Chapter 12

Bayesian Inference

While Maximum Likelihood for model-based inference in phylogenetics is common, the Bayesian framework is also widely used. Like Maximum Likelihood, it is a general approach, that can be applied in any statistical setting where a probabilistic model has been formulated. Although Bayesian methods are not new, computational difficulties often prevented their application for complex problems until recent decades.

While there are significant philosophical differences between these two frameworks, their mathematical formulations are not so far apart. Despite a long-running debate between some of the supporters of each, most practitioners adopt a more pragmatic approach, accepting both as reasonable.

12.1 Bayes' theorem and Bayesian inference

As in Chapter 8, our basic problem of statistical inference is the following: Having formulated a probabilistic model of a process, where the model depends on some unknown values of parameters, from a data set we wish to infer the parameter values. In phylogenetics, this typically would mean we have chosen to use a model, such as GTR, to describe the evolution of a collection of sequences along a tree, and the data are the observed sequences from the leaves of the tree. The unknown parameters are all of the numerical parameters of the GTR model (the stable base composition and the relative rates) as well as the topological tree and all of its branch lengths. But that is a rather complicated model, so it's best to begin with a simpler example. We return to the same one used in Chapter 8.

Example. Our experiment is the toss of a (possibly unfair) coin. Our model is simply that the toss will produce heads with probability p and tails with probability $1 - p$. The parameter here is p, which might have any numerical value from 0 to 1.

With Maximum Likelihood, we sought a single best value of p based on the data, and ultimately produced the estimate $\hat{p}$ =(number of heads)/(number of

tosses).

In Bayesian inference we adopt a more elaborate viewpoint that we should not necessarily give a single estimate for $\hat{p}$, but rather a range of values, together with a measure of how much support each has. After all, if we obtained 63 heads out of 100 it is quite reasonable that p was not 0.63. Perhaps it was 0.60, or 0.58, or even 0.50. As we consider values further from .63, we may feel that they become less reasonable, but we are not inclined to completely rule them out.

To capture this more formally, the goal of Bayesian inference is not to infer a single best choice of parameters to fit data, but rather to associate to each possible choice of parameter values a probability of those being the ones that produced the data. This probability measures the support for the parameter values, with values near 1 indicating strong support, and those near 0 indicating essentially no support.

The key to obtaining these probabilities is *Bayes' Theorem*, which has already appeared in equation (8.2), though we repeat it more carefully here: For any two events A and B,

$$\mathcal{P}(A \mid B)\mathcal{P}(B) = \mathcal{P}(A \text{ and } B) = \mathcal{P}(B \mid A)\mathcal{P}(A),$$

so solving for $\mathcal{P}(A \mid B)$ yields

$$\mathcal{P}(A \mid B) = \frac{\mathcal{P}(B \mid A)\mathcal{P}(A)}{\mathcal{P}(B)}.$$

If we think of both our data and the value of the parameters p as being probabilistically determined, then we obtain the special case

$$\mathcal{P}(p \mid \text{data}) = \frac{\mathcal{P}(\text{data} \mid p)\mathcal{P}(p)}{\mathcal{P}(\text{data})}. \tag{12.1}$$

Now the Bayesian perspective requires that we give meanings to the terms on the right hand side of this equation, in order that we can compute the left hand side.

We begin with $\mathcal{P}(p)$, the probability of a specific parameter value p. In the likelihood framework, the viewpoint is that there simply is some unknown value of p, and it doesn't really make sense to talk about its probability. From a Bayesian perspective, however, we think of $\mathcal{P}(p)$ as capturing the support we feel a value p has, or equivalently, our belief that it is the true value. Since $\mathcal{P}(p)$ does not depend on the data, in this equation it must represent support for values of p before we consider the data. It is therefore called the *prior distribution* of p, since it captures our *a priori* beliefs. In contrast $\mathcal{P}(p \mid \text{data})$ measures support for p after the data has been taken into account. It is called the *posterior distribution* of p since it captures our *a posteriori* beliefs.

Already the broad outline of the Bayesian approach has been given. We begin an analysis of data by specifying a prior distribution on the parameters, which indicates our current beliefs in what values are reasonable. Then equation

(12.1) is used to take into account both the data and our prior beliefs to produce updated, posterior beliefs.

To do this, though we must examine the other terms in equation (12.1). In the numerator of the right side, we have $\mathcal{P}(\text{data} \mid p)$, which is simply the likelihood function that formed the basis for Maximum Likelihood inference, and thus something we know how to compute.

In the denominator, we have $\mathcal{P}(\text{data})$. In the likelihood framework this is a rather nonsensical concept, since after all, we collected the data so if there is a probability associated to it, it must be 1. From the Bayesian viewpoint, however, we have prior beliefs about the parameter values, and so we could use them to compute the probability of obtaining any specific data. More specifically, we can set

$$\mathcal{P}(\text{data}) = \sum_p \mathcal{P}(\text{data} \mid p)\mathcal{P}(p),$$

where the sum is over all possible choices of the parameter values.

Thus our final formula for how we update the prior distribution to obtain the posterior one is

$$\mathcal{P}(p \mid \text{data}) = \frac{\mathcal{P}(\text{data} \mid p)\mathcal{P}(p)}{\sum_p \mathcal{P}(\text{data} \mid p)\mathcal{P}(p)}, \tag{12.2}$$

Example (1). Returning to the coin toss example, for simplicity let's suppose we know the coin that is flipped is weighted in one of 3 ways: p is either 1/4, 1/2, or 3/4. With no data collected, we might choose for a prior the probabilities 1/3, 1/3, 1/3 that p has each of these values. (Note that while p itself is a probability, the prior assigns probabilities to each of its possible values, so the prior, in this case, gives us the probability of a probability.)

Now we collect some data by flipping the coin 3 times, obtaining the sequence HHT. The likelihood is computed as in Chapter 8 to be

$$\mathcal{P}(\text{HHT} \mid p) = p^2(1-p).$$

We use this, and the prior $\mathcal{P}(p)$ to compute the denominator $\mathcal{P}(\text{data})$ in equation (12.2) as

$$\begin{aligned}\mathcal{P}(\text{HHT}) &= \mathcal{P}(\text{HHT} \mid p = 1/4)\mathcal{P}(p = 1/4) + \mathcal{P}(\text{HHT} \mid p = 1/2)\mathcal{P}(p = 1/2) \\ &\quad + \mathcal{P}(\text{HHT} \mid p = 3/4)\mathcal{P}(p = 3/4) \\ &= \left(\frac{1}{4}\right)^2\left(\frac{3}{4}\right)\left(\frac{1}{3}\right) + \left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)\left(\frac{1}{3}\right) + \left(\frac{3}{4}\right)^2\left(\frac{1}{4}\right)\left(\frac{1}{3}\right) \\ &= \frac{3+8+9}{4^3 \cdot 3} = \frac{20}{192} = 0.104166\ldots.\end{aligned}$$

Finally we use equation (12.2) for each of the possible values of p to obtain

the posterior distribution:

$$\begin{aligned}\mathcal{P}(p = 1/4 \mid \text{data}) &= \frac{\mathcal{P}(\text{data} \mid p = 1/4)\mathcal{P}(p = 1/4)}{\sum_p \mathcal{P}(\text{data} \mid p)\mathcal{P}(p)} \\ &= \frac{\left(\frac{1}{4}\right)^2 \left(\frac{3}{4}\right) \left(\frac{1}{3}\right)}{\frac{20}{192}} \\ &= \frac{3}{192}\frac{192}{20} = \frac{3}{20} = 0.15\end{aligned}$$

$$\begin{aligned}\mathcal{P}(p = 1/2 \mid \text{data}) &= \frac{\mathcal{P}(\text{data} \mid p = 1/2)\mathcal{P}(p = 1/2}{\sum_p \mathcal{P}(\text{data} \mid p)\mathcal{P}(p)} \\ &= \frac{\left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right) \left(\frac{1}{3}\right)}{\frac{20}{192}} \\ &= \frac{8}{192}\frac{192}{20} = \frac{8}{20} = 0.40\end{aligned}$$

$$\begin{aligned}\mathcal{P}(p = 3/4 \mid \text{data}) &= \frac{\mathcal{P}(\text{data} \mid p = 3/4)\mathcal{P}(p = 3/4)}{\sum_p \mathcal{P}(\text{data} \mid p)\mathcal{P}(p)} \\ &= \frac{\left(\frac{3}{4}\right)^2 \left(\frac{1}{4}\right) \left(\frac{1}{3}\right)}{\frac{20}{192}} \\ &= \frac{9}{192}\frac{192}{20} = \frac{9}{20} = 0.45.\end{aligned}$$

Thus the posterior probabilities of the coin having probability of heads 1/4, 1/2, and 3/4, are 0.15, 0.40, and 0.45, respectively. Note that these add to 1, as they must since they specify a probability distribution.

In comparison to the prior, this posterior indicates we have shifted our belief to one where p is larger. The most probable value of p is 3/4, but there is not much more support for that than for $p = 1/2$. On the other hand, the support for $p = 1/4$ has decreased significantly to less than half its prior probability.

As should be clear, computing posterior probabilities using equation (12.2) in even simple examples requires a long computation. In addition to needing the same likelihood function as was used in Chapter 8, and a prior distribution on the parameters, the denominator involves a sum of products of these over all possible parameter values. If in this example we had said the probability p of heads could have any of the values $0.1, 0.2, 0.3, \ldots, 0.9$, then there would have been 9 summands rather than 3. Computing this denominator thus becomes a major computational hurdle when there are many possible values of the parameters.

Example (2). In the coins toss example above, we began by saying there were only 3 possible values for the parameter p. For a more elaborate example, we

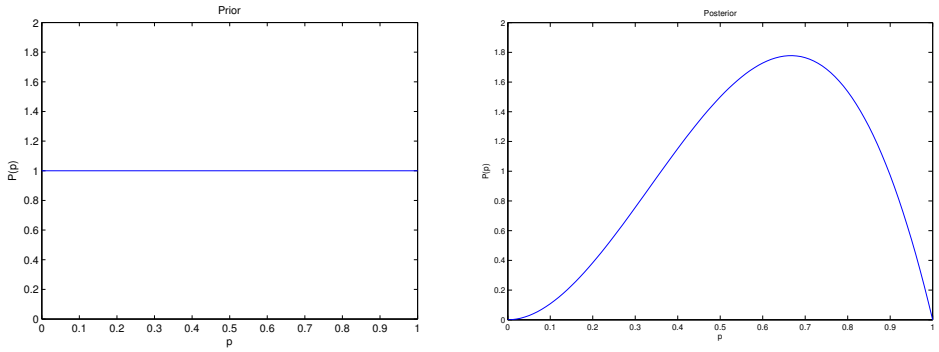

Figure 12.1: (a) A flat prior for a coin with probability p of heads, and (b) the posterior after the observation HHT.

could instead say p could have *any* of the continuum of values between 0 and 1. Then we must specify the prior distribution by a probability density function, $f(p)$, defined on this interval. The probability of the parameter lying in a small interval of size dp around the value p is then $\mathcal{P}(p) = f(p)dp$.

To express complete ignorance of the values of p, we might choose this to be a constant function $f(p) = c$. The specific value of c is determined by the need for the total probability, $\int_0^1 f(p)\,dp$ to be 1, so we use $c = 1$. (Any probability density for a continuous parameter must have total area 1 under its graph.) The formula for the posterior density then becomes

$$f(p \mid \text{data}) = \frac{\mathcal{P}(\text{data} \mid p) f(p)}{\int_0^1 \mathcal{P}(\text{data} \mid p) f(p)\,dp},$$

so

$$f(p \mid \text{HHT}) = \frac{p^2(1-p)\cdot 1}{\int_0^1 p^2(1-p)\cdot 1\,dp}.$$

Since

$$\int_0^1 p^2(1-p)\cdot 1\,dp = \int_0^1 p^2 - p^3\,dp = \frac{p^3}{3} - \frac{p^4}{4}\bigg|_0^1 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12},$$

the posterior density is

$$f(p \mid \text{HHT}) = 12p^2(1-p).$$

Graphs of the prior and posterior are show in Figure 12.1. Note how there is a significant shift toward probabilities in the vicinity of 2/3 being the most probable.

To summarize, and contrast the Bayesian inference approach to the of Maximum Likelihood, we list a few key points

1. Both are model-based methods in which a probabilistic description of the data generation process is assumed.

2. Both involve computing $\mathcal{P}(\text{data} \mid p)$ for parameter values p. In Maximum Likelihood, this is the Likelihood function we seek to maximize; in a Bayesian analysis it appears in the process of computing the posterior distribution.

3. A Bayesian analysis requires specifying a prior distribution of the parameters we seek to infer, which expresses our beliefs in what parameter values are likely before the data are considered. For Maximum Likelihood we do not need to express these prior beliefs, and in fact cannot take such prior beliefs into account.

4. Maximum Likelihood produces a single estimate, called a *point estimate*, of the parameters we infer. A Bayesian analysis produces a distribution of estimates, the posterior distribution, which indicates differing levels of support for different parameter values.

12.2 Prior Distributions

The requirement that we specify a prior for a Bayesian analysis is a major difference from what is needed to perform Maximum Likelihood. Depending on one's viewpoint and the application, this can be either a positive or negative feature.

In the majority of phylogenetic analyses, the priors are chosen to express as much ignorance as possible. Such a prior is often called *uninformative*, though it may carry some information, such as the fact that a parameter must lie in a certain range. For example, the prior depicted in Figure 12.1 indicates that p lies between 0 and 1, but its "flatness" or uniformity indicates that we have no further information on its likely value. By choosing uninformative priors, one hopes that the posterior distribution will indicate signal in the data alone, and not any assumptions that a different prior might express.

For a few of the parameters in a phylogenetic model, creating an uninformative prior is easy. For instance if an model allowing some invariable sites is used, then the proportion of invariable sites can be given such a uniform prior across the range 0 to 1, just as in the coin toss example.

For the base distribution, an uninformative prior is not much more complicated. Since $\mathbf{p} = (p_A, p_G, p_C, p_T)$ with $p_A + p_G + p_C + p_T = 1$, we can use the uniform distribution which assigns to each choice of $\mathbf{p}$ the same value, 1 divided by the volume of the region in space with $x, y, z \geq 0, x + y + z \leq 1$. Since this is a bounded region, its volume is finite. This is a special case of the *Dirichlet distribution*, which has several parameters (or, as they are often called, *hyperparameters* since these are parameters of the prior and not of the phylogenetic model) that can be varied to range from the uniform distribution to ones increasingly concentrated at any point. (See Exercise 5.) The Dirichlet

distribution can thus be used to provide as informative a prior as is desired. A similar uninformative prior can be used for substitution rates, since they can be normalized to sum to 1, by adopting an appropriate time scale.

Priors for edge lengths are more problematic, since lengths can range over all non-negative real numbers. Since this set is unbounded, a uniform distribution is not possible as a valid prior, since a constant function on an unbounded interval would not have area 1 under its graph. One possibility is to use a uniform distribution on a large interval from 0 up to a value believed to safely exceed any plausible edge length. However, Yang and Rannala (2005) have, through a mixture of theoretical analysis and simulation studies, shown this can lead to posteriors that are concentrated on longer edge lengths, and thus perhaps bias results in a way that was not intended. An alternative is to use a prior that decays exponentially with branch length. Although that concentrates more probability on short edge lengths, that may actually be a better representation of what we might consider to be uninformative than a flat distribution.

Finally, for the tree topology the simplest approach is that we make all binary topologies have the same probability. Since there are only finitely many possibilities, this gives a valid prior.

These choices of priors have all been made in the hopes of expressing little to no information that will be used in the Bayesian analysis. Whether they fully succeed is a difficult issue. For instance, do we really believe all binary trees are really equally likely? What about the prior for branch lengths where there seems to be no obvious way to even say what truly uninformative would mean? This dependence on priors when we may have no objective way of choosing them is the reason some feel discomfort with Bayesian methods in general.

From a more pragmatic perspective though, we can make reasonable choices of priors, and then examine the resulting posterior distribution. If the analysis gives posteriors that are very different from the priors, then we can be fairly confident that the data have overwhelmed whatever information we inadvertently put into the priors.

12.3 MCMC Methods

In phylogenetics, as in many other modern applications of Bayesian inference, the computation of the posterior distribution from equation (12.2) is not done directly. Attempting to evaluate the denominator in that formula would be computationally intractable, since it requires summing terms for every possible choice of parameters. The tree topology parameter alone has an enormous number of possible values if the number of taxa is at all large, and then there are many numerical parameters as well.

Instead a different computationally-intensive process, called a *Markov chain Monte Carlo* (MCMC) method, is followed that rather than giving exact values of the posterior distribution attempts to give a sample chosen from that distribution. We say 'attempts' here, since the process that is followed will, provided it runs long enough, provably converge to a process that gives such a

sample. Thus the process is first run for a 'burn in' period of many iterations, in order to get from its starting point to a point more typical of the asymptotic behavior. There are no useful theoretical results on how long the burn-in period must run before it gives a sample that approximates one from the true distribution, but there are heuristic guides or diagnostics that are provided by software developers. The sample produced during burn in are then ignored, and as the process continues to run a new large sample is collected. Provided it is large enough, this new sample should closely approximate the true distribution of the posterior.

The name 'Markov chain Monte Carlo' signifies the two main ideas behind this process. The reference to the casinos of Monte Carlo indicates that it a random process is followed, and so only by taking a large sample can we be reasonably confident that we have a good approximation to the posterior. The Markov chain refers to the more theoretical underpinnings of the process, in which there is a collection of states, and conditional probabilities indicating how we move from state to state.

In the phylogenetics applications, the *Metropolis-Hastings* algorithm is the MCMC method usually used. The states for the process are a choice of parameters, *i.e.* a tree together with specifications of all other numerical parameters. We start at one such state, and then move to a new one by a certain probabilistic rule. An overview of the slightly simpler Metropolis procedure is the following:

1. Create some probabilistic *proposal process* that given any state p_1 picks a new state p_2 with probability given by some function $\mathcal{P}(p_2 \mid p_1)$. The only requirements on this process is that $\mathcal{P}(p_2 \mid p_1) = \mathcal{P}(p_1 \mid p_2)$, so that the probability of jumping from one state to another is symmetric, and that the probability of proposing any state from any other (perhaps by a succession of proposals) is non-zero.

2. Choose some starting state p_0.

3. Repeat the following steps:

   (a) If the current state is p_n, propose a new state p according to the proposal process.

   (b) Compute the acceptance ratio

   $$\alpha = \frac{\mathcal{P}(p \mid \text{data})}{\mathcal{P}(p_n \mid \text{data})}.$$

   (c) If $\alpha \geq 1$, accept the proposal and make the current state $p_{n+1} = p$, and return to step 3a.

   (d) If $\alpha < 1$, then flip a coin that is weighted to give H with probability α, and T with probability $1 - \alpha$.

   (e) If the coin gives H, then accept the proposal by making the current state $p_{n+1} = p$. If the coin gives T, then reject the proposal by making the current state $p_{n+1} = p_n$.

(f) Return to step 3a.

Informally, the acceptance of a proposed state is done in such a way that we will always accept one that has higher probability in the posterior than our current state. However, we will also accept ones that are less probable, but not always. Thus it is plausible that the process could tend toward sampling from the posterior. This in fact can be proved.

Actually, the formula for α above makes it appear that we must already know the posterior to use the algorithm. In fact, the key point is that we do *not* need to know it, since the posterior only appears in a ratio. While the denominator in the formula (12.2) for the posterior is what is intractable to compute, it cancels out in the formula for α:

$$\alpha = \frac{\mathcal{P}(p \mid \text{data})}{\mathcal{P}(p_n \mid \text{data})} = \frac{\frac{\mathcal{P}(\text{data} \mid p)\mathcal{P}(p)}{\sum_p \mathcal{P}(\text{data} \mid p)\mathcal{P}(p)}}{\frac{\mathcal{P}(\text{data} \mid p_n)\mathcal{P}(p_n)}{\sum_p \mathcal{P}(\text{data} \mid p)\mathcal{P}(p)}} = \frac{\mathcal{P}(\text{data} \mid p)\mathcal{P}(p)}{\mathcal{P}(\text{data} \mid p_n)\mathcal{P}(p_n)}.$$

Thus computing the acceptance ratio depends only on being able to compute the likelihood function, and the prior.

Notice that almost any proposal function can be used with the same theoretical guarantee of eventually approximating the posterior distribution. The Hastings modification of the algorithm even allows a non-symmetric proposal. In practice, however, the design of a good proposal process can have a large effect on how the algorithm performs. For instance, considering only tree topologies, we could imagine proposal processes that only used NNI moves, or instead took larger steps through tree space. If the posterior turns out to be highly concentrated on a single topological tree, then using NNI moves alone may work well after the burn-in period has passed, but might result in a longer burn-in as small steps are taken from the starting tree to get to the one with high posterior probability. On the other hand, if large steps are taken, we should expect that once we are past burn-in that most proposed trees will be rejected, resulting in a longer run time to build up a good sample.

12.4 Summarizing Posterior Distributions

A great strength of Bayesian inference is that it yields a distribution of estimates rather than a single one. In simple circumstances, such as the coin toss example in section 12.1, the posterior can be communicated by a graph, as in Figure 12.1. If the area under this graph is closely concentrated above a small interval on the horizontal axis, then a glance indicates there is strong support for the parameter value being in that interval. If the area is more spread out, then we quickly see there are a large range of reasonable parameter values, and the data were not able to give us a tight conclusion.

In the phylogenetic setting, however, where a parameter choice may mean a metric tree together with a base distribution, a rate matrix, and possibly several

other numerical parameters, things are a little more complicated. Rather than a single parameter which could be placed on a single axis, we have a rather large number of parameters. Moreover, the tree topology is also a parameter, and there is no natural way to depict that on an axis in a graph.

When faced with difficulties in presenting the full posterior distribution, one alternative is to report the single choice of parameters (assuming there is only one) that maximizes the posterior. This *maximum a posteriori* (MAP) estimate is the most probable choice of parameters given the data and prior. Reporting only it, however, throws away most of the information in the posterior distribution. Most importantly, we lose an indication of the spread of the distribution across other parameter values.

For a distribution of phylogenetic trees, a reasonable approach is to report the MAP tree topology, but then further indicate the support for each split in the unrooted tree case, or for each clade if the tree is rooted, in the full distribution. Each tree topology has a associated posterior probability (obtained by restricting the distribution to trees of the given topology and then integrating over all numerical parameter values). The probability of a split or clade is then the sum of the probabilities of the topological trees displaying that split or clade. Thus a probability of 1 indicates the split or clade appears in every tree that appears with non-zero probability in the posterior, while lower probabilities indicate appearance in a lower proportion of the trees. Thus each edge of the reported tree can be assigned the probability of the associated split, indicating its support across the full posterior distribution.

Instead of the MAP tree, one could instead use a consensus tree for the splits or clades. Here the posterior probabilities of each possible split or clade in the full distribution would be used to determine which splits or clades make the cutoffs for building the consensus tree. A strict consensus thus includes only those with probability 1 in the full distribution, while a majority-rule consensus uses all those with probabilities strictly greater than 0.5.

Of course one hopes that a Bayesian phylogenetic analysis will lead to very strong support for a single tree. If the MAP tree topology has probability 1, then the posterior is entirely concentrated on a single tree, and any sort of a consensus tree will have the same topology. Even if the MAP tree topology only has probability slight larger than .5, the majority rule consensus tree will agree with it.

To give metric information and other numerical parameters for the reported tree there are also several possibilities. On the MAP tree topology one could simply report MAP edge lengths, by either fixing the MAP topology and then choosing the collection of edge lengths that simultaneously maximize the restricted posterior distribution, or alternatively, for each edge integrate the restricted distribution over all other numerical parameters and then report edge lengths maximizing this marginal distribution. For a consensus tree one can report averages (weighted by the probabilities of the posterior distribution) of edge lengths across the different trees displaying the associated split or clade. If the posterior is highly concentrated, there should be very little difference in

these approaches.

Of course not every analysis leads to a highly concentrated posterior distribution. In such a case rather than depending on some simple summaries, it is wise to look carefully at the more detailed distribution, which software will typically make available for examination.

12.5 Exercises

1. a) In example (1) of the computation of the posterior distribution in section 12.1, the data were taken to be HHT. Redo the calculation using the same uniform prior on 1/4, 1/2, 3/4, but assuming the data are a sequence of 300 coin tosses, with 200 heads and 100 tails. Compare the posterior distribution you obtain to the one in the example. Are they the same? If not, explain why any differences you see are reasonable.

   b) Repeat part (a), but for example (2), with a uniform prior on the interval $0 \leq p \leq 1$.

2. Suppose a prior is such that for some specific parameter value p_0, $\mathcal{P}(p = p_0) = 0$. From equation (12.2) explain why $\mathcal{P}(p = p_0 \mid \text{data})$, the posterior probability that $p = p_0$ will also be 0. (Thus in a Bayesian framework, if something is viewed as impossible, then no amount of data will change that view.)

3. Suppose a prior is such that for some specific parameter value p_0, $\mathcal{P}(p = p_0) = 1$. From equation (12.2) explain why $\mathcal{P}(p = p_0 \mid \text{data})$, the posterior probability that $p = p_0$ will also be 1. (Thus in a Bayesian framework, if something is viewed as definite, then no amount of data will change that view.)

4. In the continuous-parameter Example(2) of Section 12.1, the posterior is shown to be $\mathcal{P}(p \mid \text{HHT}) = 12p^2(1-p)$.

   a) For what values of p is this posterior 0? Why is this reasonable for these data?

   b) Calculate the MAP estimate of p. How does it compare to the ML estimate of p? (Note: this relationship depended on the specific choice of prior that was used.)

5. The *Dirichlet distribution* for $(x_1, x_2, \ldots, x_k)$ with $x_1 + x_2 + \cdots + x_k = 1$, $x_i \geq 0$, is specified by the probability density function

$$f(x_1, x_2, \ldots, x_k; \alpha_1, \ldots, \alpha_k) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{k} x_i^{\alpha_i - 1},$$

   where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$ is a vector of parameters and

$$B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{k} \alpha_i)}$$

is a normalizing constant so that the total probability is 1.

a) Show that if $\alpha_1 = \cdots = \alpha_k = 1$ this is a uniform distribution, and thus when $k = 4$ gives the uninformative prior for base distributions.

b) To better understand the distribution for other values of $\boldsymbol{\alpha}$, consider $k = 3$. Produce 3-dimensional plots of the unnormalized version of the distribution, $f(x_1, x_2, x_3) = x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1}(1 - x_1 - x_2)^{\alpha_3 - 1}$, for $\boldsymbol{\alpha} = (1, 1, 1)$, $(10, 10, 10)$, $(100, 100, 100)$. Roughly give the location of the peak, and explain how the concentration of probability for the Dirichlet distribution changes for these values of $\boldsymbol{\alpha}$. (Note: the only relevant portion of the graph is where $x_1 + x_2 \leq 1$.)

c) Produce 3-dimensional plots of the unnormalized version of the distribution, $f(x_1, x_2, x_3) = x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1}(1 - x_1 - x_2)^{\alpha_3 - 1}$, for $\boldsymbol{\alpha} = (1, 2, 3)$, $(10, 20, 30)$, $(100, 200, 300)$. Roughly give the location of the peak, and explain how the concentration of probability for the Dirichlet distribution changes for these values of $\boldsymbol{\alpha}$.

d) Analytically determine the location of the maximum of $f(x_1, x_2, \ldots, x_k; \boldsymbol{\alpha})$ by finding the maximum of its logarithm. (In taking derivatives, you will need to let $x_k = 1 - x_1 - \cdots - x_{k-1}$ and treat the logarithm as a function of $x_1, x_2, \ldots, x_{k-1}$ alone.)

6. Suppose you collect a data set D_1 and using a prior $\mathcal{P}(p)$ you compute the posterior distribution $\mathcal{P}(p \mid D_1)$ according to equation (12.2). You then collect a second data set D_2, and using $\mathcal{P}(p \mid D_1)$ as a prior for it, you compute a second posterior distribution. Under the assumption that D_1 and D_2 are independent regardless of the value of p, show that this two-step analysis gives exactly the same posterior as would a single-step one to find $\mathcal{P}(p \mid D_1 \text{ and } D_2)$.