \infty) = \int_{-\infty}^{\infty} P(X)dX = 1\,. \tag{4}$$

2.2 Product Rule

Suppose now there are two separate outcomes A and B. If you wish to verify the truth of $A \wedge B$, one can first verify A and then verify B assuming A, i.e.

$$P(A \wedge B) = P(A) \times P(B|A)\,, \tag{5}$$

or similarly the other way around,

$$P(A \wedge B) = P(B) \times P(A|B)\,. \tag{6}$$

We can do the same thing for continuous distributions, e.g.

$$P(X_1 < X < X_2 \wedge Y < Y_{\max}) = P(X_1 < X < X_2) \times P(Y < Y_1 | X_1 < X < X_2)\,. \tag{7}$$

If A and B are independent, then

$$P(A \wedge B) = P(A) \times P(B)\,, \tag{8}$$

and

$$P(B|A) = P(B)\,. \tag{9}$$

Since $P(A \wedge B) = P(B \wedge A)$, we can derive the following relation

$$\frac{P(A)}{P(B)} = \frac{P(A|B)}{P(B|A)}\,. \tag{10}$$

This is Bayes' theorem, although it is normally written as

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}\,. \tag{11}$$

Here $P(\theta|D)$ is the probability of θ given the data D, known as the *posterior probability*, $P(D|\theta)$ is the probability of the data given the parameters (called the *likelihood*), and $P(\theta)$ is the probability of the parameters before any data has been taken, known as the *prior*. The final term in the equation $P(D)$, is some normalising factor such that the posterior probability over all models is unity, ie.

$$\int P(\theta|D) d\theta = 1\,. \tag{12}$$

2.3 Mean and Variance

For any function $g(X)$, the expectation of $g(X)$ is defined to be

$$\langle g(X) \rangle = \int_{-\infty}^{\infty} f_X(X) g(X) dX \tag{13}$$

where f_X is the probability density of states in X. This is sometimes denoted as $E(g)$ or $g(\bar{X})$.

Any probability density function can be described through its moments, which are the expectation of the different powers of X.

$$\mu' = \langle X^n \rangle = \int_{-\infty}^{\infty} f_X(X) X^n dX \tag{14}$$

The mean is $n = 1$ and the variance $n = 2$.

3 The inverse problem

3.1 A simple case

Consider the following situation. There are identical two urns set up, both filled with an equal number of balls, in this case one hundred. The only difference between the two urns is that one (urn A) is filled with ninety-nine black balls and one white ball, while the other (urn B) is filled with ninety-nine white balls and one black. The balls themselves are exactly identical except for their colours.

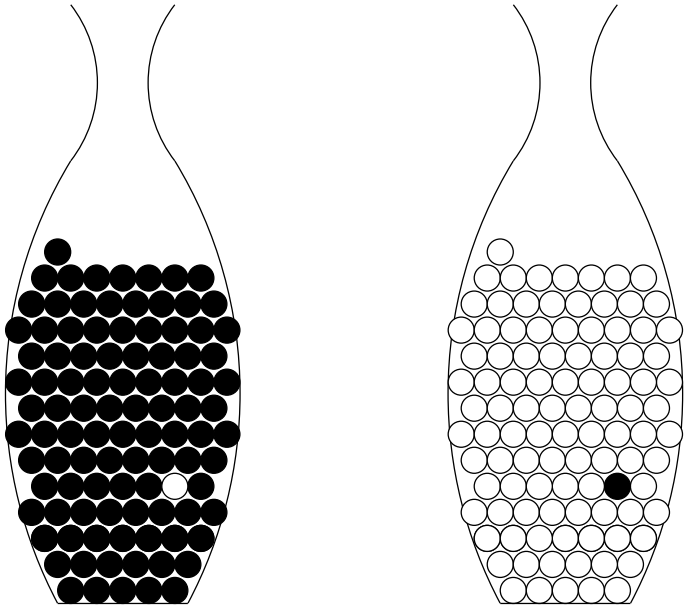

Figure 1: Two urns filled with balls. Urn A (left) contains 99 black balls and 1 white, while urn B contains 99 white balls and 1 black

If the mouth of the urn is only just large enough to admit a human hand, then it is impossible to select a particular colour when reaching for a ball. It should be obvious that the conditional probability of selecting a particular coloured ball at random out of an urn is simply the ratio of the number of balls with colour to the total number of balls in the urn. For example, the probability of selecting a black ball from urn A is

$$P(\text{black ball } |\text{urn A}) = 0.99\,. \tag{15}$$

This is an incredibly simple situation, as by simply knowing which urn you have, it is easy to predict the probabilities of getting black and white

from it. However, in science we rarely, if ever have this situation. Many times we find ourselves going not from model to data (where effectively 'we' have filled the urns), but from data to model (where someone else has done it for us). This is the **Inverse Problem**.

Imagine that the urns have been mixed up, so we do not know which is which. We take a ball of a random urn and find that it is black. The question know is what is probability of that urn being urn a (i.e. what is $P(\text{urn A }|\text{black ball})$?). We can find it using Bayes theorem.

In the case of the urns, we can rewrite Bayes' theorem as

$$P(\text{urn A }|\text{black ball}) = \frac{P(\text{black ball }|\text{urn A})P(\text{urn A})}{P(\text{black ball})}\,. \tag{16}$$

Now the calculation becomes easy. We already know that $P(\text{black ball }|\text{urn A}) = 0.99$, and the prior probability $P(\text{urn A}) = 0.5$, as there are only two of them. Finally since the number of black balls and white balls are equal over all urns, $P(\text{black ball}) = P(\text{whiteball}) = 0.5$. So

$$P(\text{urn A }|\text{black ball}) = \frac{0.99 \times 0.5}{0.5} = 0.99\,. \tag{17}$$

Since the number of urns and the number of balls and the number of urns are the same, this is a maximum symmetry case where the conditional terms vanish, and the probability of the data given the hypothesis is the same as the probability of the hypothesis give the data. However, there are many conditional cases in science when this is not the case.

3.2 A more complex problem

A more complicated situation involves balls of three different colours distributed among the urns. Here ten blacks balls have been replaced by blue balls in urn A, and six white balls have been replaced by blue balls in urn B. The ratio of white balls to total balls in urn A has not changed (nor has the ratio of black to total in urn B), so $P(\text{white ball }|\text{urnA}) = 0.01$ is still the case. However, because the probability of getting a white or black ball **overall** has changed, the posteriors have changed. First we calculate $P(D)$

$$P(\text{black ball}) = \frac{\text{No. black balls}}{\text{Total no. of balls}} = 0.45\,, \tag{18}$$

$$P(\text{white ball}) = \frac{\text{No. white balls}}{\text{Total no. of balls}} = 0.475\,, \tag{19}$$

$$P(\text{blue ball}) = \frac{\text{No. blue balls}}{\text{Total no. of balls}} = 0.075\,. \tag{20}$$

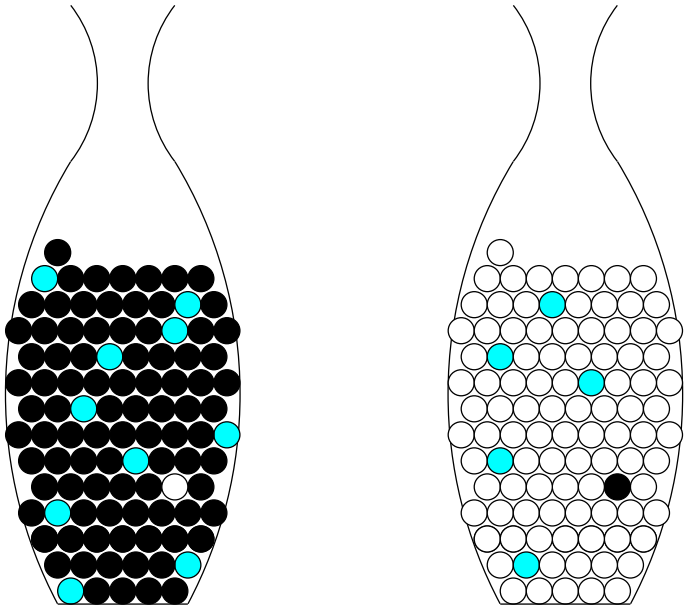

Figure 2: Two urns filled with balls. Urn A (left) contains 89 black balls, 10 blue balls and 1 white, while urn B contains 94 white balls, 5 blue balls and 1 black

	black ball	white ball	blue ball
P(urn A |colour)	$0.9\dot{8}$	0.0105	$0.6\dot{6}$
P(urn B |colour)	$0.0\dot{1}$	0.9895	$0.3\dot{3}$
Total	1.0	1.0	1.0

Table 1: The conditional probabilities of choosing from either urn, given the colour of ball chosen.

Now the probabilities of the different outcomes can easily be calculated. The are shown in table 1.

The inequality between the probability of the data given the hypothesis and the probability of the hypothesis given the data in this case is obvious. The extreme case here is $P(\text{white ball} \mid \text{urn B}) = 0.05$, while $P(\text{urn B} \mid \text{white ball}) = 1/3$, over six times larger. The difference is caused by the conditional probabilities. Although probability in Bayes' theorem seems to be less objective and physically grounded, it is in fact the only logically consistent way to deal with conditional probabilities.

4 Parameter Estimation

In the previous case there were only a finite number of models to choose between (two urns), and the properties of these models (the relative composition of each urn) were well understood. Now we consider a cases where the properties of the model are unknown.

4.1 Flipping a coin

Consider a coin, with different symbols on each side[1]. If a coin is fair, then if the coin is flipped into the air, it is equally likely to come down with either side facing upwards. If a coin is unfair however, it may have some preference to one face or the other. Because it is impossible to know what preference the coin has before you start flipping, the uncertainty can be described by a parameter, in this case represented by the symbol θ.

Here θ represents the bias towards flipping heads: if $\theta = 1$ the coin will always land heads up, whereas if $\theta = 0$, it will always land tails up. We can rewrite Bayes' theorem to describe the posterior probability distribution of this parameter,

$$P(\theta|D, \mathcal{M}) = \frac{P(D|\theta, \mathcal{M})P(\theta|\mathcal{M})}{P(D|\mathcal{M})}\,. \tag{21}$$

The prior ($P(\theta|\mathcal{M})$) is the probability distribution of θ before any data has been taken (before the first flip of the coin). Since we have no prior information, a simple prior to assume would a uniform distribution of probability between $\theta = 0, 1$, i.e.,

$$\begin{aligned} P(\theta|\mathcal{M}) &= 1\,, \quad 0 < \theta < 1 \\ &= 0 \qquad \text{otherwise}\,. \end{aligned} \tag{22}$$

Now through Bayes' theorem we see how the prior is updated by the collection of data to give the posterior probability of the data. Until the first data point is taken the posterior is the same as the prior. As each data point is taken, the likelihood function updates the prior into the posterior.

If each flip of the coin is an independent event, so that the outcome of one does not influence the probability of another, and since probability of a

[1]Most coins have a side where the imprint of a person, such as a current or former head of state, is impressed - this side is called the "heads" side (since the embossing is of the head of a person). The other side may have any imprint, or none, and is called the "tails" side. Technically, the heads and tails sides are known as the obverse and reverse, respectively.[from Wikipedia]

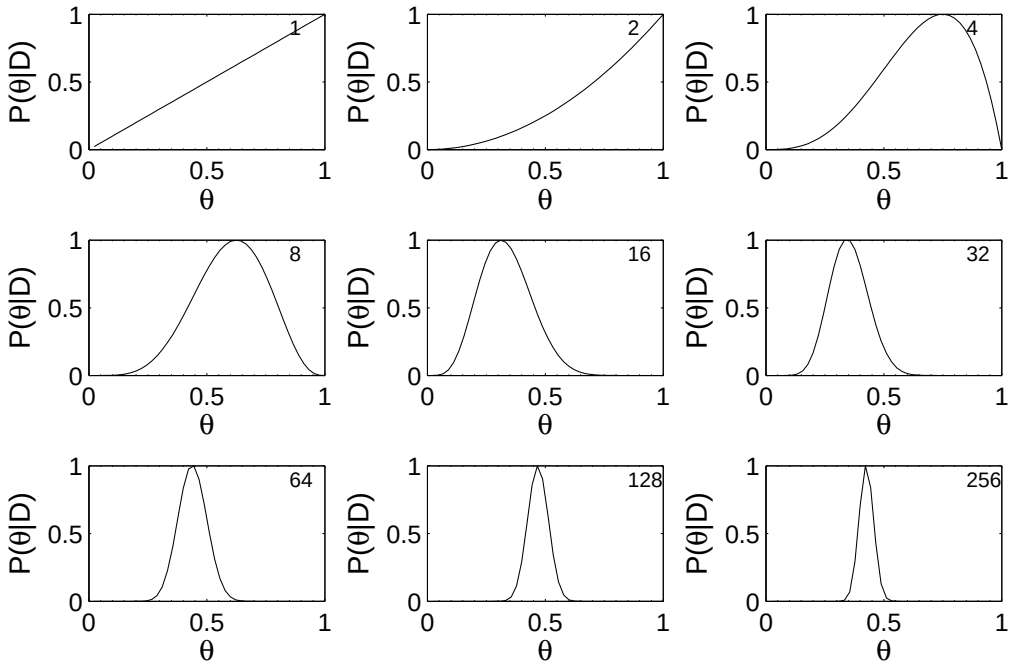

Figure 3: The posterior probabilities of the bias-weighting parameter θ as the number of coin flips increases, assuming a uniform prior on θ. The number in the top-right hand corner indicates the number of coin flips that have been performed.

single head is θ, then the probability of 'R heads in N flips' is given by the binomial distribution,

$$P(D|\theta, \mathcal{M}) \propto \theta^R (1-\theta)^{N-R} \,. \tag{23}$$

Suppose we were to flip this coin once, and it were land tails-up. The probability distribution of θ would resemble the top-left hand plot of Figure 3, with a zero value for $\theta = 1$ rising to unity for $\theta = 0$ (note that the posterior pdf's have not been normalised so that the area under the curve is unity. Instead the most likely parameter value has a posterior set to one).

Yet at this point we do not even know coin has a head (it may be a two-tailed coin). It is only by making more flips of the coin that we start to learn more about the coin, and the value of the bias-weighting parameter θ.

Now it may be that a uniform prior is not appropriate. You may have already seen someone else flipping the coin and not notice any particular bias. In this case a more complicated prior may be a better representation of your initial belief about the fairness of the coin. In Figure 4 we consider an example were a Gaussian prior on θ, centred on the fair value ($\theta = 0.5$), with

one sigma limits of 0.2. Obviously since the Gaussian integrates to unity only in the limits $\theta = -\infty \ldots + \infty$, there is a small amount of prior distribution for parameter values greater than unity or less than zero. This may lead to small errors.

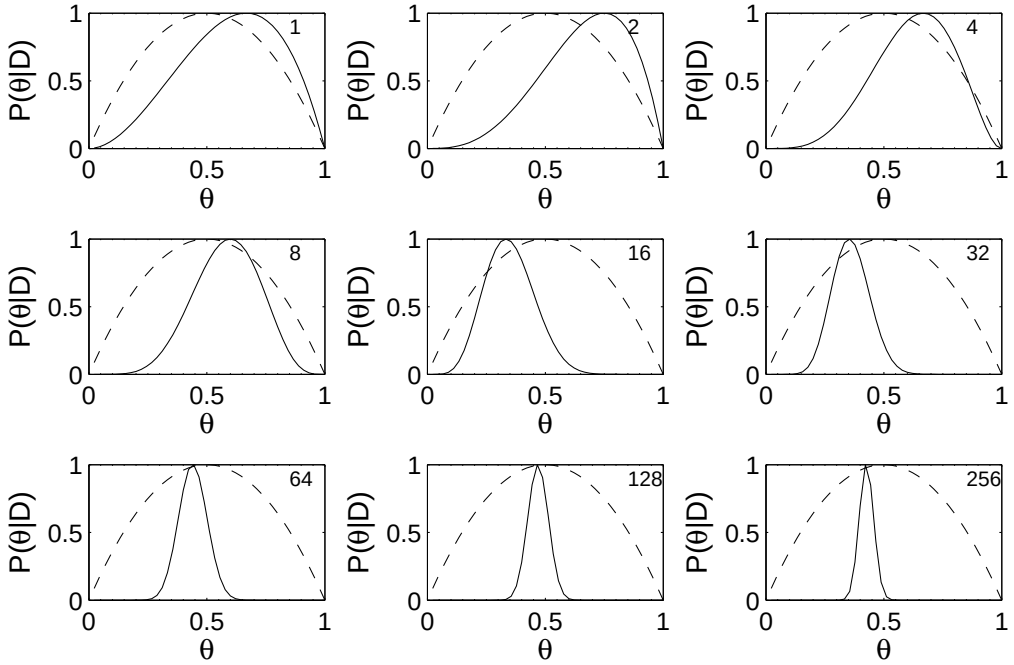

Figure 4: The posterior probabilities (solid curve) of the bias-weighing parameter θ as the number of coin flips increases, assuming a beta distribution prior with $\alpha = 2$ and $\beta = 2$ (and so centred at $\theta = 0.5$ with width $\sigma = 0.22$ (dashed curve)). The number in the top-right hand corner indicates the number of coin flips that have been performed.

A more appropriate choice of prior would be a beta distribution, defined as

$$f(\alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}\,, \tag{24}$$

where $B(\alpha, \beta)$ is the Beta function, defined as

$$B(x, y) = \int_0^1 t^{(x-1)}(1-t)^{(y-1)} dt\,, \tag{25}$$

which serves as a normalising coefficient so that the beta distribution integrates to unity. A good choice of parameters $\alpha = 2$ and $\beta = 2$, which effectively amounts to flipping the coin twice and getting one head and one tail. This way the option of a two-headed or two-tailed coin is ruled out.

As you can see, when the number of flips is small, the prior contains a lot more constraining power than the likelihood, and the posterior is prior-dominated. However, once the number of flips reaches eight or above, the posterior peak starts to shift away from the prior peak, and the data begins to dominate. By the time we reach 256 flips, the effect of the prior is negligible. This is a common occurrence in Bayesian statistics. In the case of bad or unconstraining data, the prior will dominate over the likelihood, and so the answer will often depend on the choice of prior.

4.2 Error estimation

The full probability density function contains all the information about the posterior of the parameter. However, normally it is easier and relatively accurate to just quote two numbers, the best estimate of the parameter, and the measure of the reliability. The first number, the best estimate of the parameter can be found by calculating the value of the parameter that maximises the posterior pdf, i.e.

$$\left.\frac{dP}{d\theta}\right|_{\theta_0} = 0\,, \tag{26}$$

where θ_0 is the best estimate of the parameter value θ. Also you must ensure that this the maximum, not the minimum or a saddle point, so there is a second condition,

$$\left.\frac{d^2P}{d\theta^2}\right|_{\theta_0} < 0\,. \tag{27}$$

Please notice this notation assumes that the function P is continuous in θ. If θ takes discrete values, the best estimate for θ is still the one that maximises the posterior probability, but these equations can't be used as there is no definition of *gradients* in this instance.

To obtain the reliability of the measure of the best fit parameter (sometimes called the error on the parameter value), we need to make an estimate of the width or spread of the posterior probability around θ_0. In this case it is easier to deal with log of the posterior,

$$L = \log_e[P(\theta|D,\mathcal{M})]\,, \tag{28}$$

since this varies much more slowly with θ. Expanding around $\theta = \theta_0$, we get

$$L = L(\theta_0) + \frac{1}{2}\left.\frac{d^2L}{d\theta^2}\right|_{\theta_0}(\theta-\theta_0)^2 + \ldots\,, \tag{29}$$

here the best estimate of θ is given by

$$\left.\frac{dL}{d\theta}\right|_{\theta_0} = 0\,, \tag{30}$$

which is equivalent to equation 26 because L is a monotonic function of P.

The first term is a constant in θ and tells us nothing about the shape of the pdf. The linear term is zero because we are expanding around the maximum. Therefore it is the quadratic term, the curvature, that determines the width of pdf. Ignoring the higher order terms, the posterior can be written as

$$P(\theta|D) = A\exp\left(\frac{1}{2}\left.\frac{d^2L}{d\theta^2}\right|_{\theta_0}(\theta-\theta_0)^2\right)\,, \tag{31}$$

where A is a normalisation constant. In essence the pdf has been approximated by a Gaussian function (also known as a normal distribution), written as

$$P(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\,, \tag{32}$$

where the mean of the function μ has been equated with the best estimate of the parameter θ_0, and the variance σ^2 with the curvature $d^2L/d\theta^2$. Notice the exact relation between these parameters is

$$\sigma = \left(-\left.\frac{d^2L}{d\theta^2}\right|_{\theta_0}\right)^{-1/2}\,, \tag{33}$$

since the curvature is necessarily negative, given condition in equation 27.

If we integrate the pdf over the interval $\theta_0 - \sigma < \theta < \theta_0 + \sigma$, we find that the probability in this region is 67%. Thus the parameter σ is often referred to as the 67% confidence interval, or more commonly, the 1-sigma error bar. It represents the size of the region over which we have a 67% belief that the *true value* of θ lies inside. Similarly the two-sigma limits bounds a region that we assign a 95% probability that the true value lies inside.

If we return to the coin-example, we remember that the posterior probability (assuming a uniform prior on θ stated in eqn. 22) is given by

$$P(\theta|D,\mathcal{M}) \propto \theta^R(1-\theta)^{N-R}\,. \tag{34}$$

Taking the natural log of this function,

$$L = \text{constant} + R\log_e(\theta) + (N-R)\log_e(\theta)\,. \tag{35}$$

For the best estimate of θ and its error bar, we need the first and second derivatives of this function,

$$\frac{dL}{d\theta} = \frac{R}{\theta} - \frac{(N-R)}{(1-\theta)} \qquad \text{and} \qquad \frac{d^2L}{d\theta^2} = -\frac{R}{\theta^2} - \frac{(N-R)}{(1-\theta)^2}\,. \tag{36}$$

First we can find the best estimate for θ,

$$\frac{dL}{d\theta}\Big|_{\theta_0} = \frac{R}{\theta_0} - \frac{(N-R)}{(1-\theta_0)} = 0\,, \tag{37}$$

which gives us

$$\theta_0 = R/N\,. \tag{38}$$

As you might expect, the best estimate for the bias of a coin to prefer landing heads-up to tails-up is given by the ratio of the number of heads to the total number of flips. We now calculate the error on this estimate,

$$\begin{aligned} \frac{d^2L}{d\theta^2} &= -\frac{R}{\theta_0^2} - \frac{(N-R)}{(1-\theta_0)^2}\,, \\ &= -\frac{N}{\theta_0(1-\theta_0)}\,. \end{aligned} \tag{39}$$

So the error-bar for the distribution is given by

$$\sigma = \sqrt{\frac{\theta_0(1-\theta_0)}{N}}\,. \tag{40}$$

In Figure 5 we make a comparison between the predicted Gaussian errors and the actual posterior. We see that the predicted Gaussian errors are a bit smaller than the actual posterior width.

Through this analysis we can see that the Gaussian prior we chose in the previous section, which had a mean of $\mu = 0.5$ and standard deviation of $\sigma = 0.2$, would be equivalent to a number of flips $N_{\rm prior}$, with an equal number of heads and tails $R_{\rm prior} = N_{\rm prior}/2$, made by some previous experimenter. Solving the equations for μ and σ in terms of θ_0 and $N_{\rm prior}$, we find that $N_{\rm prior} \approx 6$.

In general, the posterior will not be easily solvable, such as cases with non-uniform priors, where the derivatives have to be taken from the product of the prior and the likelihood, or where the likelihood is not such a simple analytic function. There are also cases where the posterior will be asymmetric around the best estimate of the parameter. In this case, the best course of action is to find the values θ_1 and θ_2, such that they are the smallest interval

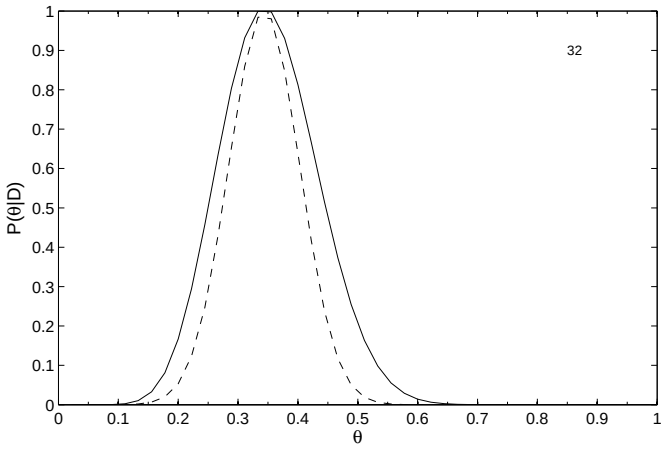

Figure 5: The posterior probabilities (solid curve) of the bias-weighting parameter θ for 32 coin flips, comparing the calculated posterior, and a posterior curve assuming a Gaussian distribution.

that encloses 95% of the area (i.e. the two sigma limits on the parameter). Assuming that the posterior has been normalised to have unit area, this means we need to find

$$P(\theta_1 \leq \theta < \theta_2|D, \mathcal{M}) = \int_{\theta_1}^{\theta_2} P(\theta|D, M)d\theta = 0.95\,, \tag{41}$$

such that the interval $\theta_2 - \theta_1$ is as small as possible. Please note that the 95% figure is just a convention in the literature, and you can easily calculate the 50%, 70% or 99% limits as you prefer.

5 Multi-parameter estimation and errors

In the previous section we considered the case where there is only unknown parameter that needs to be determined. However, more commonly the model will involve many parameters, some of which are of interest (because they describe some particular piece of science), while others will merely be a nuisance. We also generalise error bars to error ellipses, and consider correlations between errors.

5.1 A signal in the presence of a background

In many branches of science we are faced with the problem of estimating the amplitude of a signal in the presence of a background (e.g. astronomical

spectrum contaminated by stray light from the night sky). An idealised situation would have a flat background of unknown magnitude B, with the signal of unknown amplitude A but known shape and position. Such a signal is sketched in figure 6.

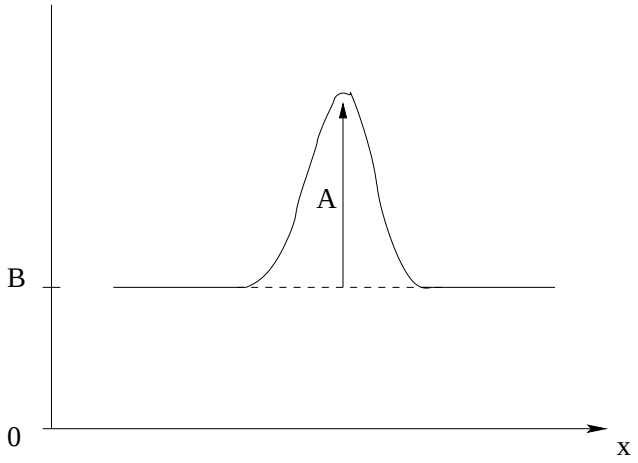

Figure 6: A schematic illustration of a signal peak of amplitude A with a flat background B.

The data for such problems are normally integer valued, corresponding to counts in a detector. The question we ask is, given a set of counts $\{N_k\}$ in a set of data channels x_k what is the best estimate for A and B?

The signal in the k-th data channel is proportional to the sum of the signal and the background at x_k; taking the signal to be a Gaussian peak centred at x_0 with a width of w the ideal data will be given by

$$D_k = n_0[Ae^{-(x-x_0)^2/2w^2} + B]\,, \tag{42}$$

where n_0 is a constant related to the amount of time data is taken. However the number of counts is an integer value, related to the signal by a Poisson distribution:

$$P(N|D) = \frac{D^N \mathrm{e}^{-D}}{N!}\,. \tag{43}$$

This equation is of course the likelihood of N counts in channel k, i.e.

$$P(N_k|A, B, \mathcal{M}) = \frac{(D_k)^{N_K} \mathrm{e}^{-D_k}}{N_k!}\,, \tag{44}$$

where the model $\mathcal{M}$ contains all information about the relationship between the expected number count D_k and the parameters A and B. For the Gaussian peak given in eq. 42 this means the values of N_0, x_0 and w. We assume

that the count in each channel is independent, so the probability of the ensemble of data $\{N_K\}$ is just just the product of the individual probabilities,

$$P(\{N_k\}|A,B,\mathcal{M}) = \prod_{k=1}^{M} P(N_K|A,B,\mathcal{M})\,, \tag{45}$$

where there are M data channels.

Once again we use Bayes' theorem to calculate the full posterior on A and B.

$$P(A,B|\{N_k\},\mathcal{M}) \propto P(\{N_k\}|A,B,\mathcal{M}) \times P(A,B|\mathcal{M}) \tag{46}$$

The likelihood has been defined in equation 45. The most simplistic prior we can assume is that both A and B are positive definite, i.e.,

$$\begin{aligned} P(A,B|\mathcal{M}) &= \text{constant}\,, \quad \text{for } A \geq 0 \text{ and } B \geq 0 \\ &= 0 \qquad\qquad\quad \text{otherwise}\,. \end{aligned} \tag{47}$$

Although we should have upper bounds on A and B, as long as they are not too small to impose a significant cut-off on the posterior, their effect will be negligible.

Given the prior given by eqn. 47 and the likelihood given in eqn. 45 we can derive the log of the posterior,

$$L = \log_e[P(A,B|\{N_k\},\mathcal{M})] = \text{constant} + \sum_{k=1}^{M} N_k \log_e(D_k) - D_k\,, \tag{48}$$

where the constant term involves all terms that do not involve A or B.

In Figure 7 we simulated four different cases, with different amounts of data and different numbers of channels. The first panel shows the number of counts detected in 15 data-bins, with n_0 chosen to give a maximum expectation of around 100 counts. The second (right hand) panel gives the corresponding posterior ellipse for A and B. The second panel down shows the data for the same set up, but only collected for one tenth of the time (so the number of counts collected will be a tenth of the first set up). The third panel down we return to the original count rate, but with 31 channels spread over twice the measurement range. And the final right hand panel shows the set-up with 7 channels spread over half the original range. In all cases the signal is centre at $x = 0$, with a full width-half maximum of five units.

As you can see, as the amount of data is decreased, either by taking measurements over less time or by decreasing the number of channels, the error on the two parameters increases, and the posterior ellipses increase in size. Also, when the data is sampled over a smaller range, the amount of

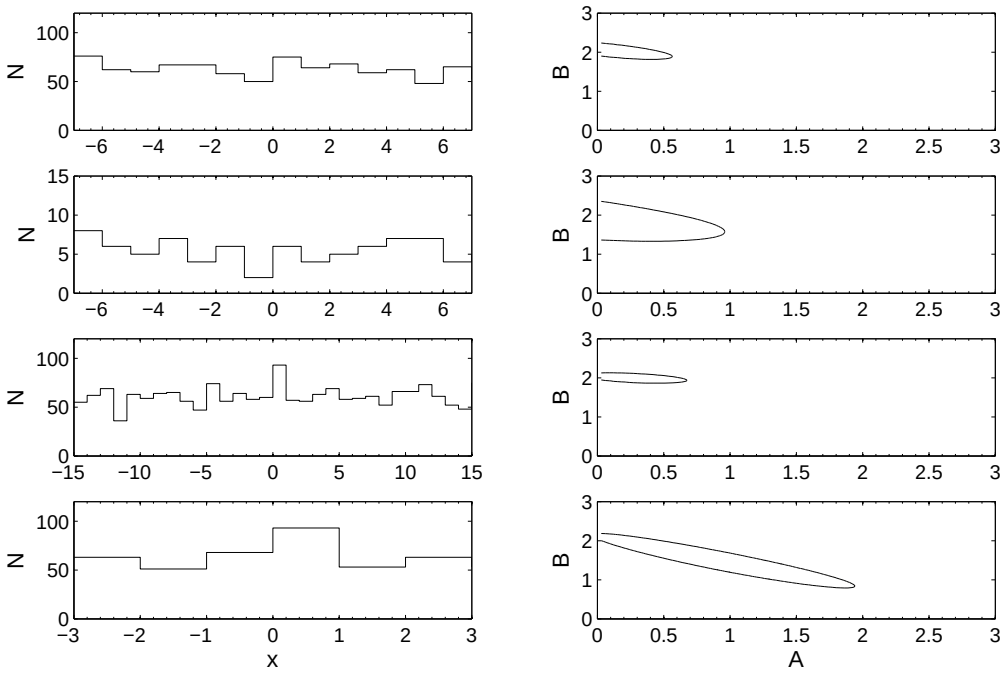

Figure 7: Poisson data and the resulting posteriors for the Amplitude A of a Gaussian signal peak, centred at the origin (x=0) with FWHM of 5 units, and the flat background B, for four different experimental set-ups.

noise being sampled compared to the amount of signal is smaller, and so the two parameters become correlated. This can be seen by comparing the third and fourth rows of Figure 7. In the posterior plot of the third pair of panels (where there are 31 channels), the error ellipse is almost horizontal, representing a near independent measurement of two parameters. In contrast the final two panels, where there are only 7 channels, has a very diagonal error ellipse, showing a very correlated measurement of the two parameters.

5.2 Marginalised posteriors

The 2-D posterior describes completely the inference about the values of A and B. Quite often we are simply not interested in the background, it is simply a nuisance parameter we have to include to do the likelihood. What we really require is the probability distribution of A by itself, which we can get by marginalising over B, in the following manner:

$$P(A|D,\mathcal{M}) = \int_0^\infty P(A,B|D,\mathcal{M})dB\,. \tag{49}$$

Alternatively we may not be interested in the peak, in which case we can do the alternative operation,

$$P(B|D,\mathcal{M}) = \int_0^\infty P(A,B|D,\mathcal{M})dA\,. \tag{50}$$

These are known as the marginal distributions of the parameters. The four pairs of marginal posteriors corresponding to the data-sets, and full posteriors, of Fig. 7, are plotted in Fig. 8

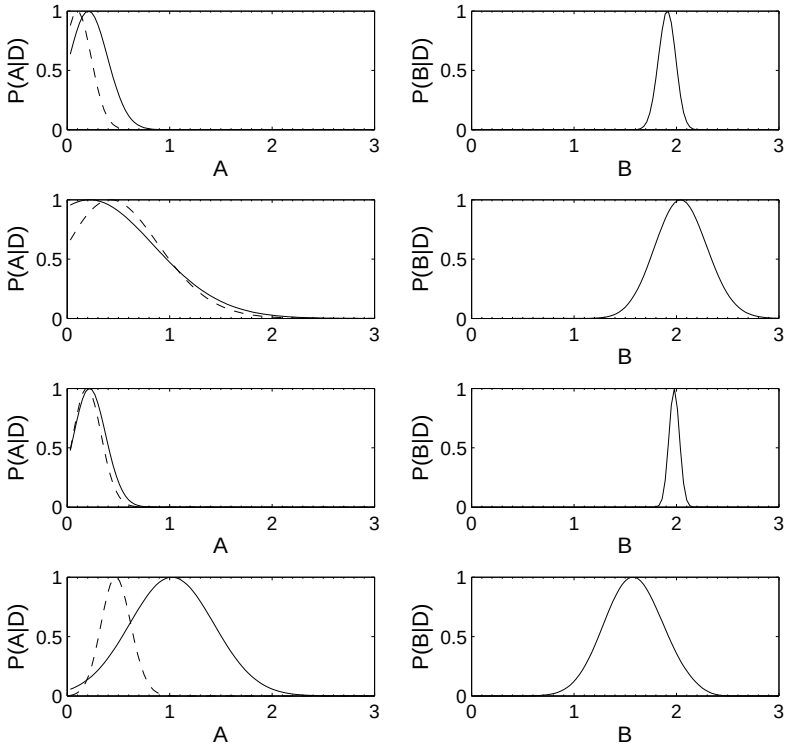

Figure 8: The marginal distributions for the amplitude A and the background B corresponding to the Poisson data, and the resulting posteriors, for the four different experiment set-ups in Fig. 7. The dashed line shows the posteriors of A conditional on knowing the true value of B.

Please note that the marginal distribution $P(A|D,\mathcal{M})$ is not the same as the conditional posterior $P(A|D,B,\mathcal{M})$. One assume a complete state of prior ignorance on B, while the other would be more appropriate when the value of B has been determined by some calibration experiment. This conditional posterior, where $B = 2$ in the likelihood analysis, is also plotted in Figure 8 for comparison. As you can see the act of marginalising over a nuisance parameter will degrade the constraints on the parameter of interest.

6 Model selection

(Note: some of this material is replicated from the paper *Measuring the effective complexity of cosmological models* by Kunz, Trotta and Parkinson, Phys.Rev. D74 (2006) 023503)

If we return now to Bayes' theorem,is takes the form:

$$P(\mathcal{M}|D) = \frac{P(D|\mathcal{M})P(\mathcal{M})}{P(D)}\,. \tag{51}$$

The expression in the denominator is a normalisation constant and can be computed by integrating over the parameters,

$$P(D|\mathcal{M}) = \int P(D|\theta,\mathcal{M})P(\theta|\mathcal{M})d\theta\,. \tag{52}$$

This corresponds to the average of the likelihood function under the prior and it is the fundamental quantity for model comparison. The quantity $P(D|\mathcal{M})$ is called *marginal likelihood* (because the model parameters have been marginalised), or alternatively, the *evidence*.

The evidence can be approximated by

$$P(D|\mathcal{M}) \approx P(D|\theta^*,\mathcal{M}) \times \frac{\delta\theta}{\Delta\theta} \tag{53}$$

where $P(D|\theta^*,\mathcal{M})$ is the likelihood of the best fit parameters to the data (θ^*), and $\delta\theta/\Delta\theta$ represents the amount of space occupied by the posterior ($\delta\theta$) relative to the prior ($\Delta\theta$). In this way, the second term embodies a kind of **Occam factor** that rewards highly predictive models with a small number of parameters and penalises models with an unnecessarily large parameter space.

The posterior probability of the model is, using Bayes theorem again,

$$P(\mathcal{M}|D) = \frac{P(D|\mathcal{M})P(\mathcal{M})}{P(d)}\,, \tag{54}$$

where $p(\mathcal{M})$ is the prior for the model. The quantity in the denominator on the right-hand side is just a normalisation factor depending on the data alone, which we can ignore. When comparing two models, $\mathcal{M}_1$ versus $\mathcal{M}_2$, one introduces the *Bayes factor* B_{12}, defined as the ratio of the models' evidences,

$$\frac{P(\mathcal{M}_1|D)}{P(\mathcal{M}_2|D)} = \frac{P(\mathcal{M}_1)}{P(\mathcal{M}_2)}B_{12}\,. \tag{55}$$

In other words, the odds of the models are updated by the data by the Bayes factor. If we do not have any special reason to prefer one model over the

other before we see the data, then $P(\mathcal{M}_1) = P(\mathcal{M}_2) = 1/2$, and the posterior odds reduce to the Bayes factor.

Although the evidences gives a rank-ordered list of models, it is still necessary to decide how big a difference in evidence is needed to be significant. If the prior probabilities of the models are assumed equal, the difference in log(evidence) can be directly interpreted as the relative probabilities of the models after the data. Even if people disagree on the relative prior probabilities, they will all agree on the direction in which the data, represented by the evidence, has shifted the balance. The usual interpretational scale employed is due to Jeffreys (from his classic 1961 book 'Theory of Probability'), which states

$\Delta \ln E < 1$	Not worth more than a bare mention.
$1 < \Delta \ln E < 2.5$	Significant.
$2.5 < \Delta \ln E < 5$	Strong to very strong.
$5 < \Delta \ln E$	Decisive.

In practise we find the divisions at 2.5 (corresponding to posterior odds of about 13:1) and 5 (corresponding to posterior odds of about 150:1) the most useful.

6.1 Fitting data with a polynomial

We consider the classic problem of fitting data drawn from a polynomial of unknown degree. The models which we test against the data are a collection of polynomials of increasing order, starting with a constant, then moving up to a line, a quadratic etc. The question is then whether our model selection can correctly recover the order m of the polynomial from which the data are actually drawn.

We assume that the noise is well modelled by a Gaussian process, and so the probability of an individual data point can be written as

$$P(D_k|\theta, \mathcal{M}) = \frac{1}{\sigma_k \sqrt{2pi}} \exp\left(-\frac{(F_k - D_k)^2}{2\sigma_k^2}\right) , \tag{56}$$

where the model $\mathcal{M}$ implicitly includes information about the expected size of the error bars, and the functional relationship, f, between the model parameters $\vec{\theta}$ and the noiseless data F,

$$F_k = f(\vec{\theta}, k) . \tag{57}$$

These two equations allow us to approximate the likelihood as

$$P(D|\theta, \mathcal{M}) \propto \exp\left(-\frac{\chi^2}{2}\right) , \tag{58}$$

where χ^2 is the sum of the *normalised residuals*

$$\chi^2 = \sum_{k=1}^{N} \left(\frac{F_k - D_k}{\sigma_k} \right)^2 . \tag{59}$$

For definitiveness, let us take $m = 6$ for the underlying model, and we generate $N = 10$ data points with noise $\sigma = 1/100$ for all data points. The prior over the polynomial coefficients $\vec{\theta}$ is a multivariate Gaussian with covariance matrix given by the identity matrix. The relationship between the parameters and the noiseless data is then simply

$$y = \sum_{i=0}^{m} \theta_i x^i \, , \tag{60}$$

where the zeroth element represents a constant term.

We plot in Figure 9 the model likelihood as a function of the polynomial order, n. The evidence of the models increases rapidly until $n = m$ and then stabilising, signalling that $n > 6$ parameters are not justified (in the figure we plot $\log P(D|\mathcal{M})$, which increases rapidly and then flattens). We conclude that the model with $n = 6$ is the one preferred by data.

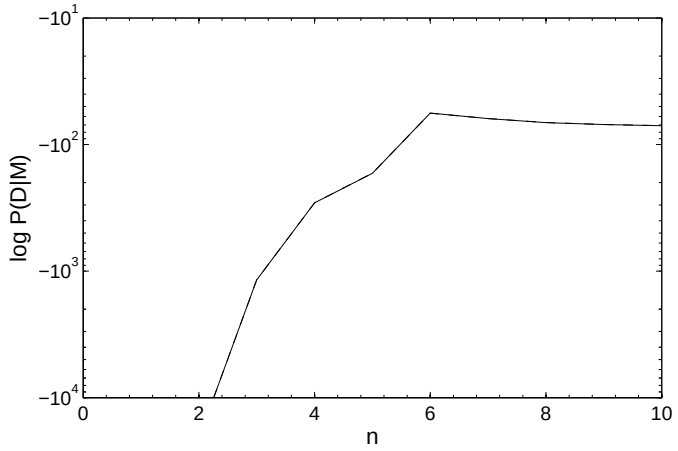

Figure 9: The evidence for the model as a function of the model index, n, which gives the order of the polynomial to the fit the data generated from a polynomial of order m=6. The curve rises steeply at the beginning, until n=6 when the data can be well fit by the model. After that, the extra parameters, which are unnecessary to fit the data well, slowly decrease the evidence through the 'Occam factor'.