Supplementary material for MVE550 Stochastic processes and Bayesian inference Autumn 2019

Petter Mostad

December 2, 2019

The course "MVE550 Stochastic processes and Bayesian inference" is mostly about stochastic processes. When introducing the course in 2018, we wanted it to also have an element of inference, i.e., theory about how one can find stochastic process models appropriate for given data. Further, we wanted this inference to use a Bayesian framework. However, we could not find existing teaching material that perfectly fitted our plans. The solution was to use "Dobrow: Introduction to Stochastic Processes with R" as the main textbook, but to supplement it with some additional material, contained in these lecture notes.

Petter Mostad

Autumn 2019

Contents

Chapter 1

Basics of Bayesian inference

This course is mostly about stochastic processes. Such processes can function as models for many real phenomena where some uncertainty is involved. In the simplest cases, we can set up a precise stochastic model based only on reasonable assumptions, and then go on to make predictions from these models. For example, when throwing a dice, it is reasonable to assume that the probaility of obtaining each of the outcomes 1 through 6 is 1/6. From this we can compute such things as the probability of obtaining a total of 9 in the first three throws, or the expected wait until we get 3 consecutive sixes. Similarly, using a deck of 52 playing cards, it is reasonable to assume that each draw from it is independent, and we can compute such things as the probaility of being dealt a straight flush.

However, for most potential applications of stochastic processes, and of mathematical statistics in general, the situation is more complex. We cannot make predictions of future observations based only on reasonable assumptions, we must also use earlier observations, *data*, to find a reasonable stochastic model. Then we can make predictions from this model. For example, if we want to predict the range of an electric car on full batteries, we could use data for the ranges of similar cars. Using this data, we would build a model for the range of the car in question, and we could then use the model to make predictions. A simple model in this situation could be a normal distribution, with *parameters* μ and σ, representing the expectation and standard deviation of the range.

Building a stochastic model using data can be called *inference*. There are two quite common ways of thinking, or *paradigms*. One is the *classical* or *frequentist* approach. In this approach, we start with building a stochastic model which can be used for making the predictions we want, defining the model in terms of some *parameters* which are regarded as *unknown*. We then use the data to *estimate* these parameters. Finally, plugging in these estimates in the model, we can use it to make the predictions we want.

The alternative approach is *Bayesian inference*. In this approach, we build a stochastic model using only general reasonable assumptions, but we include in the model random variables representing both the data we have observed and the future observations we want to predict. We then compute the conditional

distribution for the future observations given that the variables representing data are fixed to the observed values. This conditional distribution is used for prediction. So we start with a stochastic model, we *update* it using the observations from the data (in a way we *learn* from the data), and then we use the updated model for prediction.

An example may make this clearer. Consider repeated throws of a six-sided dice. If you know that the dice is fair, all sequences of equal length of outcomes will have the same probability: Observing 1,1,1 will have the same probability as observing 2,4,1, namely $(1/6)^3 = 0.00463$. But what if you suspect the dice is not fair? In practice, you would throw the dice a number of times, and if one outcome appears more often than other outcomes, you might start to suspect that the dice is loaded in favour of this outcome. This would increase your belief that this outcome would appear again, in your next throw. So if you ave thrown 1,2,3,1,1,4,1,2,1,1,5, continuing this sequence with a 1 would seem more probable than continuing it with a 6. Even a short sequence like 1,1,1 would be slightly more probable than a sequence like 2,4,1. In other words, if you suspect a loaded dice, the outcomes are no longer independent.

Using a Bayesian approach, one could here set up a model for how the dice could be loaded, which would yield predicted probabilities for any sequence of observed outcomes. Let $p_1, p_2, \ldots, p_6$ denote the probabilities of observing six sequences that are identical except for the last throw, and end with the outcomes $1, 2, \ldots, 6$, respectively. If you have observed the throws these sequences have in common, the conditional probabilities for the next throw to be either of $1, 2, \ldots, 6$ given the observed outcomes would be $p_1/(p_1 + \cdots + p_6), \ldots, p_6/(p_1 + \cdots + p_6)$, respectively.

1.1 Tossing a coin

Let us explore the idea above in more detail in a slightly simpler setting: A loaded coin has probability 0.7 for either H (heads) or T (tails), but you do not know which. In fact, you think there is an equal probability that it is loaded in favour of heads or tails. The probability for observing k heads in n throws is now

$$\Pr(k) = 0.5 \cdot \text{Binomial}(k; n, 0.7) + 0.5 \cdot \text{Binomial}(k; n, 0.3) \tag{1.1}$$

Here we write, for example, $\text{Binomial}(k; n, 0.7)$ for the value at k of the Binomial probability mass function with parmeters n and 0.7. We obtain the formula above by conditioning on whether the coin is loaded in favour of either H or T: The probability of each of these possibilities is 0.5, and, given each choice, the probabilities of k heads in n throws is either $\text{Binomial}(k; n, 0.7)$ or $\text{Binomial}(k; n, 0.3)$, respectively.

Figure 1.1 illustrates a particular sequence of observations of heads (represented as crosses at 1) and tails (represented as crosses at 0). Together with each observation, we also plot the probability, *before* this observation is made but *given* all the previous observations, of observing heads. Before any observations are made, the probabilities of observing either heads of tails is 0.5,

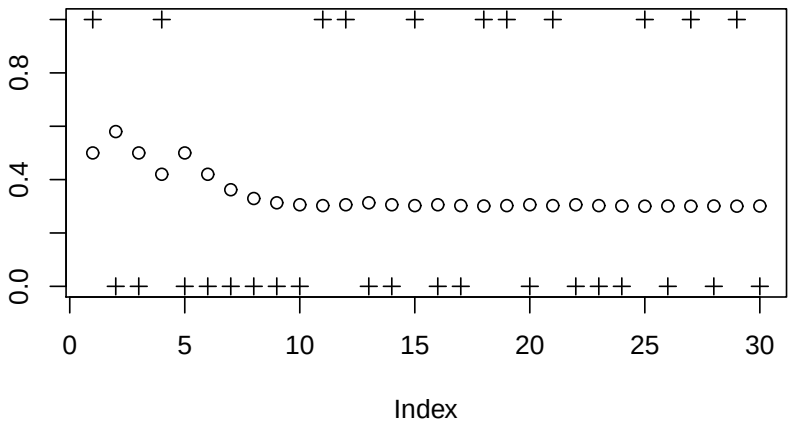

Figure 1.1: The probability of heads at each point in a sequence of observations, conditioning on the previous observations of heads and tails. The prior used is $\theta = 0.7^x \cdot 0.3^{1-x}$ where $x \sim \text{Bernoulli}(0.5)$.

because of the symmetry of the situation. Once we start making observations, the prediction for the next observation will jump up and down a bit, depending on those observations. However, after a while, it settles close to 0.3: we then have such a substantial overweight of tails in the data that it seems reasonable to believe that the coin is loaded towards tails. In a way, we have *learned* from the data that the coin is loaded this way.

The probabilities displayed in Figure 1.1 may be computed as follows: The conditional probability of heads after observing a specific sequence of heads and tails is equal to the probability of observing this sequence, followed by observing heads, divided by the probability of observing the sequence. Thus, if the sequence contains y_H heads and y_T tails, the probability is given by

$$\Pr(\text{heads} \mid \text{sequence}) = \frac{0.5 \cdot 0.7^{y_H+1}0.3^{y_T} + 0.5 \cdot 0.3^{y_H+1}0.7^{y_T}}{0.5 \cdot 0.7^{y_H}0.3^{y_T} + 0.5 \cdot 0.3^{y_H}0.7^{y_T}} \tag{1.2}$$

Here, we have used that the probability of observing a specific sequence with y_H heads and y_T tails is $\theta^{y_H}(1-\theta)^{y_T}$, where θ is the probability of observing heads each time. So we have implicitly used the order of the observations of heads and tails as part of the data. However, we get the same results by assuming that the data only contains the counts y_H and y_T, see Exercise 4 of Section 1.7.

To sum up, Bayesian inference is done with the following steps:

1. Based on reasonable assumptions, create a stochastic model containing

random variables representing observed data and observations you would like to predict.

2. Use for prediction the conditional distribution of prediction variables given that the the data variables are fixed to their observed values.

A frequentist approach in the situation above might focus on estimating whether the coin is loaded towards heads or towards tails. With n low, one might conclude that there is not enough information to reliably make an estimate, and continue to predict heads in the next throw with probability 0.5. Then, as n increases, one would reach a point where one would conclude that one could reliably estimate the direction of the bias. From that point on, the predicted probability for heads would be exactly 0.7 or 0.3, depending on the estimate.

Although the Bayesian approach above is not formulated in terms of estimating the direction of the bias, we may reformulate it as stepwise learning about a stochastic variable θ representing the true probability of heads. Thus in our current setup θ has the two possible values 0.7 and 0.3, with some probability for each. We may reformulate the model as encompassing two dependent random variables, θ and k (the count of heads after n trials):

$$\begin{aligned} x &\sim \text{Bernoulli}(0.5) \\ \theta &= 0.7^x \cdot 0.3^{1-x} \\ k \mid \theta &\sim \text{Binomial}(n, \theta). \end{aligned}$$

Here, we write Bernoulli(0.5) for a random variable which has value 1 with probability 0.5 and otherwise value 0. We also write Binomial(n, θ) for a random variable which is Binomially distributed with parameters n and θ. We write $x \sim$ Bernoulli(0.5) and $k \mid \theta \sim$ Binomial(n, θ) to indicate that the random variables x and $k \mid \theta$ have the given distributions.

Now, let y be 1 or 0 depending on whether the $n+1$'st throw is heads or not. Using standard probability theory formulas, the conditional distribution of y given the count k of heads in the n first throws can be written

$$\begin{aligned} \Pr(y \mid k) &= \sum_{\theta=0.3,0.7} \Pr(y, \theta \mid k) = \sum_{\theta=0.3,0.7} \Pr(y \mid \theta, k) \Pr(\theta \mid k) \\ &= \sum_{\theta=0.3,0.7} \Pr(y \mid \theta) \Pr(\theta \mid k). \end{aligned} \tag{1.3}$$

The key step above is that we can write $\Pr(y \mid \theta, k) = \Pr(y \mid \theta)$. This is because once we know the value of θ, the probability that y is 1 does not depend on k: In fact, the probability that $y = 1$ is exactly θ. According to the computation above, we can compute the probability $\Pr(y = 1 \mid k)$ by first computing the two probabilities $\Pr(\theta = 0.3 \mid k)$ and $\Pr(\theta = 0.7 \mid k)$, and then multiplying these as weights with the probabilities $\Pr(y = 1 \mid \theta = 0.3) = 0.3$ and $\Pr(y = 1 \mid \theta = 0.7) = 0.7$.

At this point, we introduce a generic notation for probability mass functions which may also be used for probability density functions: For example, we

write $\pi(k)$ instead of $\Pr(k)$ and $\pi(y \mid \theta, k)$ instead of $\Pr(y \mid \theta, k)$. If $z \sim$ Exponential(λ) so that z is a continuous random variable with an Exponential distribution with parameter λ, we also write $\pi(z) = \lambda \exp(-\lambda z)$ for the density function. This generic notation is helpful, as so many probability computations are the same whether the underlying functions are probability mass functions or probaility density functions.

The distribution $\pi(\theta \mid k) = \Pr(\theta \mid k)$ is called the *posterior* for θ. The unconditional distribution for θ, in which θ has probability 0.5 for both 0.7 and 0.3 in our case, is called the *prior*. We may compute the posterior using *Bayes formula*:

$$\pi(\theta \mid k) = \frac{\pi(k \mid \theta)\pi(\theta)}{\pi(k)} = \frac{\pi(k \mid \theta)\pi(\theta)}{\sum_\theta \pi(k \mid \theta)\pi(\theta)}$$

That Bayes formula appears in our approach is the reason why we call it Bayesian inference. In our case, we get

$$\begin{aligned}
\pi(\theta \mid k) &= \frac{\pi(k \mid \theta)\pi(\theta)}{\sum_\theta \pi(k \mid \theta)\pi(\theta)} \\
&= \frac{\text{Binomial}(k; n, \theta) \cdot 0.5}{\text{Binomial}(k; n, 0.3) \cdot 0.5 + \text{Binomial}(k; n, 0.7) \cdot 0.5} \\
&= \frac{\binom{n}{k}\theta^k(1-\theta)^{n-k}}{\binom{n}{k}0.3^k 0.7^{n-k} + \binom{n}{k}0.7^k 0.3^{n-k}} = \frac{\theta^k(1-\theta)^{n-k}}{0.3^k 0.7^{n-k} + 0.7^k 0.3^{n-k}}
\end{aligned}$$

Using Equation 1.3 we now get

$$\pi(y \mid k) = 0.3\frac{0.3^k 0.7^{n-k}}{0.3^k 0.7^{n-k} + 0.7^k 0.3^{n-k}} + 0.7\frac{0.7^k 0.3^{n-k}}{0.3^k 0.7^{n-k} + 0.7^k 0.3^{n-k}}$$

Comparing with Equation 1.2, we see that we have arrived at exactly the same result as we got there.

1.2 The Beta and Binomial distributions

Above, we made the rather curious assumption that θ was either equal to 0.3 or to 0.7. A more realistic assumption is that θ is just some real number between 0 and 1. Specifically, let us now assume that θ has as prior the uniform distribution on the interval $[0, 1]$, so that $\pi(\theta) = 1$.

Assume further that k out of n observations are heads. Even if θ is now a continuous variable instead of a discrete one, we can still compute the posterior

using Bayes formula, we just need to use an integral instead of a sum:

$$\begin{aligned}
\pi(\theta \mid k) &= \frac{\pi(k \mid \theta)\pi(\theta)}{\pi(k)} \\
&= \frac{\pi(k \mid \theta)\pi(\theta)}{\int_\theta \pi(k,\theta)\,d\theta} \\
&= \frac{\pi(k \mid \theta)\pi(\theta)}{\int_\theta \pi(k \mid \theta)\pi(\theta)\,d\theta} \\
&= \frac{\text{Binomial}(k;n,\theta)}{\int_\theta \text{Binomial}(k;n,\theta)\,d\theta} \\
&= \frac{\binom{n}{k}\theta^k(1-\theta)^{n-k}}{\int_\theta \binom{n}{k}\theta^k(1-\theta)^{n-k}\,d\theta} \\
&= \frac{\theta^k(1-\theta)^{n-k}}{\int_\theta \theta^k(1-\theta)^{n-k}\,d\theta}
\end{aligned} \tag{1.4}$$

To go on, we might compute the integral in the denominator. To do so, we may use a shortcut, looking up the density of for the Beta distribution. In fact $x \in [0,1]$ has a Beta distribution with parameters $\alpha > 0$ and $\beta > 0$ if its density is

$$\pi(x \mid \alpha, \beta) = \frac{1}{\text{B}(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

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

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

where the *Gamma function* $\Gamma(t)$ in turn is defined for $t > 0$ by

$$\Gamma(t) = \int_0^\infty x^{t-1}\exp(-x)\,dx.$$

Right now, the important thing for us is that the Beta density integrates to 1, so that, for all $\alpha > 0$ and $\beta > 0$,

$$\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}\,d\theta = \text{B}(\alpha,\beta).$$

Plugging this into the computations above by setting $\alpha - 1 = k$ and $\beta - 1 = n - k$, we get

$$\pi(\theta \mid k) = \frac{\theta^k(1-\theta)^{n-k}}{B(k+1, n-k+1)}$$

However, we can now use the definition of the Beta density again, to recognize that the posterior density $\pi(\theta \mid k)$ is in fact a Beta density, specifically a Beta density with parameters $k+1$ and $n-k+1$.

Our goal is to find the probability of heads in the $n+1$'st throw assuming that k out of the n first throws were heads. Similar to above, we may compute

$$\begin{aligned}\pi(y=1 \mid k) &= \int_\theta \pi(y=1,\theta \mid k)\,d\theta \\ &= \int_\theta \pi(y=1 \mid \theta,k)\pi(\theta \mid k)\,d\theta \\ &= \int_\theta \pi(y=1 \mid \theta)\pi(\theta \mid k)\,d\theta \\ &= \int_\theta \theta\pi(\theta \mid k)\,d\theta\end{aligned} \tag{1.5}$$

We could compute this integral. But we may also notice that the integral is the expectation of the posterior distribution $\theta \mid k$. Having derived that this posterior is the Beta$(k+1, n-k+1)$ distribution, we can look up its expectation, finding that it is $(k+1)/(k+1+n-k+1) = (k+1)/(n+2)$, so that

$$\pi(y=1 \mid k) = \frac{k+1}{n+2}.$$

In Figure 1.2 we show, as in the previous section, a sequence of observed heads and tails, together with the probability of observing heads given all *previous* observations. The difference compared to Figure 1.1 is that we now use the prior Uniform$(0,1)$ for θ, while in Figure 1.1 we use a prior where θ is equal to 0.3 or 0.7 with equal probability. Thus the result does not stabilize as easily. But it does seem to stabilize eventually.

Notice how we repeatedely took advantage of knowledge about the Beta distribution in the computations above. This simplified our computations, but, in fact, we may go one step further in simplification: As we are computing a posterior density for θ, we know that whatever we compute will always integrate to 1 when we integrate it as a function of θ over the possible values for θ, i.e., the interval $[0,1]$. Thus, there is no loss of information if we multiply or divide by factors that do not depend on θ. These factors can always be reinstated in the end, by using the requirement that our density must integrate to 1. To take advantage of this idea, we define the symbol $\propto_\theta$ (read "proportional to") to mean that two expressions are identical up to a factor not involving θ. Thus we can write for example $\theta \propto_\theta 3\theta$ and $\theta/(1+\theta) \propto_\theta \alpha\theta/(1+\theta)$. Using this notation, the computations of Equation 1.4 can be written

$$\pi(\theta \mid k) \propto_\theta \pi(k \mid \theta)\pi(\theta) = \text{Binomial}(k; n, \theta) \propto_\theta \theta^k(1-\theta)^{n-k}. \tag{1.6}$$

By comparing with the density for a Beta distribution, we see that the posterior $\pi(\theta \mid k)$ must necessarily be a Beta$(k+1, n-k+1)$ density. This trick of removing uninteresting factors until we need them is going to be used repeatedly in the rest of this text.

Above, we assume a Uniform$(0,1)$ prior for θ. However, the computation of Equation 1.6 suggests that, as long as the prior has the form $\theta^{\text{something}}(1-$

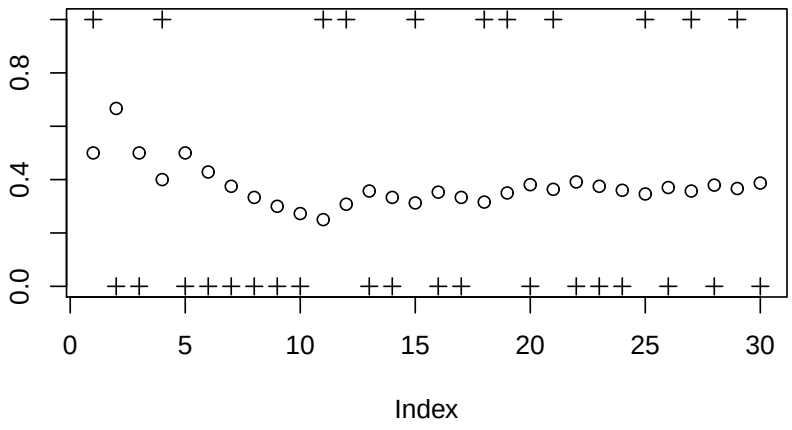

Figure 1.2: The probability of heads at each point in a sequence of observations, conditioning on the previous observations of heads and tails. The prior used is Uniform$(0, 1)$.

$\theta)^{\text{something else}}$, we will get a posterior that has the form of a Beta density. So, specifically, we now assume that θ has a Beta(α, β) prior for some parameters $\alpha > 0$ and $\beta > 0$. The computation of the posterior density now becomes

$$\pi(\theta \mid k) \propto_\theta \pi(k \mid \theta)\pi(\theta) \propto_\theta \theta^k(1-\theta)^{n-k}\theta^{\alpha-1}(1-\theta)^{\beta-1} = \theta^{\alpha+k-1}(1-\theta)^{\beta+n-k-1}$$

from which we can read off that the posterior $\theta \mid k$ is a Beta$(\alpha + k, \beta + n - k)$ distribution. The computations of Equation 1.5 apply unchanged to our more general situation. Referring again to what we know about the expectation of the Beta distribution, we get

$$\pi(y = 1 \mid k) = \frac{\alpha + k}{\alpha + \beta + n}$$

We now have the possibility to do slightly more realistic learning about the biasedness of the coin. Neither guessing that θ is either 0.3 or 0.7 or assuming that it can be any number between 0 and 1 with equal probability seems very realistic. Rather, one might guess that the coin is not too far away from fair, but it might be slightly unfair, i.e., θ might for example most likely be in the interval from 0.4 to 0.6. Selecting a Beta density that is symmetric and has approximately 90% of its probability mass in this interval (see Exercise 5 of Section 1.7), we find that we may use the prior Beta$(33.4, 33.4)$. Figure 1.3 is comparable to Figures 1.1 and 1.2, but now using this new prior. Because we

have now put much more information into the prior, the prediction probabilities are much more stable from the start. Another way to illustrate what is going on is with Figure 1.4. It shows the prior density Beta$(33.4, 33.4)$ and the posterior density Beta$(33.4 + 11, 33.4 + 19)$ after considering all of the 30 observations illustated in Figure 1.3. Notice how the posterior is slightly narrower than the prior, as it is based on more information and thus represents less uncertainty. It is also slightly shifted to the left, as there are 11 heads and 19 tails in the data.

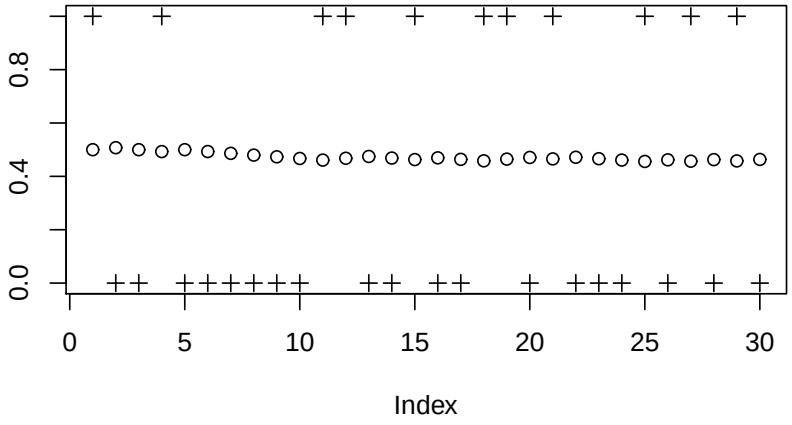

Figure 1.3: The probability of heads at each point in a sequence of observations, conditioning on the previous observations of heads and tails. The prior used is Beta$(33.4, 33.4)$.

1.3 The Poisson Gamma conjugacy

In the previous section, we saw that when considering data with a Binomial(n, θ) likelihood and using a prior for θ with a Beta distribution, the posterior for θ also became a Beta distribution. This kind of situation is in fact quite common in basic Bayesian inference. We say that a family of distributions is *conjugate* to a likelihood if selecting the prior in the family leads to a posterior in the same family. Thus the Beta family of distributions is conjugate to the Binomial likelihood, when considering the probability θ as the parameter. In this section, we will look at the Poisson Gamma conjugacy. Note that the Appendix in Chapter 8 contains an overview of several examples of conjugacy.

Assume you are monitoring the number of incoming requests for data to

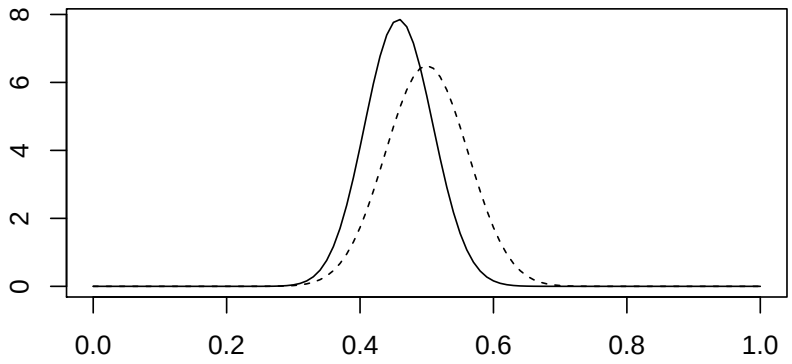

Figure 1.4: The the prior and posterior probability for θ when the prior Beta$(33.4, 33.4)$ is used. The prior is the dotted line.

some internet database. Assuming these requests happen independently, we see in Chapter 6 of Dobrow how it may be reasonable to model the number of requests per time unit as Poisson distributed. In other words, if k is the number of requests per time unit, the probability mass function is

$$\pi(k \mid \theta) = e^{-\theta}\frac{\theta^k}{k!},$$

where θ is the expected number of requests for this time unit. Assume you count the number of such requests for successive time units. After a number of such counts, you want to predict the count for the coming time unit.

The situation is quite similar to the one in the previous section, and indeed, we can use much of the same thinking. As successive counts are independent if we know the true value of θ, we can handle the situation as follows: We set up a prior density $\pi(\theta)$ for θ, representing our knowledge about the expected count before we have made any actual counts. Then, we find a way to compute the posterior density $\pi(\theta \mid k_1)$ given an observed count k_1. Notice that, if we then make another count, k_2, we can update our knowledge about θ again, but now using $\pi(\theta \mid k_1)$ as the prior and obtaining a posterior density $\pi(\theta \mid k_1, k_2)$. Continuing like this for some counts, we obtain a final posterior $\pi(\theta \mid k_1, \ldots, k_n)$. We can then use this posterior to make a prediction for the next count k_{n+1}, using computations like those in Equation 1.5:

$$\pi(k_{n+1} \mid k_1, \ldots, k_n) = \int_\theta \pi(k_{n+1} \mid \theta)\pi(\theta \mid k_1, \ldots, k_n)\, d\theta. \tag{1.7}$$

To make computations in practice, we need to decide on a prior $\pi(\theta)$ to start with. In the previous example, we used the uniform distribution on the interval $[0, 1]$ to indicate starting out with "no knowledge". In our current situation, θ can be any positive number. One possibility might be to try to use a uniform distribution on the interval $[0, \infty)$. Notice that there is no such thing, as the integral of any positive function constant on this interval is infinite. However, without going into the technical arguments here, it turns out that in Bayesian statistics, we may use "densities" that integrate to infinity: We call these densities "improper". This will work fine as long as the posterior density we compute is an ordinary "proper" density.

So we might use a "constant density" on $[0, \infty)$; we would denote this as $\pi(\theta) \propto_\theta 1$, for $\theta \geq 0$. However, this improper density may actually not correspond very well to "having no knowledge" about θ. In fact, it would for example appear to assign the same probability to the intervals $[1, 2]$ $[10, 11]$, and $[1000000, 1000001]$. A better representation of "no knowledge" about the parameter θ might be that the intervals $[1, 2]$, $[10, 20]$ and $[1000000, 2000000]$ have the same prior probability. It would then be reasonable to use the prior

$$\pi(\theta) \propto_\theta \frac{1}{\theta}.$$

Notice that this prior is also improper, as the integral of $1/\theta$ over $[0, \infty)$ is infinite.

Assuming we have observed a count k, we can now get, using Bayes formula,

$$\pi(\theta \mid k) \propto_\theta \pi(k \mid \theta)\pi(\theta) \propto_\theta e^{-\theta}\theta^k \cdot \frac{1}{\theta} = e^{-\theta}\theta^{k-1}$$

Using a similar trick as in the previous section, we look up the density for a Gamma distribution:

$$\text{Gamma}(\theta; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1} \exp\left(-\beta\theta\right).$$

Thus we see that we must have

$$\pi(\theta \mid k) = \text{Gamma}(\theta; k, 1).$$

Continuing with the script from the previous section, we can now try out using the prior

$$\pi(\theta) = \text{Gamma}(\theta; \alpha, \beta)$$

Computations with Bayes formula give

$$\pi(\theta \mid k) \propto \pi(k \mid \theta)\pi(\theta) \propto_\theta e^{-\theta}\theta^k\theta^{\alpha-1} \exp(-\beta\theta) = \theta^{\alpha+k-1} \exp(-(\beta + 1)\theta)$$

so the posterior is a $\text{Gamma}(\theta; \alpha + k, \beta + 1)$ distribution, and we have proved that the Gamma family of distributions is conjugate to the Poisson likelihood.

It is now easy to determine the effect of observing counts $k_1, k_2, \ldots, k_n$. Starting with the distribution $\text{Gamma}(\alpha, \beta)$ for θ, each time we observe a count

k_i we add k_i to the first parameter and 1 to the second parameter. Thus, after n repeated updates, we get a Gamma $(\alpha + \sum_{i=1}^n k_i, \beta + n)$ distribution. It is worth noticing that the two improper densities for θ that we considered at the start can fit into this framework: The prior $\pi(\theta) \propto_\theta 1$ could be named a "Gamma$(1, 0)$" density, resulting in a Gamma $(1 + \sum_{i=1}^n k_i, n)$ posterior. Similarly, the prior $\pi(\theta) \propto_\theta 1/\theta$ could be named a "Gamma$(0, 0)$" density, resulting in a Gamma $(\sum_{i=1}^n k_i, n)$ posterior.

We can also make the computation of Equation 1.7 explicit:

$$\begin{aligned} & \pi(k_{n+1} \mid k_1, \ldots, k_n) \qquad (1.8) \\ = & \int_\theta \pi(k_{n+1} \mid \theta)\pi(\theta \mid k_1, \ldots, k_n)\, d\theta \\ = & \int_\theta e^{-\theta} \frac{\theta^{k_{n+1}}}{k_{n+1}!} \frac{(\beta+n)^{\alpha+\sum_{i=1}^n k_i}}{\Gamma(\alpha + \sum_{i=1}^n k_i)} \theta^{\alpha+\sum_{i=1}^n k_i - 1} \exp\left(-(\beta+n)\theta\right)\, d\theta \\ = & \frac{(\beta+n)^{\alpha+\sum_{i=1}^n k_i}}{\Gamma(\alpha + \sum_{i=1}^n k_i) k_{n+1}!} \int_\theta \theta^{\alpha+\sum_{i=1}^{n+1} k_i - 1} \exp\left(-(\beta+n+1)\theta\right)\, d\theta \\ = & \frac{(\beta+n)^{\alpha+\sum_{i=1}^n k_i}}{\Gamma(\alpha + \sum_{i=1}^n k_i) k_{n+1}!} \cdot \frac{\Gamma(\alpha + \sum_{i=1}^{n+1} k_i)}{(\beta+n+1)^{\alpha+\sum_{i=1}^{n+1} k_i}}. \end{aligned}$$

In the last step, we have again compared with the density for a Gamma distribution to compute the integral.

Let us illustrate our results so far in a concrete example. We start with the prior $\pi(\theta) \propto_\theta 1/\theta$, and we then make the consequtive observations $k_1 = 20$, $k_2 = 24$, and $k_3 = 23$. The posteriors after one, two, and three observations are Gamma$(20, 1)$, Gamma$(44, 2)$, and Gamma$(67, 3)$, respectively. These posteriors are illustrated in Figure 1.5. We see that our knowledge about θ is increasing in each step, as the variances of the distributions are decreasing. Note that the expectations of the Gamma distributions are $20/1 = 20$, $44/2 = 22$, and $67/3 = 22.33$, respectively.

Figure 1.6 illustrates predictions we may make for the fourth observation k_4 after observing k_1, k_2, k_3. The pluses represent probabilities for various values of k_1 computed according to the formula for $\pi(k_4 \mid k_1, k_2, k_3)$ above. The circles represents a more classical prediction: Using the three observations k_1, k_2, k_3, the maximum likelihood estimate for the parameter θ is $67/3 = 22.33$. The circles plots a Poisson mass density function with parameter 22.33. We see that the predicted distribution is then narrower than the one using Equation 1.8. The reason is that with the classical prediction, we have "thrown away" our remaining uncertainty about θ: We believe that its value is exactly 22.33 instead of believing that its value is given by the posterior Gamma$(67, 3)$ depicted in Figure 1.5. In this sense, we can say that the frequentist model is *overfitted*.

The computations shown in Equation 1.8 may seem a bit messy. Let us close this section with showing how we can do such computations in a more structured way that make them simpler to follow. In general, if we are in a situation with conjugacy, so that all the densities $\pi(k \mid \theta)$, $\pi(\theta)$, and $\pi(\theta \mid k)$ are expressed in nice analytic formulas, the last relevant density, $\pi(k)$, can also

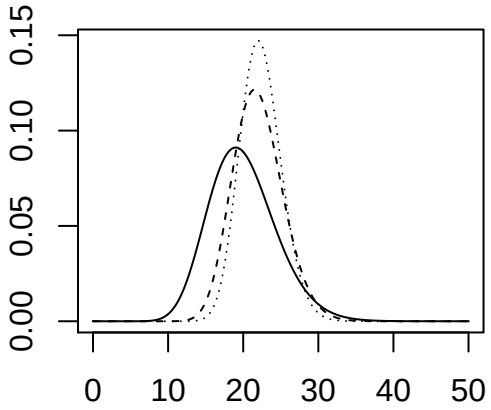

Figure 1.5: The three posterior distributions Gamma$(20, 1)$, Gamma$(44, 2)$, and Gamma$(67, 3)$.

be expressed in a nice analytic formula using the following formula, which can be derived immediately from Bayes rule:

$$\pi(k) = \frac{\pi(k \mid \theta)\pi(\theta)}{\pi(\theta \mid k)} \tag{1.9}$$

If all the densities on the right have nice formulas, we get a nice formula also for $\pi(k)$ on the left. Moreover, as the right-hand side contains θ while the left-hand side does not, we know that if we put in explicit formulas on the left-hand side, the θ must somehow disappear from the formula after simplifications.

Let us assume that $\pi(k \mid \theta) = \text{Poisson}(k; \theta)$, $\pi(\theta) = \text{Gamma}(\alpha, \beta)$, and $\pi(\theta \mid k) = \text{Gamma}(\alpha + k, \beta + 1)$. We then get

$$\begin{aligned} \pi(k) &= \frac{\pi(k \mid \theta)\pi(\theta)}{\pi(\theta \mid k)} \\ &= \frac{\text{Poisson}(k; \theta)\,\text{Gamma}(\theta; \alpha, \beta)}{\text{Gamma}(\theta; \alpha + k, \beta + 1)} \\ &= \frac{e^{-\theta}\frac{\theta^k}{k!} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}\exp(-\beta\theta)}{\frac{(\beta+1)^{\alpha+k}}{\Gamma(\alpha+k)}\theta^{\alpha+k-1}\exp(-\beta\theta - \theta)} \\ &= \frac{\beta^\alpha \Gamma(\alpha + k)}{(\beta + 1)^{\alpha+k}\Gamma(\alpha)k!} \end{aligned} \tag{1.10}$$

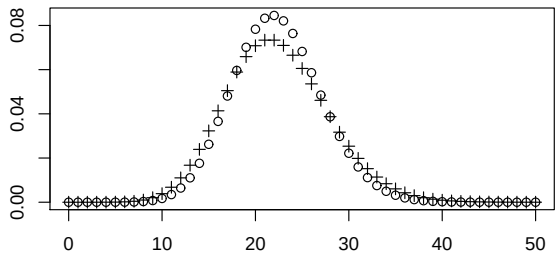

Figure 1.6: Two possible predictions for the fourth observation k_4: One derived in our Bayesian computations (shown with pluses) and one derived with a classical approach (shown in circles).

As we knew it should, the θ's disappeared from our computations, and we are left with what is called the *prior predictive* distribution for k, given the prior $\theta \sim \text{Gamma}(\alpha, \beta)$. If we instead start with a posterior, for example the posterior $\theta \sim \text{Gamma}(\alpha + \sum_{i=1}^{n} k_i, \beta + n)$, the formula above gives the *posterior predictive*. Replacing α with $\alpha + \sum_{i=1}^{n}$ and β with $\beta + n$ in Equation 1.10, we see that we get exactly the result of Equation 1.8.

The trick we just used to compute the predictive distribution can be used in all situations where you have conjugacy. Often the resulting predictive distribution turns out to be in a well-known family of distributions. In our case, the probability mass function for $\pi(k)$ found in Equation 1.10 is actually a Negative Binomial distribution: A stochastic variable x taking on as possible values any positive integer has a Negative Binomial distribution if its probability mass function is given by (NOTE: The definition below has now been changed to conform with the definition used for example in R):

$$\pi(x \mid r, p) = \binom{x + r - 1}{x} \cdot (1 - p)^x p^r = \frac{\Gamma(x + r)}{\Gamma(x + 1)\Gamma(r)} (1 - p)^x p^r \qquad (1.11)$$

where $r > 0$ and $p \in (0, 1)$ are parameters. Thus we see that the prior predictive density computed in Equation 1.10 is a Negative-Binomial$(\alpha, \beta/(1 + \beta))$ distribution. (NOTE: This formula has now bee updated to conform with the updated definition).

1.4 Summary so far

Bayesian inference can be summarized in the following way: Let y represent a vector of random variables which you have observed, and let y_{new} represent a vector of random variables you would like to predict. There are then two steps:

1. Based on reasonable assumptions, create a stochastic model relating y and y_{new}.
2. Make predictions for y_{new} using the conditional distribution $y_{new} \mid y$, where y is fixed to its observed values.

Most often, the stochastic model is formulated using an additional stochastic variable θ, a parameter or vector of parameters, so that y and y_{new} are conditionally independent given θ, i.e., for the densities,

$$\pi(y_{new} \mid \theta, y) = \pi(y_{new} \mid \theta).$$

Then,

$$\pi(y_{new} \mid y) = \int_\theta \pi(y_{new}, \theta \mid y)\, d\theta = \int_\theta \pi(y_{new} \mid \theta)\pi(\theta \mid y)\, d\theta \tag{1.12}$$

and computation of $\pi(y_{new} \mid y)$ is done by first computing $\pi(\theta \mid y)$ and then computing the integral above.

To find $\pi(\theta \mid y)$ one can generally formulate the joint distribution $\pi(\theta, y) = \pi(y \mid \theta)\pi(\theta)$ and then use Bayes theorem:

$$\pi(\theta \mid y) = \frac{\pi(y \mid \theta)\pi(\theta)}{\pi(y)} \propto_\theta \pi(y \mid \theta)\pi(\theta)$$

The distributions $\pi(\theta)$ and $\pi(\theta \mid y)$ are called the *prior* and *posterior*, respectively.

In many cases, the data y is a *random sample* $y = (y_1, \ldots, y_n)$, where the y_i are conditionally independent given the parameter θ, in other words,

$$\pi(y \mid \theta) = \prod_{i=1}^{n} \pi(y_i \mid \theta).$$

We get for the posterior

$$\pi(\theta \mid y) \propto_\theta \prod_{i=1}^{n} \pi(y_i \mid \theta)\pi(\theta),$$

and it can be obtained by stepwise updating the prior $\pi(\theta)$, using the data values y_i in any order, and using the posterior from one update as the prior for the next update.

A parametric family of probability distributions for a parameter θ is *conjugate* to a likelihood $\pi(y \mid \theta)$ if, when the prior is in the family, the posterior is also in the family. We have so far looked at two examples of conjugacy: The Beta-Binomial conjugacy and the Gamma-Poisson conjugacy. Whenever we have conjugacy, the *prior predictive density*

$$\pi(y) = \int_\theta \pi(y, \theta)\, d\theta = \int_\theta \pi(y \mid \theta)\pi(\theta)\, d\theta$$

has a simple closed form, which may be computed using Equation 1.9. The *posterior predictive density* of Equation 1.12 can also be computed from the same equation, replacing the prior $\pi(\theta)$ with the posterior $\pi(\theta \mid y)$. Explicitly,

$$\pi(y_{new} \mid y) = \frac{\pi(y_{new} \mid \theta)\pi(\theta \mid y)}{\pi(\theta \mid y_{new}, y)}.$$

1.5 Bayesian inference using discretization

We saw in the previous section that Bayesian inference requires computation of the posterior $\pi(\theta \mid y)$ and the predictive distribution $\pi(y_{new} \mid y)$ using (most often) Equation 1.12. However, in many practical applications, these computations cannot be done using conjugacy; there simply does not exist a conjugate prior to the likelihood $\pi(y \mid \theta)$ one would like to use. In this and the next section we look at some simple alternative computational approaches.

We first turn to discretization, which can be a very good alternative if θ has only one dimension (i.e., component) and $\pi(\theta)$ is positive only within some bounded interval. Assume $\theta_1, \theta_2, \ldots, \theta_k$ are equally-spaced values within this interval, so that the prior density $\pi(\theta)$ can reasonably be approximated by a discrete distribution on $\{\theta_1, \ldots, \theta_k\}$ specified by

$$a_i = \Pr(\theta = \theta_i) = \frac{\pi(\theta_i)}{\sum_{i=1}^{k} \pi(\theta_i)}.$$

Writing $b_i = \pi(y \mid \theta_i)$ for $i = 1, \ldots, k$ we can then approximate the posterior with a discrete distribution on $\{\theta_1, \ldots, \theta_k\}$ specified by

$$c_i = \Pr(\theta = \theta_i \mid y) = \frac{\pi(y \mid \theta = \theta_i)\Pr(\theta = \theta_i)}{\sum_{j=1}^{k} \pi(y \mid \theta = \theta_j)\Pr(\theta = \theta_j)} = \frac{a_i b_i}{\sum_{j=1}^{k} a_j b_j}$$

Finally, for a specific value of y_{new} we may approximate the predictive distribution as

$$\begin{aligned} \pi(y_{new} \mid y) &= \int_\theta \pi(y_{new} \mid \theta)\pi(\theta \mid y)\, d\theta \\ &\approx \sum_{i=1}^{k} \pi(y_{new} \mid \theta_i)\Pr(\theta = \theta_i \mid y)\; d\theta = \sum_{i=1}^{k} \pi(y_{new} \mid \theta_i) c_i. \end{aligned}$$

How good this approximation is depends of course on how large k is, as well as the regularity of the functions involved.

Example

Assume

$$\begin{aligned} p &\sim \text{Beta}(2.3, 4.1) \\ y \mid p &\sim \text{Binomial}(17, p) \\ y_{new} \mid p &\sim \text{Binomial}(3, p). \end{aligned}$$

Assume we would like to compute the probability $\pi(y_{new} = 1)$ given that $y = 4$. Using the theory developed above, we get the posterior

$$p \mid (y = 4) \sim \text{Beta}(2.3 + 4, 4.1 + 17 - 4) = \text{Beta}(6.3, 17.1).$$

According to the results of Exercise 8 of Section 1.7, the predictive distribution is Beta-Binomial, so we get

$$\begin{aligned} \pi(y_{new} = 1 \mid y = 4) &= \frac{B(6.3 + 1, 17.1 + 3 - 1)}{B(6.3, 17.1)} \binom{3}{1} \\ &= \frac{\Gamma(7.3)\Gamma(20.1)\Gamma(6.3 + 17.1)}{\Gamma(7.3 + 20.1)\Gamma(6.3)\Gamma(17.1)} \cdot \frac{3!}{1!2!} \\ &= 0.403364 \end{aligned}$$

The following R code approximates this result using discretization:

```
p <- seq(0, 1, length.out=20)
a <- dbeta(p, 2.3, 4.1)
b <- dbinom(4, 17, p)
c <- a*b/sum(a*b) #No need to divide a by its sum before this step
d <- dbinom(1, 3, p)
sum(c*d)
```

The code results in 0.4033704, which we see is a good approximation even if k is only 20. The advantage with the R code is of course that one may use any prior density on $[0, 1]$, not just a Beta density.

1.6 Numerical integration

Instead of discretizing one may apply numerical integration. After all, the answers we seek can be expressed as integrals:

$$\begin{aligned} \pi(y_{new} \mid y) &= \int_\theta \pi(y_{new} \mid \theta)\pi(\theta \mid y)\, d\theta = \int_\theta \pi(y_{new} \mid \theta) \frac{\pi(y \mid \theta)\pi(\theta)}{\int_\theta \pi(y \mid \theta)\pi(\theta)\, d\theta} \\ &= \frac{\int_\theta \pi(y_{new} \mid \theta)\pi(y \mid \theta)\pi(\theta)\, d\theta}{\int_\theta \pi(y \mid \theta)\pi(\theta)\, d\theta} \end{aligned}$$

Example

For simplicity we continue with the example from the previous section. So assume we would like to compute, as above, the probability that $y_{new} = 1$ given that $y = 4$. We get

$$\pi(y_{new} = 1 \mid y = 4) = \frac{\int_0^1 \text{Binomial}(1; 3, \theta)\, \text{Binomial}(4; 17, \theta)\, \text{Beta}(\theta; 2.3, 4.1)\, d\theta}{\int_0^1 \text{Binomial}(4; 17, \theta)\, \text{Beta}(\theta; 2.3, 4.1)\, d\theta}$$

and the R code

```
f1 <- function(theta) {dbinom(1, 3, theta)*dbinom(4, 17, theta)*
        dbeta(theta, 2.3, 4.1)}
f2 <- function(theta) {dbinom(4, 17, theta)*dbeta(theta, 2.3, 4.1)}
integrate(f1, 0, 1)$value/integrate(f2, 0, 1)$value
```

which produces the answer 0.403364, i.e., an even better approximation than with the discretization. As with discretization, the computations above can be done in principle with any densities, one is not limited to using conjugate priors.

Discretization or numerical integration works well in the simple example above, yielding accurate results. When variables are defined on unbounded intervals, one may need to make transformations before doing discretization. However, the biggest limitation to these methods is the dimension of the θ vector.

Let's imagine that, to get some kind of reasonable accuracy when discretizing a real variable, you cannot use fewer than 10 values to represent it. A density in n dimensions will then need 10^n points to represent it. For many problems n will be higher than, say, 10, giving at least 10^{10} gridpoints, which is unfeasible to handle. In fact we may very well want to study problems with millions of dimensions. Cleary, neither discretization nor numerical integration are then useful tools.

1.7 Exercises

1. A survey has been made about the type of living conditions and the political opinions of people in a city. Probabilities for observing each combination have been estimated and is listed in the following table

	Party A	Party B	Party C	Party D
Rental flat	0.11	0.03	0.08	0.01
Self-owned flat	0.09	0.01	0.14	0.03
House	0.13	0.04	0.09	0.24

   If you observe that a person lives in a rental flat, what is the probability that the person votes for party B?

2. A disease is affecting 0.7% of the population. Initial diagnosis is done with a somewhat unreliable test. If a person is affected, the test will be positive with a 95% probability. However, if the person is not affected, there is still a 5% chance that the test is positive. Given that the test is positive, what is the probability that the person is affected?

3. Assume you are making repeated independent experiments with a probability of success θ in each experiment. Initially, you make 12 experiments, of which 9 are successful.

   (a) Using a prior for θ that is uniform on the interval $[0, 1]$, what is the posterior for θ given the results of the 12 experiments?

(b) Assume now that you continue with doing 19 more experiments, of which 11 are successful. Given the combined information from all your 31 experiments, what is the posterior for θ?

(c) Given all the information above, what is the probability for success in your thirtysecond experiment?

4. Refer to the coin-flipping example in the beginning of the chapter.

   (a) Write down the probability of observing y_H heads and y_T tails during a sequence with $y_H + y_T$ coinflips.

   (b) Write down the probability of observing y_H heads and y_T tails followed by observing heads, during $y_H + y_T + 1$ coinflips.

   (c) Write down the conditional probability of observing heads after having observed y_H heads and y_T tails. Verify that your result is the same as that in Equation 1.2.

5. Write an R program that computes the parameter α for a Beta(α, α) distribution which has 90% of its density in the interval $[0.4, 0.6]$.

6. Assume $y \sim$ Negative-Binomial(α, p), where α is fixed and known and p is the unknown parameter. Prove that the Beta family is a conjugate prior family.

7. Assume $y \sim$ Normal(μ, τ^{-1}) where μ is fixed and τ is the unknown parameter. Prove that the Gamma family of distributions is a conjugate prior family.

8. Find the formula for the probability mass function of the prior predictive distribution for the Beta Binomial conjugacy. Try to look up if this distribution has a standard name.

9. In the beginning of this chapter, we discussed an example with a loaded dice. We will not make some explicit computations for this example. We will use a parameter vector $\theta = (\theta_1, \theta_2, \theta_3)$, where each $\theta_i \in (0, 1)$. Given this parameter, and based on the geometry of standard dice, we model the probability of obtaining the value k with the dice as

$$\pi(k \mid \theta) = \begin{cases} \frac{1}{3}\theta_1 & k = 1 \\ \frac{1}{3}\theta_2 & k = 2 \\ \frac{1}{3}\theta_3 & k = 3 \\ \frac{1}{3}(1-\theta_3) & k = 4 \\ \frac{1}{3}(1-\theta_2) & k = 5 \\ \frac{1}{3}(1-\theta_1) & k = 6 \end{cases}$$

   For the prior on θ, we use

$$\pi(\theta) = \text{Beta}(\theta_1; 20, 20) \cdot \text{Beta}(\theta_2; 20, 20) \cdot \text{Beta}(\theta_3; 20, 20)$$

(a) Find a formula for the posterior for θ given a specific observed sequence of throws $k_1, k_2, \ldots, k_n$.

(b) Find the probabilities for each of the outcomes $1, 2, \ldots, 6$ conditional on having observed the following sequence: 2, 4, 1, 6, 3, 6, 6, 3, 4, 2, 2.

10. Assume you have defined a likelihood function $\pi(y \mid \theta)$ and are given a family of priors $q_\gamma(\theta)$, parametrized by a vector γ of parameters, with $\gamma \in \Omega$. Assume now that this family is conjugate, so that, if the prior is $q_\gamma(\theta)$ for some $\gamma \in \Omega$, then the posterior $\theta \mid y$ has density $q_{f(\gamma)}(\theta)$ for some other $f(\gamma) \in \Omega$.

    (a) Fix an integer $k > 1$ and describe a new family of priors as consisting of all densities

    $$r(\theta) = \sum_{i=1}^{k} \lambda_i q_{\gamma_i}(\theta)$$

    where $\lambda_1, \ldots, \lambda_k$ are nonnegative real numbers summing to 1, and for all i, $\gamma_i \in \Omega$. Prove that this family is a conjugate family. Derive explicit formulas for the posterior given a prior like the one above.

    (b) Compute an explicit formula for the prior predictive distribution in this case.

    (c) Can you imagine an application where using this kind of *mixture prior* as a model could be advantageous?

11. An example of conjucacy is the Normal Normal conjugacy; see the Appendix in Chapter 8. Assume that $x \mid \mu \sim \text{Normal}(\mu, \tau^{-1})$ and $\mu \sim \text{Normal}(\mu_0, \tau_0^{-1})$ for fixed and known τ and τ_0. Then we known theoretically that the prior predictive distribution for x is normal. Use this fact, together with what is called in Dobrow the Law of Total Expectation and the Law of Total Variance, to find the parameters of this normal distribution.

Chapter 2

Hidden Markov Models

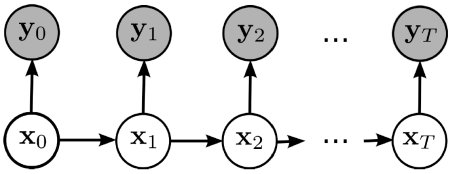

Figure 2.1: A hidden Markov model.

Let us now consider Hidden Markov Models, or HMMs. A general example of the following: A ("hidden") Markov chain $X_0, X_1, \ldots$, and conditionally on the values of this chain, another Markov chain $Y_0, Y_1, \ldots$. In fact, one generally assumes that

$$\pi(Y_i \mid Y_0, \ldots Y_{i-1}, X_0, X_1, \ldots) = \pi(Y_i \mid Y_{i-1}, X_i)$$

so that the dependence of Y_i is only on the corresponding term X_i in the Markov chain $X_0, X_1, \ldots$.

In this section we will consider a special case of this, so that the values of the Y_i do not depend other Y_j for $j \neq i$ when X_i is given. Specifically, we asume we have a stochastic process consisting of random variables $X_1, X_2, X_3, \ldots, Y_1, Y_2, Y_3, \ldots$ so that

- $X_1, X_2, X_3, \ldots$ is a Markov chain,
- the $Y_1, Y_2, \ldots$ are independent given $X_1, X_2, X_3, \ldots,$,
- $\pi(Y_i \mid X_1, X_2, \ldots) = \pi(Y_i \mid X_i)$ for all $i \geq 0$.

An example is illustrated in Figure 2.1.

Many different inference questions can be raised in this context. Here, we will restrict ourselves to the situation where the distributions $\pi(X_{i+1} \mid X_i)$ and $\pi(Y_i \mid X_i)$ are known for all i, and where the initial distribution $\pi(X_0)$ is

known. We also assume that we have observed the sequence $y_0, y_1, \ldots, y_T$ for the variables $Y_0, \ldots, Y_T$, and that our main objective is to find the marginal posterior distribution for each of the variables $X_0, \ldots, X_T$. These posterior distributions may be of interest in themselves or one may use the posterior distributions to make other predictions of interest.

The Forward algorithm

The objective is to compute, and store, for $i = 0, 1, \ldots, T$, the distributions $\pi(X_i \mid Y_0, \ldots, Y_i)$. This is done recursively, starting with $i = 0$, and at each step using the results of the previous step.

We first compute $\pi(X_0 \mid Y_0)$ using Bayes formula:

$$\pi(X_0 \mid Y_0) = \frac{\pi(Y_0 \mid X_0)\pi(X_0)}{\pi(Y_0)} \propto_{X_0} \pi(Y_0 \mid X_0)\pi(X_0).$$

Then, assuming that we have computed and (somehow) stored $\pi(X_i \mid Y_0 \ldots, Y_i)$ we compute $\pi(X_{i+1} \mid Y_0 \ldots, Y_{i+1})$ again using Bayes formula:

$$\begin{aligned} \pi(X_{i+1} \mid Y_0, \ldots, Y_{i+1}) &\propto_{X_{i+1}} \pi(Y_{i+1} \mid X_{i+1}, Y_0, \ldots, Y_i)\pi(X_{i+1} \mid Y_0, \ldots, Y_i) \\ &= \pi(Y_{i+1} \mid X_{i+1}) \int \pi(X_{i+1} \mid X_i)\pi(X_i \mid Y_0 \ldots, Y_i)\, dX_i \end{aligned}$$

The details of how these computations are done, and how the results are stored, depend on the particular types of distributions involved. In an alternative description of the Forward algorithm, one recursively computes $\pi(X_i \mid Y_0, \ldots, Y_{i-1})$ instead of $\pi(X_i \mid Y_0, \ldots, Y_i)$. The same ideas as above are used and the same computations are done; they are simply subdivided in a slightly different manner.

The Backward algorithm

The objective now is to compute and store, for $i = T, \ldots, 0$, the probabilities $\pi(Y_{i+1}, \ldots, Y_T \mid X_i)$. Note that when $i = T$, this expression is not really meaningful; however we will interpret $\pi(Y_{i+1} \ldots, Y_T \mid X_i)$ for $i = T$ as a function equal to 1 for all values of X_T. Starting with this function, we go backwards, stepwise decreasing the index i, and computing $\pi(Y_{i+1}, \ldots, Y_T \mid X_i)$ in terms of $\pi(Y_{i+2}, \ldots, Y_T \mid X_{i+1})$. We can to this by averaging out over X_{i+1}:

$$\begin{aligned} \pi(Y_{i+1}, \ldots, Y_T \mid X_i) &= \int \pi(Y_{i+1}, \ldots, Y_T, X_{i+1} \mid X_i)\, dX_{i+1} \\ &= \int \pi(Y_{i+1} \mid X_{i+1})\pi(Y_{i+2}, \ldots, Y_T \mid X_{i+1})\pi(X_{i+1} \mid X_i)\, dX_{i+1} \end{aligned}$$

The details of how these computations are done, and how the results are stored, depend on the particular types of distributions involved. In an alternative description of the Backward algorithm, one recursively computes $\pi(Y_i \ldots, Y_T \mid X_i)$ instead of $\pi(Y_{i+1}, \ldots, Y_T \mid X_i)$. the same ideas as above are used and the same computations are done; they are simply subdivided in a slightly different manner.

The Forward Backward algorithm

There are several interesting ways of putting together the two algorithms above. Let us for example assume that we would like to compute, for all $i = 0, \ldots, T$, the marginal posterior distributions $\pi(X_i \mid Y_0, \ldots, Y_T)$. We can do this by using (surprise!) Bayes formula:

$$\begin{aligned} \pi(X_i \mid Y_0 \ldots, Y_T) &\propto_{X_i} \pi(Y_{i+1}, \ldots, Y_T \mid X_i, Y_0, \ldots, Y_i)\pi(X_i \mid Y_0, \ldots, Y_i) \\ &= \pi(Y_{i+1}, \ldots, Y_T \mid X_i)\pi(X_i \mid Y_0, \ldots, Y_i) \end{aligned}$$

The distributions in the last line can be computed with the Backward and Forward algorithms, respectively.

Instead of finding the marginal distributions above, one might be interested in the joint distribution of all $X_0, \ldots, X_T$ given the observed values for $Y_0, \ldots, Y_T$. As this is a high-dimensional distribution, it is easier to focus on obtaining a sequence $x_0, x_1, \ldots, x_T$ which is a sample from this distribution. This can be done, for example, as follows: First, draw x_0 from $\pi(X_0 \mid Y_0, \ldots, Y_T)$ found as above. Then, for $i = 1, 2, \ldots, T$, draw x_i according to the density

$$\begin{aligned} & \pi(X_i \mid Y_0, \ldots, Y_T, X_0 = x_0, \ldots, X_{i-1} = x_{i-1}) \\ \propto_{X_i} \quad & \pi(Y_i, \ldots, Y_T \mid X_i, Y_0, \ldots, Y_{i-1}, X_0 = x_0, \ldots, X_{i-1} = x_{i-1}) \\ & \cdot\pi(X_i \mid Y_0, \ldots, Y_{i-1}, X_0 = x_0, \ldots, X_{i-1} = x_{i-1}) \\ = \quad & \pi(Y_i \mid X_i)\pi(Y_{i+1}, \ldots, Y_T \mid X_i)\pi(X_i \mid X_{i-1} = x_{i-1}) \end{aligned}$$

As always, the details depend on the types of distributions involved.

Implementation when the state space is finite

Let us now assume the state space for X is finite, with the possible values $1, \ldots, s$. The Forward algorithm can be implemented as follows: To compute $\pi(X_0 \mid Y_0) \propto_{X_0} \pi(Y_0 \mid X_0)\pi(X_0)$, compute $\pi(X_0)$ and $\pi(Y_0 \mid X_0)$ for all s possible values of X_0, to obtain two vectors of lenght s. Multiply these two vectors termwise to obtain a new vector of length s, and divide by its sum to obtain a vector of length s representing $\pi(X_0 \mid Y_0)$.

For the recursive part of the computation, let v_i denote the probability vector representing $\pi(X_i \mid Y_0 \ldots, Y_i)$. Then $v_i P$ is the probability vector representing $\int \pi(X_i + 1 \mid X_i)\pi(X_i \mid Y_0, \ldots, Y_i)\, dX_i$ as a function of X_{i+1}. Multiplying termwise with the probability vector representing $\pi(Y_{i+1} \mid X_{i+1})$ as a function of X_{i+1} and normalizing so that the sum becomes 1, we get the probability vector representing $\pi(X_{i+1} \mid Y_0, \ldots, Y_{i+1})$.

The Backward algorithm is implemented similarly. See the exercise which concerns writing an R implementation of these algorithms.

2.1 Exercise

1. Assume we have a Hidden Markov Model where the Markov chain has a state space consisting of 1, 2, 3, 4, a transition matrix

$$P = \begin{bmatrix} 0.7 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.7 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.7 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.7 \end{bmatrix}$$

and a distribution for the initial state X_0 given by the probability vector $(0.2, 0.4, 0.1, 0.3)$. Assume also that the possible values for Y_i are also 1, 2, 3, 4, and that the probability matrix for transitions from X_i to Y_i is given by

$$Q = \begin{bmatrix} 0.9 & 0.06 & 0.03 & 0.01 \\ 0.04 & 0.9 & 0.04 & 0.02 \\ 0.02 & 0.04 & 0.9 & 0.04 \\ 0.01 & 0.03 & 0.06 & 0.9 \end{bmatrix}$$

Assume we have observed the sequence 3, 4, 1, 1, 4, 3, 4, 3, 2, 2, 1 for $Y_0, \ldots, Y_{10}$.

   (a) Implement in R the Forward algorithm for this situation. Store the computed distributions.

   (b) Implement in R the Backward algorithm for this situation. Store the computed vectors.

   (c) For each $i = 0, \ldots, 10$, compute in R the vector representing the marginal posterior $X_i \mid Y_0, \ldots, Y_{10}$, when the Y_i have the values above.

   (d) Implement in R a function generating a sequence $x_0, \ldots, x_{10}$ representing a sample from $\pi(X_0, \ldots, X_{10} \mid Y_0, \ldots, Y_{10})$, when the Y_i have the values above.

Chapter 3

Some basic inference for Markov chains

Assume you would like to use a Markov chain $X_0, X_1, \ldots$, as your model in an applied setting where some data is available. The parameters of the Markov chain are the transition matrix P and the probability distribution p on the initial state X_0. You would then like to learn about these parameters from the data. This data could take many forms, for example, you could have observed only some specific selection of the variables X_i. In this chapter, we will assume that you have observed the whole sequence $X_0, X_1, \ldots, X_n$ up to some number n. After learning about the parameters, you can use this to predict further steps $X_{n+1}, X_{n+2}, \ldots$ of the chain.

We will also consider inference for Hidden Markov Models (HMMs). There is then a wider range of what your data could consist of. In this chapter, we will assume you have observed the hidden states $X_0, X_1, \ldots, X_n$ up to some n, and also the corresponding states $Y_0, \ldots, Y_n$, and that you would like to use this information to learn about the parameters of the HMM. We will also restrict ourselves to the case where the Y_i variables only depend on the underlying X_i variables, and not on each other.

3.1 The Multinomial Dirichlet conjugacy

The *Multinomial distribution* counts the number of outcomes in each of k possible classes when n independent trials are performed and the probability of ending up in each of the classes is given by the probability vector $p = (p_1, \ldots, p_k)$. (Recall that a probability vector p of length k is a vector of non-negative real numbers such that $\sum_{i=1}^{k} p_i = 1$.) In other words, a vector $x = (x_1, \ldots, x_k)$ of non-negative integers has a Multinomial distribution with parameters n and p

if $\sum_{i=1}^k x_i = n$ and the probability mass function is given by

$$\pi(x \mid n, p) = \binom{n}{x_1\ x_2\ x_3\ \dots\ x_k} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}.$$

Recall that the *multinomial coefficient* above is given by

$$\binom{n}{x_1\ x_2\ x_3\ \dots\ x_k} = \frac{n!}{x_1! x_2! \dots x_k!}.$$

Note that the Multinomial distribution with $k = 2$ can be identified with the Binomial distribution.

A vector $\theta = (\theta_1, \dots, \theta_k)$ of non-negative numbers satisfying $\sum_{i=1}^k \theta_i = 1$ has a *Dirichlet* distribution with parameter vector $\alpha = (\alpha_1, \dots, \alpha_k)$, with each $\alpha_i > 0$, if it has probability density function

$$\pi(\theta \mid \alpha) = \frac{\Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_k)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\dots\Gamma(\alpha_k)} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \dots \theta_k^{\alpha_k - 1}.$$

Note that the Dirichlet distribution with $k = 2$ can be identified with the Beta distribution.

As the Beta family is conjugate to the Binomal likelihood, it is natural to check if the Dirichlet family is conjugate to the Multinomial likelihood. So assume θ has the prior $\theta \mid \alpha \sim \text{Dirichlet}(\alpha)$ for some α, and assume we have the Multinomial likelihood $x \mid n, \theta \sim \text{Multinomial}(n, \theta)$. Bayes formula gives

$$\begin{aligned} \pi(\theta \mid x) &\propto_\theta \pi(x \mid \theta)\pi(\theta) \\ &\propto_\theta \theta_1^{x_1} \theta_2^{x_2} \dots \theta_k^{x_k} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \dots \theta_k^{\alpha_k - 1} \\ &= \theta_1^{\alpha_1 + x_1 - 1} \theta_2^{\alpha_2 + x_2 - 1} \dots \theta_k^{\alpha_k + x_k - 1} \end{aligned}$$

from which we deduce that

$$\theta \mid x \sim \text{Dirichlet}(\alpha_1 + x_1, \alpha_2 + x_2, \dots, \alpha_k + x_k)$$

and we have shown conjugacy. For the predictive distribution we get

$$\begin{aligned} \pi(x) &= \frac{\pi(x \mid \theta)\pi(\theta)}{\pi(\theta \mid x)} \qquad (3.1) \\ &= \frac{\text{Multinomial}(x; n, \theta)\, \text{Dirichlet}(\theta; \alpha)}{\text{Dirichlet}(\theta; \alpha + x)} \\ &= \frac{\frac{n!}{x_1! \dots x_k!} \theta_1^{x_1} \dots \theta_k^{x_k} \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_k)} \theta_1^{\alpha_1 - 1} \dots \theta_k^{\alpha_k - 1}}{\frac{\Gamma(\alpha_1 + \cdots + \alpha_k + x_1 + \cdots + x_k)}{\Gamma(\alpha_1 + x_1) \dots \Gamma(\alpha_k + x_k)} \theta_1^{\alpha_1 + x_1 - 1} \dots \theta_k^{\alpha_k + x_k - 1}} \\ &= \frac{n!}{x_1! \dots x_k!} \cdot \frac{\Gamma(\alpha_1 + x_1)}{\Gamma(\alpha_1)} \cdots \frac{\Gamma(\alpha_k + x_k)}{\Gamma(\alpha_k)} \cdot \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1 + \cdots + \alpha_k + x_1 + \cdots + x_k)} \end{aligned}$$

which is a generalization of the Beta Binomial distribution.

See Exercises 1 and 2 for more about these distributions.

3.2 Inference for time-homogeneous Markov chains with finite state space

Consider a time-homogeneous markov chain consisting of random variables $X_0, X_1, \ldots$, with finite state space S with s elements. The parameters of this model are p_0, the probability vector describing the distribution on X_0, and P, the transition matrix of the chain. Assume first that p_0 is known while P is unknown, and assume we would like to learn about P using a sequence of observations $x_0, x_1, \ldots, x_k$ for the variables $X_0, \ldots, X_k$. Let us write P_i for the i'th row of P: We now consider it a random variable, in fact, a probability vector with non-negative entries summing to 1. The probability of the data $x_0, \ldots, x_k$ for a fixed P is

$$\begin{aligned}
\pi(x_0, \ldots, x_k \mid P) &= \pi(x_0) \prod_{r=1}^{k} \pi(X_r \mid X_{r-1}, P) \\
&= \pi(x_0) \prod_{r=1}^{k} P_{x_{r-1}, x_r} \\
&= \pi(x_0) \prod_{i=1}^{s} \prod_{j=1}^{s} (P_{ij})^{c_{ij}}
\end{aligned}$$

where c_{ij} is the count of times the chain $x_0, x_1, \ldots, x_k$ transits from state i to state j. We now define a prior on P with

$$\pi(P) = \prod_{i=1}^{s} \text{Dirichlet}(P_i; \alpha_i)$$

where $\alpha_i = (\alpha_{i1}, \ldots, \alpha_{is})$ is a vector of parameters for $i = 1, \ldots, s$. For the posterior, we get

$$\begin{aligned}
\pi(P \mid x_0, \ldots, x_k) &\propto_P \pi(x_1, \ldots, x_k \mid P)\pi(P) \\
&\propto_P \prod_{i=1}^{s} \prod_{j=1}^{s} (P_{ij})^{c_{ij}} \prod_{i=1}^{s} \prod_{j=1}^{s} (P_{ij})^{\alpha_{ij}-1} \\
&\propto_P \prod_{i=1}^{s} \prod_{j=1}^{s} (P_{ij})^{\alpha_{ij}+c_{ij}-1}
\end{aligned}$$

from which we read off that

$$\pi(P \mid x_0, \ldots, x_k) = \prod_{i=1}^{s} \text{Dirichlet}(P_i; \alpha_i + c_i)$$

where $c_i = (c_{i1}, \ldots, c_{is})$.

Example

Assume you have a Markov chain with three possible states, and that a sequence of values $x_0, \ldots, x_{20}$ have been observed. Assume the counts c_{ij} of transitions from state i to state j are given by the following table:

3	4	1
3	3	0
2	0	4

A classical inference approach might try to use the observed frequencies of transitions as the values in the transition matrix P. But we see that some values of P_{ij} would then become zero. We have seen that whether the entries of the transition matrix are zero or positive can have a decisive influence on the properties of the Markov chain, and it may seem rash to conclude that some transitions have probability zero simply because they have not been observed in a short sequence of the chain. Even more fundamentally, with other data, some states might not have been visited at all. All counts in the vector c_i would then be zero, and it would be impossible to compute frequencies summing to one from such a vector. Note that even if such problems might be overcome with more data, the amount of data needed for good frequency estimates increases dramatically with the number s of states in the state space.

In practice, to get a useful result for P, one may need to use more information than that available in the counts c_{ij}. In the Bayesian inference above, such information is provided in the parameters α_i of the prior for P. Note that if $x \sim \text{Dirichlet}(\lambda)$, then $\text{E}[x] = \lambda / \sum_{i=1}^{k} \lambda_i$. Thus, in our case,

$$\text{E}(P_i \mid x_0, \ldots, x_k) = \frac{\alpha_i + c_i}{\alpha_{i1} + \cdots + \alpha_{is} + c_{i1} + \cdots + c_{is}}.$$

Thus, as long as we use parameters $\alpha_{ij} > 0$, all posterior expectations of values in P will be nonzero. In many situations, a reasonable choice may be $\alpha_{ij} = 1$ for all i, j, leading to

$$\text{E}(P_i \mid x_0, \ldots, x_k) = \frac{(1, 1, \ldots, 1) + c_i}{s + c_{i1} + \cdots + c_{is}}.$$

In the example above, the posterior expectation for P would be the matrix

$$\text{E}(P \mid x_0, \ldots, x_k) = \begin{bmatrix} 4/11 & 5/11 & 2/11 \\ 4/9 & 4/9 & 1/9 \\ 3/9 & 1/9 & 5/9 \end{bmatrix}. \tag{3.2}$$

The values α_{ij} are sometimes called *pseudocounts*; however, they do not need to be integers.

The situation where the distribution p_0 for X_0 is unknown can be handled in a similar way. Note however that if we have observed only one sequence $x_0, \ldots, x_k$, only x_0 informs us about p_0, so unless we have observed a number of sequences from the chain, the distribution for p_0 will be more or less determined by the prior.

3.2.1 Prediction

Let us assume we would like to predict the observation x_{k+1} of X_{k+1} based on the sequence $x_0, \ldots, x_k$. We can write

$$\begin{aligned} \pi(x_{k+1} \mid x_0, \ldots, x_k) &= \int \pi(x_{k+1} \mid x_k, P)\pi(P \mid x_0 \ldots, x_k)\, dP \\ &= \int P_{x_k, x_{k+1}} \pi(P_{x_k} \mid x_0 \ldots, x_k)\, dP_{x_k} \end{aligned} \tag{3.3}$$

If we use a prior and compute the posterior for P_{x_k} as in the last subsection, we get

$$P_{x_k} \mid x_0, \ldots, x_k \sim \text{Dirchlet}(\alpha_{x_k} + c_{x_k}).$$

According to Equation 3.3, the predictive distribution is given as the Expectation vector of this Dirichlet:

$$\pi(x_{k+1} \mid x_0, \ldots, x_k) = \frac{\alpha_{x_k} + c_{x_k}}{\alpha_{x_k,1} + \cdots + \alpha_{x_k,s} + c_{x_k,1} + \cdots + c_{x_k,s}}$$

Consider the example of the previous section, where $x_0, \ldots, x_{20}$ were observed; assume that $x_{20} = 2$. We showed in Equation 3.2 that the expectation of P_2, the second row of the transition matrix, was $(4/9, 4/9, 1/9)$. Thus we get that

$$\pi(x_{k+1} \mid x_0, \ldots, x_k) = (4/9, 4/9, 1/9)\,.$$

Let us also consider the prediction of a whole chain of observations $x_{k+1}, \ldots, x_{k+r}$ based on the sequence $x_0, \ldots, x_k$. In the same way as above, we get

$$\begin{aligned} & \pi(x_{k+1}, \ldots, x_{k+r} \mid x_0, \ldots, x_k) \\ = & \int \left[\prod_{s=1}^{r} \pi(x_{k+s} \mid x_{k+s-1}, P) \right] \pi(P \mid x_0 \ldots, x_k)\, dP \\ = & \int \left[\prod_{s=1}^{r} P_{x_s, x_{k+s-1}} \right] \pi(P \mid x_0 \ldots, x_k)\, dP. \end{aligned} \tag{3.4}$$

When the posterior for P is a product of Dirichlet distributions, it is in fact possible to compute the value of this integral, in a simiar way as in computations for the predictive distribution for the Multinomial Dirichlet conjugate pair given in Equation 3.1. See Exercise 4 for concrete computations.

3.2.2 Extensions

The Dirichlet priors we have considered here assumes that the transition matrix is positive. However, there may be situations where certain transitions may be ruled out apriori. In such cases, an alternative is to use Dirichlet distributions on the parameters in each line that could be non-zero. See Exercise 3 for an example.

In other situations, the assumption used above that the lines of the transition matrix are a priori independent may be unreasonable. In such cases, a prior reflecting this situation could be used.

In yet other situations, it may be known that the Markov chain is time reversible, so that inference about its parameters should be done under this restriction. A possibility is then to represent the Markov chain as a random walk on a weighted undirected graph, and infer the weights from data. It is even possible to use a conjugate analysis in this case.

The above discussion on predictions can also be extended in many directions. For example, making predictions for long stretches of a Markov chain may best be done by first simulating its transition matrix from the posterior and then continuing the Markov chain simulating with this transition matrix. Finally, one may study how a stationary distribution derived from a transition matrix changes when taking into account the posterior uncertainty in this transition matrix.

3.3 Inference for HMMs

Assume $X_0, X_1, \ldots, X_n, \ldots$, is a Markov chain with a discrete state space, transition matrix P, and probability distribution p on the initial state X_0. Assume also $Y_0, Y_1, \ldots, Y_n, \ldots$, are discrete random variables so that

$$\pi(Y_n \mid X_0, X_1, \ldots, Y_0, \ldots, Y_{n-1}, Y_{n+1}, \ldots) = \pi(Y_n \mid X_n).$$

Assume further that $\Pr(Y_n = j \mid X_n = i) = Q_{ij}$ is independent of n, so there is a single matrix Q describing the dependence of Y_n on X_n. Finally, assume one has observed $X_0, \ldots, X_n$ and $Y_0, \ldots, Y_n$. How can we learn about the parameters P, p and Q of our model?

This will depend on what prior distribution we use for P, p, and Q. Generally, one will use independent priors for these. Then, the learning for P and p will be done as for any Markov chain $X_0, \ldots, X_n, \ldots$, see the previous section. The learning for Q will be based on the $n+1$ observed pairs (X_0, Y_0), (X_1, Y_1), $\ldots$, (X_n, Y_n). In the example below, we try out two different priors for Q, to illustrate how this choice influences results.

Example

Assume the following values have been observed for $X_0, \ldots, X_{20}$ and $Y_0, \ldots, Y_{20}$:

X	3	3	3	1	2	1	1	2	2	1	1	3	3	3	1	1	2	2	1	2	2
Y	1	4	3	2	3	2	1	1	4	1	1	3	3	4	0	0	3	0	0	2	2

We assume the X variable has possible values 1, 2, 3. The counts of transitions are exactly the same as those of the Example of the previous section, and we can learn about the transition matrix P in exactly the same way as in that section.

Let us first assume that Y can only take on the values 0, 1, 2, 3, 4, and use as a prior for Q a product of Dirichlet distributions. Specifically,

$$\pi(Q) = \prod_{i=1}^{3} \text{Dirichlet}(Q_i; \beta_i)$$

where Q_i is the i'th row of the Q matrix, and the vector β_i is the corresponding set of *pseudocounts* for the transitions from state i. The counts of transitions from the possible values $1, 2, 3$ for X to the possible values $0, 1, 2, 3, 4$ for Y is given in the following table:

	0	1	2	3	4
1	3	3	2	0	0
2	1	1	2	2	1
3	0	1	0	3	2

Just like in the previous section, we get that the posterior for Q is also a product of Dirichlet distributions:

$$\pi(Q \mid \text{data}) = \prod_{i=1}^{3} \text{Dirichlet}(Q_i; \beta_i + d_i)$$

where d_i is the vector of counts of transitions from state i. So, for example, $d_1 = (3, 3, 2, 0, 0)$ and $d_2 = (1, 1, 2, 2, 1)$. Setting all the pseudocounts equal to 1, we get, explicitly,

$$\pi(Q \mid \text{data}) = \text{Dirchlet}(Q_1; 4, 4, 3, 1, 1)\,\text{Dirchlet}(Q_2; 2, 2, 3, 3, 2)\,\text{Dirichlet}(Q_3; 1, 2, 1, 4, 3).$$

Computing expectations as in the previous section, if a state X_i for $i > 20$ has value 2, the probabilities for Y_j are given by the vector $(2/12, 2/12, 3/12, 3/12, 2/12)$, i.e.,

```
[1] 0.1666667 0.1666667 0.2500000 0.2500000 0.1666667
```

We can make a prediction for Y_{21} by conditioning on the probabilities for X_{21} found in the previous section. We get, for example,

$$\begin{aligned}
\Pr(Y_{21} = 4) &= \mathrm{E}[\Pr(Y_{21} = 4) \mid X_{21}] \\
&= \Pr(X_{21} = 1)\,\mathrm{E}(Q_{14}) + \Pr(X_{21} = 2)\,\mathrm{E}(Q_{24}) + \Pr(X_{21} = 3)\,\mathrm{E}(Q_{34}) \\
&= \frac{4}{9} \cdot \frac{1}{13} + \frac{4}{9} \cdot \frac{2}{12} + \frac{1}{9} \cdot \frac{3}{11} \\
&= 0.1385651.
\end{aligned}$$

For illustration, we try out a second, more structured prior: $Y_i \sim \text{Poisson}(\lambda X_i)$, where we use a $\text{Gamma}(2, 2)$ prior for λ. Such a prior might be chosen if there is a reason to believe that each Y_i is Poisson distributed with some underlying parameter that might be proportional to the value X_i of the "hidden chain".

When X_i and λ are known all Y_i are independent and we can update our knowledge about λ stepwise, using information from one observed Y_i at the time. Specifically, if $\lambda \sim \text{Gamma}(\alpha, \beta)$ and $y \sim \text{Poisson}(x\lambda)$, then

$$\begin{aligned} \pi(\lambda \mid y) &\propto_\lambda \pi(y \mid \lambda)\pi(\lambda) \\ &\propto_\lambda e^{-\lambda x}\frac{(\lambda x)^y}{y!}\lambda^{\alpha-1}\exp(-\lambda\beta) \\ &\propto_\lambda \lambda^{\alpha+y-1}\exp(-\lambda(\beta+x)) \end{aligned}$$

so $\lambda \mid y \sim \text{Gamma}(\alpha + y, \beta + x)$. Thus, to find the posterior for λ given all the data, we take the first parameter and add the sum of all the Y_i observed, obtaining $2 + 40 = 42$. To the second parameter we add 1 times the number of observations with $X_i = 1$, 2 times the number of observations with $X_i = 2$, and 3 times the number of obervations with $X_i = 3$, obtaining $2+1\cdot 8+2\cdot 7+3\cdot 6 = 42$. Thus the have the posterior $\lambda \mid \text{data} \sim \text{Gamma}(42, 42)$.

If a state X_i for $i > 20$ has value 2, the probabilities for Y_i will be given by the Poisson distribution with parameter 2λ, where $\lambda \sim \text{Gamma}(42, 42)$. Writing $\lambda' = 2\lambda$, it is fairly direct to show that $\lambda' \sim \text{Gamma}(42, 42/2) = \text{Gamma}(42, 21)$. From the general conjugacy theory of Chapter 1 we have that if $y \sim \text{Poisson}(\lambda)$ and $\lambda \sim \text{Gamma}(\alpha, \beta)$, then $y \sim \text{Negative-Binomial}(\alpha, \beta/(\beta+1))$. Thus, in our case $Y_i \mid \text{data} \sim \text{Negative-Binomial}(41, 21/(21+1))$, and we can for example use the R command `dnbinom(0:4, 42, 21/22)` to compute some probabilities for Y_i, resulting in

```
[1] 0.1417287 0.2705730 0.2644236 0.1762824 0.0901444
```

The probability for $Y_{21} = 4$ using the distribution for X_{21} can be found with

```
sum(dnbinom(4, 42, 42/(1:3)/(42/(1:3) + 1))*c(4, 4, 1)/9)
```

producing 0.06530846.

3.4 Exercises

1. Assume an experiment can have one of three outcomes; let us name the outcomes 1, 2, and 3. Assume the probabilities for these outcomes are p_1, p_2, and p_3, respectively, but that these probabilities are unknown. Assume 13 independent experiments are performed, of which 3 have outcome 1, 9 have outcome 2, and 1 has outcome 3.

   (a) Using a Dirichlet(α) prior for $p = (p_1, p_2, p_3)$, where $\alpha = (\alpha_1, \alpha_2, \alpha_3) = (1, 1, 1)$, find the posterior for p.

   (b) Find the expected posterior value for p. (See the Appendix on probability distributions).

   (c) Still using the same prior, compute the probability that, among the next 4 experiments, there will be 1 with outcome 1, 2 with outcome 2, and 1 with outcome 3.

2. (a) Convince yourself that a Dirichlet distribution with $k = 2$ is the same as a Beta distribution, just using different notation.

   (b) Define a function on the set of non-negative vectors $\theta = (\theta_1, \ldots, \theta_k)$ with $\sum_{i=1}^{k} \theta_i = K$ by

   $$\pi(\theta \mid \alpha, K) = \frac{1}{K^{\alpha_1+\cdots+\alpha_k-1}} \cdot \frac{\Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_k)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\ldots\Gamma(\alpha_k)} \theta_1^{\alpha_1-1}\theta_2^{\alpha_2-1}\ldots\theta_k^{\alpha_k-1}.$$

   Show that π is a density on this set, i.e., that it integrates to 1. We will use the notation $\theta \mid \alpha, K \sim \text{Dirichlet}_K(\alpha)$.

   (c) Assuming that $(\theta_1, \ldots, \theta_k) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k)$, use a proportionality argument to show that for any i with $2 < i < k$,

   $$\theta_1 \ldots, \theta_{i-1} \mid \theta_i, \ldots, \theta_k \sim \text{Dirichlet}_{1-\theta_i-\cdots-\theta_k}(\alpha_1, \ldots, \alpha_{i-1}).$$

   (Note also that the ordering of the indexes does not matter in our context).

   (d) For i with $1 < i < k$, use the identity

   $$\pi(\theta_1, \ldots, \theta_i) = \frac{\pi(\theta_1, \ldots, \theta_k)}{\pi(\theta_{i+1}, \ldots, \theta_k \mid \theta_1, \ldots, \theta_i)}$$

   to compute the marginal density $\pi(\theta_1, \ldots, \theta_i)$ up to a constant not depending on $\theta_1, \ldots, \theta_i$.

   (e) Assuming that $(\theta_1, \ldots, \theta_k) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k)$, find the distribution of the random vector $(\theta_1, \ldots, \theta_i, \theta_{i+1} + \cdots + \theta_k)$.

   (f) Assuming that $(\theta_1, \ldots, \theta_k) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k)$, show that $\theta_1 + \cdots + \theta_i$ has a Beta distribution, and find the parameters of this distribution.

3. Assume a Markov chain with state space containing the numbers 1, 2, 3, 4 has been observed for 26 steps. The values in these steps are 1, 2, 3, 2, 2, 3, 4, 4, 3, 2, 3, 2, 1, 1, 2, 1, 2, 3, 4, 3, 4, 3, 3, 2, 1, 1.

   (a) Write down an estimate for the transition matrix P based only on frequencies of observed transitions.

   (b) Using a prior for the transition matrix consisting of a product of Dirchlet distributions with all pseudo-counts equal to 1, find the expectation of the posterior for the transition matrix given the observed sequence above.

   (c) Given the same prior, compute the posterior distribution for $P_{11} \mid$ data. (Hint: You may need to look up, or solve Exercise 2 above, for the marginal distribution for the components of a Dirchlet distriution). In particular compute $P(P_{11} > 0.3 \mid \text{data})$.

(d) Assume now that you have prior information that transitions in the Markov chain cannot happen to states whose value differs more than one compared to the current state. Reformulate a new prior for P incorporating this information. Then, recompute the results from questions (b) and (c) above using this new prior.

4. We will now compute the explicit value for the predictive distribution in Equation 3.4.

   (a) Assume your data is a *specific sequence* $z = (z_1, z_2, \ldots, z_n)$ of outcomes from n independent trials where each trial can have one of k outcomes, with probabilities of the outcomes given by a probability vector $p = (p_1, \ldots, p_k)$. Show that the probability mass function for z is given by

   $$\pi(z \mid n, p) = p_1^{y_1} p_2^{y_2} \cdots p_k^{y_k}$$

   where y_i is the count of values in the sequence z equal to the i'th outcome. Go through the discussion in this chapter about the Multinomial Dirichlet conjugacy and show what needs to be changed when the Multinomial density is replaced with the density above. In particular, prove that the Dirichlet family is a conjugate family, and compute the predictive distribution corresponding to Equation 3.1.

   (b) Show that the value of Equation 3.4 can be computed as a product of predictive distributions like those you found in (a). Find the formula for this product.

   (c) Consider the data of Exercise 3 above, and the prior used in (3b) and (3c). Given this prior, compute the probability of observing the sequence 1,2,2,3 after the sequence given in 3.

Chapter 4

Some basic inference for Branching processes

Assume you want to use a Branching process, as defined in Dobrow, as a model in some applied setting. The parameter defining such a Branching process is the vector a of probabilities in the offspring distribution of having $0, 1, 2, \ldots,$ offspring. Various types of data that one might learn about this parameter are conceivable. Below, we will simply assume that the data consists of n independent observations of the offspring distribution, i.e., counts $y_1, y_2, \ldots, y_n$ of the number of actual offspring in n different cases.

Within the Bayesian paradigm, one would start with defining a prior $\pi(a)$ based on the context of the situation. The likelihood of the data is the product $\prod_{i=1}^{n} a_{y_i}$ and we can derive a posterior distribution $\pi(a \mid \text{data})$. This posterior can then be used in any kind of prediction one might want to pursue: It could be for example simulation or computation of coming generation sizes or simulation or computation of the probability of extinction.

In this chapter we will limit ourselves to presenting two examples of possible priors to use for a, while a third example will be presented in the Exercise. We will focus on how to derive the posterior $\pi(a \mid \text{data})$ and only mention briefly how such a posterior can then be used in predictions.

4.1 Example: A Binomial model

Assume from the context it is natural to assume that the number of offspring is between 0 and N, distributed according to a Binomial distribution with some parameter p. A possibility is to use a Beta(α, β) prior for p. The posterior for p after the observations $y_1, \ldots, y_n$ then becomes Beta $(\alpha + S, \beta + nN - S)$ where $S = \sum_{i=1}^{n} y_i$. This distribution can then be used in predictions for the further growth of the branching process.

For illustrational purposes, let us instead assume we use the following prior

$\pi(p) = f(p)$ where

$$f(p) = \begin{cases} 100(p - 0.1) & 0.1 \leq p \leq 0.2 \\ 100(0.3 - p) & 0.2 \leq p \leq 0.3 \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

Then we can compute the posterior (note that $\text{Beta}(p; 1, 1) = 1$ for $0 \leq p \leq 1$)

$$\begin{aligned} \pi(p \mid \text{data}) &\propto_p \pi(\text{data} \mid p)\pi(p) \propto_p \left(\prod_{i=1}^{n} \text{Binomial}(y_i; N, p) \right) \text{Beta}(p; 1, 1) f(p) \\ &\propto_p \text{Beta}(p; 1 + S, 1 + nN - S) f(p). \end{aligned}$$

Computational example

Let us assume there can be a maximum of $N = 6$ offspring and that the number of offspring is distributed according to a Binomial$(6, p)$ distribution where p has the prior of Equation 4.1. Assume the number of offspring in 342 observed cases are

Number of offspring	0	1	2	3	4	5	6
Number of cases	117	138	58	25	3	0	1

so that the total number of offspring is $S = 347$. The posterior for p is proportional to $\text{Beta}(p; 1 + 347, 1 + 342 \cdot 6 - 347) f(p) = \text{Beta}(p; 348, 1706) f(p)$.

What is the probability that the branching process is supercritical? As the offspring process is Binomial its expectation is $6p$, and we would like to compute the probability that $6p > 1$, i.e., that $p > 1/6$. This can be done for example with

```
prior <- function(p) {
  if (p<0.1) 0 else if (p<0.2) 100*(p-0.1) else
    if (p<0.3) 100*(0.3-p) else 0
}
f <- function(p) dbeta(p, 348, 1706)*Vectorize(prior)(p)
integrate(f, 1/6, 1)$value/integrate(f, 0, 1)$value
```

producing 0.671135.

4.2 Example: Using the Multinomial Dirichlet conjugacy

Again for illustration, we include another example of how the prior for the offspring distribution parameter vector a can be chosen. Now, we use a Dirichlet distribution on possible values 0, 1, 2, 3, 4, 5, 6, with pseudocounts 1. In other words, we use the prior

$$\pi(a) = \text{Dirichlet}(a; (1, 1, 1, 1, 1, 1, 1))$$

Re-using the data from the previous section we get the posterior

$$\pi(a \mid \text{data}) = \text{Dirichlet}(a; 1{+}117, 1{+}138, 1{+}58, 1{+}25, 1{+}3, 1, 1{+}1) = \text{Dirchlet}(a; 118, 139, 59, 26, 4, 1, 2)$$

As we have 342 cases, the sum of the parameters becomes $342 + 7 = 349$ and the expected parameter vector becomes

$$\text{E}[a] = \left(\frac{118}{349}, \frac{139}{349}, \frac{59}{349}, \frac{26}{349}, \frac{4}{349}, \frac{1}{349}, \frac{2}{349}\right).$$

With this expected parameter vector, the expectation of the offspring process becomes

$$\text{E}[\mu] = 0 \cdot \frac{118}{349} + 1 \cdot \frac{139}{349} + 2 \cdot \frac{59}{349} + 3 \cdot \frac{26}{349} + 4 \cdot \frac{4}{349} + 5 \cdot \frac{1}{349} + 6 \cdot \frac{2}{349} = 1.054441$$

which indicates that the Branching process may be supercritical. However, there is an uncertainty in μ, and we can ask for the probability that the branching process is supercritical, i.e., that $\mu > 1$. Below, we use simulation as a quick and simple way to estimate this probability:

```
mean(rdirichlet(1000000, c(118,139,59,26,4,1,2))%*%(0:6)>1)
```

produces 0.835042. The function `rdirichlet` simulates from the Dirichlet distribution and can be found in the R package LearnBayes.

4.3 Exercise

1. In this exercise we use the same data as in the examples above. However, we now assume that the offspring distribution is Geometric(p), with a uniform prior $\pi(p) \sim \text{Uniform}(0, 1)$.

   (a) If $y \sim \text{Geometric}(p)$ find a family of distributions for p that is a conjugate family. Do this by guessing and trying out if your guess is correct. (If this fails you may also look up a conjugate family for the Geometric distribution).

   (b) Compute the posterior distribution for the parameter p of the offspring distribution when the prior above is used.

   (c) Compute the probability that $\mu > 1$ in this case.

Chapter 5

Markov chain Metropolis Hastings (MCMC)

As described in Section 1.4, the overall idea of Bayesian inference is to describe a stochastid model with variables representing both y, the observed data, and y_{new}, whatever you would like to predict, and then use the resulting conditional distribution of $y_{new} \mid y$ for predictions. Usually, one describes the model in terms of a vector of parameters θ, describing models for likelihoods $y \mid \theta$ and $y_{new} \mid \theta$ and a prior distribution for θ in such a way that y and y_{new} are conditionally independent given θ, i.e.,

$$\pi(y_{new} \mid \theta, y) = \pi(y_{new} \mid \theta).$$

We can then write

$$\pi(y_{new} \mid y) = \int_{\theta} \pi(y_{new} \mid \theta)\pi(\theta \mid y)\, d\theta \tag{5.1}$$

and Bayesian inference resolves into three steps: First describing the stochastic model (i.e., the likelihoods and the prior), second deriving from this the posterior $\pi(\theta \mid y)$, and finally using the posterior to to make predictions according to Equation 5.1.

Computing the integral in Equation 5.1 can be a big challenge in many types of models. In Chapter 1, we saw how it can be done numerically when the number of dimensions of θ is very low (in practice 1-3). We have also seen a number of cases where the integral can be computed analytically, i.e., where we can use conjugacy. However, in most realistic models, none of these options are available, and one needs to turn to approximate numerical approaches.

Assume we can generate a sample $\theta_1, \theta_2, \ldots, \theta_m$ from the posterior distribution $\pi(\theta \mid y)$. Then we can approximate the integral above with the average of the numbers $\pi(y_{new} \mid \theta_i)$. More precisely, the Strong Law of Large Numbers

gives that, with probability 1,

$$\int_\theta \pi(y_{new} \mid \theta)\pi(\theta \mid y)\, d\theta = \mathrm{E}_{\theta|y}\left[\pi(y_{new} \mid \theta)\right] = \lim_{m\to\infty} \frac{1}{m}\sum_{i=1}^{m} \pi(y_{new} \mid \theta_i). \quad (5.2)$$

Importantly for us, Equation 5.2 holds even in the case where $\theta_1, \theta_2, \ldots, \theta_i, \ldots$ is not a random sample but instead the values in a Markov chain with limiting distribution $\pi(\theta \mid y)$. Equation 5.2 is then called the Strong Law of Large Numbers for Markov chains. The second and third steps of Bayesian inference can now be done as follows: Generate a sequence $\theta_1, \ldots, \theta_m$ from a Markov chain with the posterior $\pi(\theta \mid y)$ as limiting distribution, and approximate

$$\pi(y_{new} \mid y) \approx \frac{1}{m}\sum_{i=1}^{m} \pi(y_{new} \mid \theta_i). \quad (5.3)$$

This technique for Bayesian inference is called Markov chain Monte Carlo, or MCMC. To use it, we need to

1. Define and simulate from a Markov chain with our posterior $\pi(\theta \mid y)$ as a limiting distribution.
2. Do this in a way so that we get a good approximation in Equation 5.3, or at least, do it so that we know how good our approximation is.

It turns out that the first point above is often surprisingly easy. We will briefly describe the Metropolis Hastings algorithm in the next section. However, the second point is surprisingly difficult. We will discuss it in the remaining parts of this chapter.

5.1 The Metropolis Hastings algorithm

Assume given a likelihood function $\pi(y \mid \theta)$ and a prior $\pi(\theta)$. Define a *proposal function* which, for every θ in the set Ω of possible parameters describes a probability density $q(\theta^* \mid \theta)$ on $\theta^* \in \Omega$. Assume you have an algorithm to simulate θ^* from this proposal distribution. Then the Metropolis Hastings algorithm for Bayesian inference is:

1. Simulate θ_0 from some distribution on Ω.
2. For $i = 1, \ldots, m$:
   (a) Generate θ^* from $q(\theta^* \mid \theta_{i-1})$.
   (b) Generate $U \sim \mathrm{Uniform}(0, 1)$.
   (c) If
   $$U < a_{\theta,\theta^*} = \frac{\pi(y \mid \theta^*)\pi(\theta^*)q(\theta_{i-1} \mid \theta^*)}{\pi(y \mid \theta_{i-1})\pi(\theta_{i-1})q(\theta^* \mid \theta_{i-1})}$$
   set $\theta_i = \theta^*$, otherwise set $\theta_i = \theta_{i-1}$.

This will generate a Markov chain $\theta_0, \theta_1, \ldots$. In Dobrow it is proven that as long as this Markov chain is ergodic, it will have limiting distribution $\pi(\theta \mid y)$. Note how

$$\pi(y \mid \theta^*) = \frac{\pi(y \mid \theta^*)\pi(\theta^*)}{\pi(y)} \quad \text{and} \quad \pi(y \mid \theta_{i-1}) = \frac{\pi(y \mid \theta_{i-1})\pi(\theta_{i-1})}{\pi(y)}$$

implies

$$\frac{\pi(\theta^* \mid y)}{\pi(\theta_{i-1} \mid y)} = \frac{\pi(y \mid \theta^*)\pi(\theta^*)}{\pi(y \mid \theta_{i-1})\pi(\theta_{i-1})}.$$

This is why we can use the quotient on the right instead of the quotient on the left when computing a_{θ,θ^*}.

The data y will often consist of a random sample $y_1, y_2, \ldots, y_n$ of observations that are independent given θ. Then $\pi(y \mid \theta) = \prod_{j=1}^n \pi(y_i \mid \theta)$ and numerically this number may come extremely close to zero, in particular when n is large. Writing

$$\begin{aligned} a_{\theta,\theta^*} &= \frac{\pi(y \mid \theta^*)\pi(\theta^*)q(\theta_{i-1} \mid \theta^*)}{\pi(y \mid \theta_{i-1})\pi(\theta_{i-1})q(\theta^* \mid \theta_{i-1})} \\ &= \frac{\prod_{j=1}^n \pi(y_j \mid \theta^*)\pi(\theta^*)q(\theta_{i-1} \mid \theta^*)}{\prod_{j=1}^n \pi(y_j \mid \theta_{i-1})\pi(\theta_{i-1})q(\theta^* \mid \theta_{i-1})} \\ &= \exp\left(\sum_{j=1}^n [\log \pi(y_j \mid \theta^*) - \log \pi(y_j \mid \theta_{i-1})]\right. \\ &\quad \left. + \log \pi(\theta^*) - \log \pi(\theta_{i-1}) + \log q(\theta_{i-1} \mid \theta^*) - \log q(\theta^* \mid \theta_{i-1})\right) \end{aligned}$$

and computing a_{θ,θ^*} according to the last expression is a way to avoid underflow problems on the computer.

5.2 Assessing convergence of the Metropolis Hastings algorithm

If whe have chosen a proposal function and proven that the resulting Metropolis Hastings Markov chain is ergodic, Equation 5.2 shows us that, for large enough m, Equation 5.3 can be used as an approximation. However, this does not tell us how large m needs to be, or if we choose a particular m, how good the approximation in Equation 5.3 is. It turns out that this is the major drawback with MCMC. Except for a few cases (see Dobrow) one is left with ad-hoc methods for making decisions about m and about the convergence. In this section we briefly discuss some of these methods.

First of all, as the sequence $\theta_0, \theta_1, \ldots, \theta_m$ is *approaching* the limiting distribution, it seems reasonable to put most emphasis on the last part of the sequence when using Equation 5.3. Indeed, a standard practice is to throw away a first part $\theta_0, \theta_1, \ldots, \theta_k$ of the sequence and only use $\theta_{k+1}, \theta_{k+2}, \ldots, \theta_m$

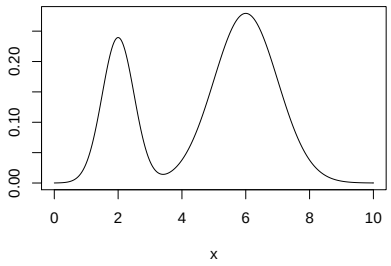

Figure 5.1: The density used in Section 5.3.

when computing this average. The sequence $\theta_0, \ldots, \theta_k$ is called the *burn-in*. But how should one choose k? An ad-hoc method is to use *trace plots*. A trace plot plots a component of the vector θ_i as a function of the index i. Often, one will see that for smaller indices, the trace plot shows some kind of trend, but after a while, the values settle into a pattern of movements that are stable over time. The part with a trend is then designated the burn-in. See the example below.

Still, the question remains how one can be sure that "the values have settled into a pattern of movements that are table over time". Many proposal functions are created so that they propose only minor changes to θ. Thus, trace plots will show how the Markov chain moves around in the posterior $\pi(\theta \mid y)$ using limited step sizes. A common problem in MCMC is that the chain may not even visit all "parts" of the posterior, where the "parts" are defined as sets with high posterior density separated by sets with low posterior density.

5.3 Example

We look at a toy example where θ has only one dimension and we assume the posterior distribution has been found to be

$$\pi(\theta \mid y) = 0.3 \cdot \text{Normal}(\theta; 2, 0.5^2) + 0.7 \cdot \text{Normal}(\theta; 6, 1^2).$$

Figure 5.1 shows a plot of the density. Cleary, it is easy to simulate directly from this distribution (by simulating a random variable which decides whether the first or the second normal density is used). For comparison, the following minimalist code uses Metropolis Hastings to generate a value drawn from this distribution:

```
f <- function(theta) 0.3*dnorm(theta, 2, 0.5)+0.7*dnorm(theta,6,1)
theta <- 1
for (i in 1:1000) {
```

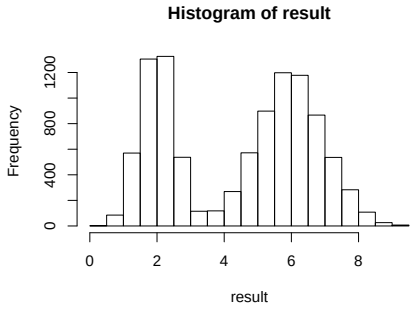

Figure 5.2: The histogram from MCMC sampling.

```
  prop <- theta + runif(1, -0.5, 0.5)
  if (runif(1) < f(prop)/f(theta)) theta <- prop
}
```

In the code, the initial value of the Metropolis Hastings Markov chain is 1. The proposal function is a uniform density on the interval from $\theta - 0.5$ to $\theta + 0.5$. Note how $q(\theta_{i-1} \mid \theta^*)/q(\theta^* \mid \theta_{i-1}) = 1$ for all θ_{i-1} and θ^*, so that this quotient disappears from the computation of a_{θ,θ^*}. The last value of the chain, for example 6.662043, is an approximate sample from the posterior. We can check this by running the code 10000 times and making a histogram:

```
result <- rep(1,10000)
for (i in 1:10000) {
  result[i] <- 1
  for (j in 1:1000) {
    prop <- result[i] + runif(1, -0.5, 0.5)
    if (runif(1) < f(prop)/f(result[i])) result[i] <- prop
  }
}
hist(result)
```

This produces Figure 5.2; the histogram is similar to the density in Figure 5.1. We can now go on and use the approximate sample in `result` to make inference. Let's say one would like to predict y_{new} where

$$\pi(y_{new} \mid \theta) = \text{Normal}(y_{new}; \theta, 1^2)$$

The code

```
hist(rnorm(10000,result, 1), probability = TRUE)
```

produces the histogram in Figure 5.3. In this toy example, the density of the predictive distribution can be computed directly, and it has been added to the figure with the code

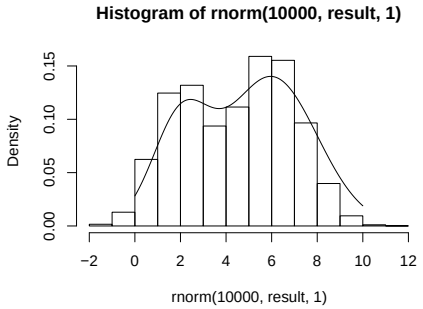

Figure 5.3: The predictive density: The histogram is produced from the MCMC sample while the curve is the theoretically correct density.

```
lines(x, 0.3*dnorm(x, 2, 0.5^2+1^2) + 0.7*dnorm(x, 6, 1^2+1^2))
```

Let us look at how well the simulated results approximate the correct ones. For example, for the probability of a prediction above 6,

```
1 - mean(pnorm(6, result, 1))
```

gives 0.3021683, while the exact result, computed with

```
1 - 0.3*pnorm(6, 2, 1.25) - 0.7*pnorm(6, 6, 2)
```

is 0.3502061. So the approximation is not very good. What can we do to improve it? Let us rewrite our MCMC code as

```
theta <- rep(1, 1000)
for (i in 2:1000) {
  prop <- theta[i-1] + runif(1, -0.5, 0.5)
  if (runif(1) < f(prop)/f(theta[i-1]))
    theta[i] <- prop
  else theta[i] <- theta[i-1]
}
plot(theta, type="l")
```

The main difference is that we now store all the θ values we generate along the way, so that we can plot them in *trace plots*. Running this code 4 times might give plots like those shown in Figure 5.4. Comparing with the density of Figure 5.1, remembering that the proposal distribution is uniform on the interval $[\theta-0.5, \theta+0.5]$, and remembering we always start our chain with the value $\theta = 1$, we see that sometimes the chain "gets stuck" in the leftmost density of the two normal densities. Sometimes it moves to the rightmost density. Sometimes it may move to the rightmost density and then back to the leftmost. But generally this happens quite rarely. In the long run, we know from the theoretical results about Metropolis Hastings that the chain will visit the two parts of the density

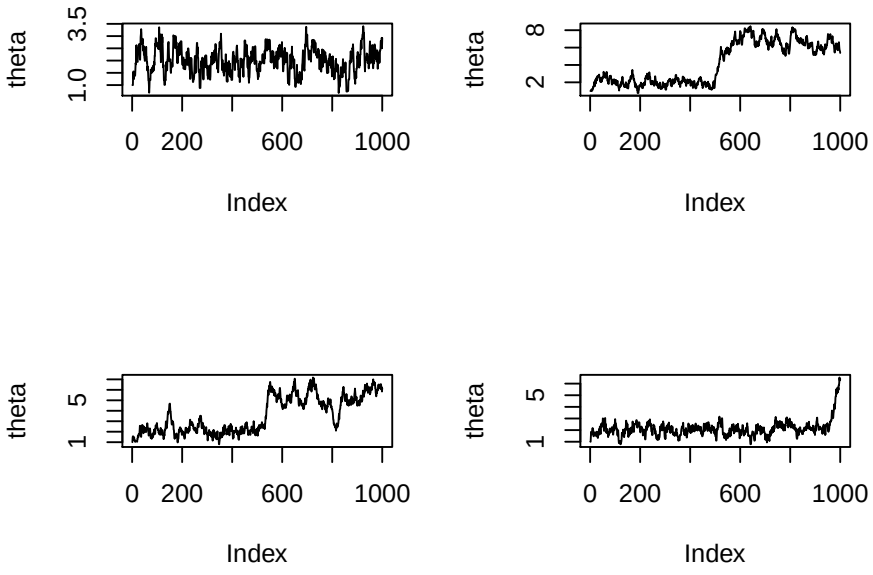

Figure 5.4: Four trace plots for θ.

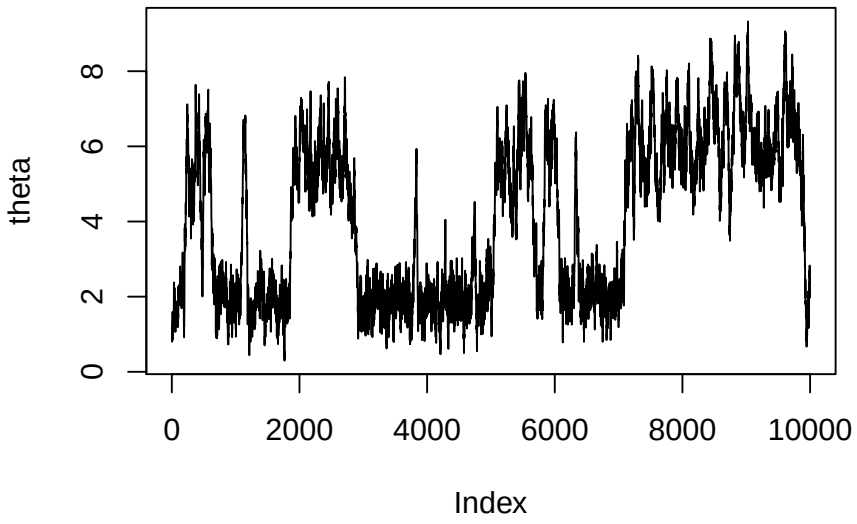

Figure 5.5: A longer trace plot for θ.

the right proportion of time, and the sample will be exactly from the density of Figure 5.1. However, clearly this will not happen in only 1000 steps; the final value of the chain will then to some extent be influenced by the starting value $\theta = 1$, which, by definition, means that it has not reached its limiting distribution. So in this case, the trace plots tell us that using 1000 steps is too few to reach the limiting distribution.

Let us rerun the code above using 10000 steps instead of 1000. Figure5.5 shows a trace plot. We can see that the two main parts of the distribution are now visited rougly in the right proportion of time. Indeed, if we rerun all our code with 10000 longer chains and estimate the probability that $y_{new} > 6$, we now get 0.3470309, which is a much closer approximatioin to the true value 0.3502061.

The problem now is that the code takes several minutes to run. How can we make it faster? A main problem is that we are generating each θ_i as the last value of an independently simulated chain. This makes the values $\theta_1, \theta_2, \ldots, \theta_m$ independent, but, as we saw in the beginning of this chapter, this is not necessary for Equation 5.2 to hold. We can instead use all the generated values in a single chain, except for some burn-in, in our predictions. Based on what we have seen above, we use a burn-in length of 1000. The code

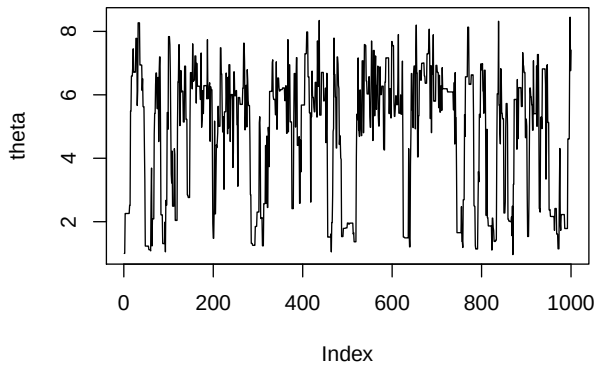

Figure 5.6: A trace plot for θ using a proposal function with larger steps.

```
result <- rep(1,1000000)
result[1] <- 1
for (i in 2:1000000) {
  prop <- result[i-1] + runif(1, -0.5, 0.5)
  if (runif(1) < f(prop)/f(result[i-1]))
    result[i] <- prop
  else result[i] <- result[i-1]
}
1 - mean(pnorm(6, result[1001:1000000], 1))
```

produces (for example) 0.3480443 in a much shorter time than the previous code with 10000 independent chains, even if we are now simulating the single chain for a much longer time.

5.4 Choosing the proposal function

Even with the improvements we made in the previous section, and even after simulating a Markov chain of length one million, the result is still not very accurate. And that was for a toy example. Generally, in order to obtain reasonable accuracy within reasonable computational time, one needs to choose the proposal function in such a way that convergence of the Metropolis Hastings Markov chain is fast. Choosing such a proposal function is not an easy task. In this section, we will simply continue the example from the previous section to illustrate some issues concerning proposal function choice.

In the example above, we saw in Figures 5.4 and 5.5 how the Markov chain moves fairly slowly between the two main parts of the density. What if we made

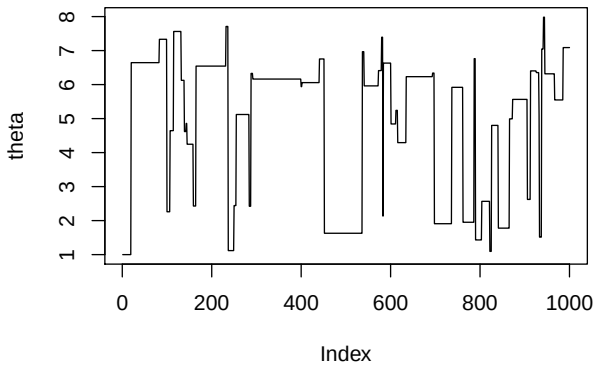

Figure 5.7: A trace plot for θ using a proposal function with too large steps.

it move faster? In other words, what if we used a proposal function with larger steps. If we re-run the code above using a proposal function that is uniform on the interval $[\theta - 3, \theta + 3]$ instead of on the interval $[\theta - 0.5, \theta + 0.5]$, we get a trace plot like Figure 5.6. We see that the chain now visits all parts of the density quite frequently; one says that it has good *mixing*. If we re-compute the probability that $y_{new} > 6$, using a single chain with length one million, we now get (for example) 0.3509797, and we are getting closer to the true value 0.3502061.

Encouraged by our results, we might think that "bigger is better", and go for a proposal function that is uniform on the interval $[\theta - 50, \theta + 50]$. A trace plot is showin in Figure 5.7. What happens here is that the acceptance probability $a_{\theta,\theta*}$ is generally much closer to zero. The reason is that new proposed values θ^* are often quite far from the main densities. With low acceptance probability, the chain is "stuck" at the same value for many steps. This again means that it moves around more slowly, i.e., it has bad mixing.

We may explore the effect of this bad mixing by computing the estimate for the probability that $y_{new} > 6$. Instead of just looking at a single estimated value, one should run the process many times and explore the variance in the estimates, compared to the variance in estimates using shorter step sizes.

It should be clear that there are many avenues to explore to obtain better mixing and better accuracy in estimates. Indeed, results in our toy example can be obtained much faster and with better accuracy than what we have achieved above. However, we leave the example at this point.

5.5 Advantages and disadvantages with MCMC for Bayesian inference

MCMC is a very flexible technique with a huge range of applications. A big advantage is that, for many models, it is quite simple and fast to program an implementation of Metropolis Hastings that works with reasonable accuracy, at least if the Markov chain is run long enough. A big disadvantage is that some experience and skill, and some understanding of the posterior you are trying to simulate from, may be needed in order to select a proposal function that leads to reasonably fast and accurate results. An even bigger disadvantage may be that, except for some limited cases, there are no general mathematical proofs about the accuracy of results.

Although the Metropolis Hastings algorithm is very flexible, and has several important special cases such as Gibbs sampling, it is not the last word in how one can generate approximate samples from posteriors. Finding imporved algorithms for such situations is still an active research field.

Many software packages exist which tries to make it easier to run MCMC even for those without expert knowledge of the relevant algorithms. A state-of-the-art package is Stan (https://mc-stan.org).

Finally, it should be mentioned that MCMC algorithms tend to be less effective when the dimension of θ increases above a few thousands. In modern machine learning applications, where parameter spaces have dimensions in the millions, other approaches, such as Variational Bayes, seem to be more effective.

Chapter 6

Some solutions to some Exercises

6.1 Exercises from Chapter 1

1. We want to compute the conditional probability that a person votes for B given that he or she lives in a rental flat. This is the quotient of the probability that the person lives in a rental flat and votes for B, divided by the probability that he or she lives in a rental flat:

$$\frac{0.03}{0.11 + 0.03 + 0.08 + 0.01} = 0.1304 = 13\%.$$

2. Let A denote that the person is affected, and P that the test is positive. Using Bayes formula, we get

$$\begin{aligned} \pi(A \mid P) &= \frac{\pi(P \mid A)\pi(A)}{\pi(P \mid A)\pi(A) + \pi(P \mid A^c)\pi(A^c)} = \frac{0.95 \cdot 0.007}{0.95 \cdot 0.007 + 0.05 \cdot 0.993} \\ &= 0.118 \approx 12\%. \end{aligned}$$

3. (a) A direct computation gives

$$\begin{aligned} \pi(\theta \mid \text{data}) &\propto_\theta \pi(\text{data} \mid \theta)\pi(\theta) \propto_\theta \text{Binomial}\,(9; 12, \theta) \\ &\propto_\theta \theta^9(1-\theta)^{12-9} \propto_\theta \text{Beta}\,(\theta; 10, 4)\,. \end{aligned}$$

   A more direct argument uses that the uniform distribution on $[0, 1]$ is identical to the Beta $(1, 1)$ distribution; with this the formulas of the lecture notes can be used to derive the posterior Beta $(1 + 9, 1 + 3)$ directly.

   (b) The posterior becomes Beta $(10 + 11, 4 + 19 - 11) =$ Beta $(21, 12)$.

   (c) The probability for success is the expectation of the Beta $(21, 12)$ density, which is $\frac{21}{21+12} = \frac{21}{33}$.

4. (a)
$$\binom{y_H + y_T}{y_H} (0.5 \cdot 0.3^{y_H} 0.7^{y_T} + 0.5 \cdot 0.7^{y_H} 0.3^{y_T})$$
(b)
$$\binom{y_H + y_T}{y_H} \left(0.5 \cdot 0.3^{y_H+1} 0.7^{y_T} + 0.5 \cdot 0.7^{y_H+1} 0.3^{y_T}\right)$$
(c) We get, as before
$$\frac{0.3^{y_H+1} 0.7^{y_T} + 0.7^{y_H+1} 0.3^{y_T}}{0.3^{y_H} 0.7^{y_T} + 0.7^{y_H} 0.3^{y_T}}$$

5. For example

```
> fn <- function(alpha) {(pbeta(0.4, alpha, alpha)-0.05)^2}
> optimize(fn, c(1, 1000))
$‘minimum‘
[1] 33.38651
```

6. Assume $p \sim \text{Beta}(\alpha_0, \beta_0)$. Then
$$\begin{aligned} \pi(p \mid y) &\propto_p \pi(y \mid p)\pi(p) \\ &\propto_p (1-p)^{\alpha} p^{y} \cdot p^{\alpha_0 - 1}(1-p)^{\beta_0 - 1} \\ &\propto_p p^{\alpha_0 + y - 1}(1-p)^{\beta_0 + \alpha - 1} \end{aligned}$$
so
$$p \mid y \sim \text{Beta}(\alpha_0 + y, \beta_0 + \alpha)$$
and we have proved conjugacy.

7. Assume that $\tau \sim \text{Gamma}(\alpha, \beta)$. Then
$$\begin{aligned} \pi(\tau \mid y) &\propto_\tau \pi(y \mid \tau)\pi(\tau) \\ &\propto_\tau \tau^{1/2} \exp\left(-\frac{\tau}{2}(y-\mu)^2\right) \tau^{\alpha-1} \exp(-\beta\tau) \\ &\propto_\tau \tau^{\alpha+1/2-1} \exp(-(\beta + \frac{1}{2}(y-\mu)^2)\tau) \end{aligned}$$
so
$$\tau \mid y \sim \text{Gamma}\left(\alpha + \frac{1}{2}, \beta + \frac{1}{2}(y-\mu)^2\right)$$
and we have proved conjugacy.

8. Writing
$$\begin{aligned} k \mid p &\sim \text{Binomial}(n, p) \\ p &\sim \text{Beta}(\alpha, \beta) \end{aligned}$$

we get

$$\begin{aligned}
\pi(k) &= \frac{\pi(k \mid p)\pi(p)}{\pi(p \mid k)} \\
&= \frac{\text{Binomial}(k; n, p) \cdot \text{Beta}(p; \alpha, \beta)}{\text{Beta}(\alpha + k, \beta + n - k)} \\
&= \frac{\binom{n}{k} p^k (1-p)^{n-k} \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}}{\frac{1}{B(\alpha+k,\beta+n-k)} p^{\alpha+k-1}(1-p)^{\beta+n-k-1}} \\
&= \frac{B(\alpha + k, \beta + n - k)}{B(\alpha, \beta)} \binom{n}{k}
\end{aligned}$$

This is a probability mass function on the set of integers $\{0, \ldots, n\}$ for all real numbers $\alpha > 0$, $\beta > 0$, and integers $n > 0$. In fact, it is called the Beta-Binomial distribution.

9. (a) Writing (for $i = 1, \ldots, 6$) c_i for the count of throws with outcome i among the throws $k_1, \ldots, k_n$, we get

$$\begin{aligned}
& \pi(\theta \mid k_1, \ldots, k_n) \\
\propto_\theta \quad & \pi(k_1, \ldots, k_n \mid \theta)\pi(\theta) \\
\propto_\theta \quad & \left(\frac{1}{3}\theta_1\right)^{c_1} \left(\frac{1}{3}\theta_2\right)^{c_2} \left(\frac{1}{3}\theta_3\right)^{c_3} \left(\frac{1}{3}(1-\theta_3)\right)^{c_4} \left(\frac{1}{3}(1-\theta_2)\right)^{c_5} \left(\frac{1}{3}(1-\theta_1)\right)^{c_6} \\
& \theta_1^{20-1}(1-\theta_1)^{20-1}\theta_2^{20-1}(1-\theta_2)^{20-1}\theta_3^{20-1}(1-\theta_3)^{20-1} \\
\propto_\theta \quad & \theta_1^{20+c_1-1}(1-\theta_1)^{20+c_6-1}\theta_2^{20+c_2-1}(1-\theta_2)^{20+c_5-1}\theta_3^{20+c_3-1}(1-\theta_3)^{20+c_4-1}.
\end{aligned}$$

Thus the posterior is

$$\text{Beta}(\theta_1; 20+c_1, 20+c_6) \cdot \text{Beta}(\theta_2; 20+c_2, 20+c_5) \cdot \text{Beta}(\theta_3; 20+c_3, 20+c_4)$$

(b) We have the counts

$$\begin{aligned}
c_1 &= 1 \\
c_2 &= 3 \\
c_3 &= 2 \\
c_4 &= 2 \\
c_5 &= 0 \\
c_6 &= 3
\end{aligned}$$

It is possible to use the predictive distribution found in Exercise 8 to answer the question, but the easiest approach may be to use the same thinking as in the start of Chapter 1. Let k_{n+1} be the outcome of the $(n+1)$'st throw. Then

$$\begin{aligned}
& \pi(k_{n+1} \mid k_1, \ldots, k_n) = \int \pi(k_{n+1} \mid \theta)\pi(\theta \mid k_1, \ldots, k_n)\, d\theta \\
= \quad & \begin{cases} \frac{1}{3}\int \theta_{k_{n+1}} \pi(\theta \mid k_1, \ldots, k_n)\, d\theta & k_{n+1} = 1, 2, 3 \\ \frac{1}{3}\left[1 - \int \theta_{7-k_{n+1}} \pi(\theta \mid k_1, \ldots, k_n)\, d\theta\right] & k_{n+1} = 4, 5, 6 \end{cases}
\end{aligned}$$

As we have the Expectations, for $i = 1, 2, 3$,

$$\begin{aligned} & \int \theta_i \pi(\theta \mid k_1, \ldots, k_n)\, d\theta = \mathrm{E}\left[\mathrm{Beta}(20 + c_i, 20 + c_{7-i})\right] \\ = & \frac{20 + c_i}{20 + c_i + 20 + c_{7-i}}, \end{aligned}$$

we get

$$\begin{aligned} \pi(k_{n+1} = 1 \mid k_1, \ldots, k_n) &= \frac{1}{3} \cdot \frac{20 + 1}{20 + 1 + 20 + 3} = 0.159 \\ \pi(k_{n+1} = 2 \mid k_1, \ldots, k_n) &= \frac{1}{3} \cdot \frac{20 + 3}{20 + 3 + 20 + 0} = 0.178 \\ \pi(k_{n+1} = 3 \mid k_1, \ldots, k_n) &= \frac{1}{3} \cdot \frac{20 + 2}{20 + 2 + 20 + 2} = 0.167 \\ \pi(k_{n+1} = 4 \mid k_1, \ldots, k_n) &= \frac{1}{3} \left(1 - \frac{20 + 2}{20 + 2 + 20 + 2}\right) = 0.167 \\ \pi(k_{n+1} = 5 \mid k_1, \ldots, k_n) &= \frac{1}{3} \left(1 - \frac{20 + 3}{20 + 3 + 20 + 0}\right) = 0.155 \\ \pi(k_{n+1} = 6 \mid k_1, \ldots, k_n) &= \frac{1}{3} \left(1 - \frac{20 + 1}{20 + 1 + 20 + 3}\right) = 0.174 \end{aligned}$$

10. (a) Assume we use a prior $\pi(\theta) = r(\theta)$ as defined in the exercise. For each $i = 1, \ldots, k$, define the prior predictive

$$r_i(y) = \int_\theta \pi(y \mid \theta) q_{\gamma_i}(\theta)\, d\theta = \frac{\pi(y \mid \theta) q_{\gamma_i}(\theta)}{q_{f(\gamma_i)}(\theta)}$$

and we can also write

$$r_i(y) q_{f(\gamma_i)}(\theta) = \pi(y \mid \theta) q_{\gamma_i}(\theta).$$

Then,

$$\begin{aligned} \pi(\theta \mid y) &\propto_\theta \pi(y \mid \theta)\pi(\theta) \\ &= \pi(y \mid \theta) \sum_{i=1}^{k} \lambda_i q_{\gamma_i}(\theta) \\ &= \sum_{i=1}^{k} \lambda_i \left[\pi(y \mid \theta) q_{\gamma_i}(\theta)\right] \\ &= \sum_{i=1}^{k} \lambda_i \left[r_i(y) q_{f(\gamma_i)}(\theta)\right] \end{aligned}$$

To get the actual posterior, we need to normalize this so that it integrates to 1 over θ. But the densities $q_{f(\gamma_i)}(\theta)$ all integrate to 1,

so we get

$$\pi(\theta \mid y) = \frac{\sum_{i=1}^{k} \lambda_i r_i(y) q_{f(\gamma_i)}(\theta)}{\sum_{i=1}^{k} \lambda_i r_i(y)}.$$

We see that this is a new density in the family defined in the exercise, with $\gamma_i' = f(\gamma_i)$ and

$$\lambda_i' = \frac{\lambda_i r_i(y)}{\sum_{j=1}^{k} \lambda_j r_j(y)}.$$

This proves conjugacy.

(b) We have

$$\pi(y) = \int_\theta \pi(y \mid \theta) \sum_{i=1}^{k} \lambda_i q_{\gamma_i}(\theta)\, d\theta = \sum_{i=1}^{k} \lambda_i \int_\theta r_i(y) q_{f(\gamma_i)}(\theta)\, d\theta = \sum_{i=1}^{k} \lambda_i r_i(y).$$

(c) Conjugate priors are very simple and practical to use, but they are not very flexible. For example, in situations where the normal family is conjugate, it may not always be reasonable to use a normal prior. However, weighted sums of normals is represents a much more flexible class of densities, and thus can often be used when a single normal prior cannot.

11. Knowing that the marginal distribution for X is normal, we only have to compute the expection and variance of this random variable to find its distribution. We get

$$\mathrm{E}(X) = \mathrm{E}_\mu\left(\mathrm{E}_{X|\mu}(X)\right) = \mathrm{E}_\mu(\mu) = \mu_0$$

and

$$\mathrm{Var}(X) = \mathrm{E}_\mu\left(\mathrm{Var}_{X|\mu}(X)\right) + \mathrm{Var}_\mu\left(\mathrm{E}_{X|\mu}(X)\right) = \mathrm{E}_\mu(\tau^{-1}) + \mathrm{Var}_\mu(\mu) = \tau^{-1} + \tau_0^{-1}.$$

Thus we have the prior predictive

$$x \sim \mathrm{Normal}(\mu_0, \tau^{-1} + \tau_0^{-1}).$$

6.2 Exercise from Chapter 2

6.3 Exercises from Chapter 3

1. (a) According to the formulas in this chapter, the poterior is

$$p \mid \text{data} \sim \mathrm{Dirichlet}(1+3, 1+9, 1+1) = \mathrm{Dirichlet}(4, 10, 2)$$

(b) We get

$$\mathrm{E}(p \mid \text{data}) = \frac{c(4,10,2)}{4+10+2} = \left(\frac{4}{16}, \frac{10}{16}, \frac{2}{16}\right)$$

(c) According to Equation 3.1 we get

$$\begin{aligned}\pi(x=(1,2,1)) &= \frac{4!}{1!2!1!}\cdot\frac{\Gamma(4+1)}{\Gamma(4)}\cdot\frac{\Gamma(10+2)}{\Gamma(10)}\cdot\frac{\Gamma(2+1)}{\Gamma(2)}\frac{\Gamma(16)}{\Gamma(16+4)}\\ &= 12\cdot 4\cdot 10\cdot 11\cdot 2\cdot\frac{1}{16\cdot 17\cdot 18\cdot 19}=0.1135\end{aligned}$$

2. (a)

(b) To compute the integral below, we use the change of variables $\theta_i = Ku_i$ for $i=1,\ldots,k-1$. Writing also $\theta_k = Ku_k$, we get $\sum_{i=1}^k u_i = 1$. Note that θ_k is completely determined by $\theta_1,\ldots,\theta_{k-1}$, so the density is $(k-1)$-dimensional. We get

$$\begin{aligned}&\int \frac{1}{K^{\alpha_1+\cdots+\alpha_k-1}}\frac{\Gamma(\alpha_1+\cdots+\alpha_k)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_k)}\theta_1^{\alpha_1-1}\cdots\theta_k^{\alpha_k-1}\,d\theta_1\ldots d\theta_{k-1}\\ =&\int \frac{1}{K^{\alpha_1+\cdots+\alpha_k-1}}\frac{\Gamma(\alpha_1+\cdots+\alpha_k)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_k)}(Ku_1)^{\alpha_1-1}\cdots(Ku_k)^{\alpha_k-1}K^{k-1}\,du_1\ldots du_{k-1}\\ =&\int \frac{\Gamma(\alpha_1+\cdots+\alpha_k)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_k)}u_1^{\alpha_1-1}\cdots u_k^{\alpha_k-1}\,du_1\ldots du_{k-1}\\ =&\ 1\end{aligned}$$

where in the last step we use that the standard Dirichlet density integrates to 1.

(c) We get

$$\pi(\theta_1,\ldots,\theta_{i-1}\mid\theta_i,\ldots,\theta_k)\propto_{\theta_1,\ldots,\theta_{i-1}}\pi(\theta_1,\ldots,\theta_k)\propto_{\theta_1,\ldots,\theta_{i-1}}\theta_1^{\alpha_1-1}\cdots\theta_{i-1}^{\alpha_{i-1}-1}$$

Now, $\theta_1+\cdots+\theta_{i-1}=1-\theta_i-\cdots-\theta_k$. Comparing with the densities defined in (b), we get that

$$\theta_1\ldots,\theta_{i-1}\mid\theta_i,\ldots,\theta_k\sim\text{Dirichlet}_{1-\theta_i-\cdots-\theta_k}(\alpha_1,\ldots,\alpha_{i-1}).$$

(d) We get

$$\begin{aligned}\pi(\theta_1,\ldots,\theta_i) &= \frac{\pi(\theta_1,\ldots,\theta_k)}{\pi(\theta_{i+1},\ldots,\theta_k\mid\theta_1,\ldots,\theta_i)}\\ &= \frac{\text{Dirichlet}((\theta_1,\ldots,\theta_k);(\alpha_1,\ldots,\alpha_k))}{\text{Dirichlet}_{1-\theta_1-\cdots-\theta_i}((\theta_{i+1},\ldots,\theta_k);(\alpha_{i+1},\ldots,\alpha_k))}\\ &\propto_{\theta_1,\ldots,\theta_i} \frac{\theta_1^{\alpha_1-1}\cdots\theta_i^{\alpha_i-1}}{1/(1-\theta_1-\cdots-\theta_i)^{\alpha_{i+1}+\cdots+\alpha_k-1}}\\ &= \theta_1^{\alpha_1-1}\cdots\theta_i^{\alpha_i-1}(1-\theta_1-\cdots-\theta_i)^{\alpha_{i+1}+\cdots+\alpha_k-1}\end{aligned}$$

(e) When $(\theta_1,\ldots,\theta_k)\sim\text{Dirichlet}(\alpha_1,\ldots,\alpha_k)$, we have that $\theta_{i+1}+\cdots+\theta_k$ is completely determined by $(\theta_1,\ldots,\theta_i)$: It is equal to 1 minus the sum of these numbers. Thus the density for the vector

$(\theta_1, \ldots, \theta_i, \theta_{i+1} + \cdots + \theta_k)$ is equal to the density found in (d). From this, and the fact that $\theta_1 + \cdots + \theta_i + (\theta_{i+1} + \cdots + \theta_k) = 1$, we can read off that

$$\theta_1, \ldots, \theta_i, \theta_{i+1} + \cdots + \theta_k \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_i, \alpha_{i+1} + \cdots + \alpha_k)$$

(f) Applying the result from (e) twice, we get that

$$(\theta_1 + \cdots + \theta_i, \theta_{i+1} + \cdots + \theta_k) \sim \text{Dirichlet}(\alpha_1 + \cdots + \alpha_i, \alpha_{i+1} + \cdots + \alpha_k)$$

Using the result from (a) we see that

$$\theta_1 + \ldots \theta_i \sim \text{Beta}(\alpha_1 + \cdots + \alpha_i, \alpha_{i+1} + \cdots + \alpha_k)$$

3. (a)

$$\hat{P} = \begin{bmatrix} 2/5 & 3/5 & 0 & 0 \\ 3/8 & 1/8 & 4/8 & 0 \\ 0 & 4/8 & 1/8 & 3/8 \\ 0 & 0 & 3/4 & 1/4 \end{bmatrix}.$$

(b)

$$\text{E}[P \mid \text{data}] = \begin{bmatrix} 3/9 & 4/9 & 1/9 & 1/9 \\ 4/12 & 2/12 & 5/12 & 1/12 \\ 1/12 & 5/12 & 2/12 & 4/12 \\ 1/8 & 1/8 & 4/8 & 2/8 \end{bmatrix}.$$

(c) We have

$$P_1 \mid \text{data} \sim \text{Dirichlet}(3, 4, 1, 1).$$

According to Exercise 2f, we thus get

$$P_{11} \mid \text{data} \sim \text{Beta}(3, 6).$$

And `pbeta(0.3, 3, 6, lower.tail=FALSE)` produces 0.5517738.

(d) Choosing to keep pseudocounts equal to 1 for the states that are possible, we get

$$\pi(P) = \text{Dirichlet}(P_1; (1, 1, 0, 0))\,\text{Dirichlet}(P_2; (1, 1, 1, 0))\,\text{Dirichlet}(P_3; (0, 1, 1, 1))\,\text{Dirichlet}(0, 0, 1, 1)$$

With this prior, the expectation of the posterior for P becomes

$$\text{E}[P \mid \text{data}] = \begin{bmatrix} 3/7 & 4/7 & 0 & 0 \\ 4/11 & 2/11 & 5/11 & 0 \\ 0 & 5/11 & 2/11 & 4/11 \\ 0 & 0 & 4/6 & 2/6 \end{bmatrix}$$

and the command `pbeta(0.3, 3, 4, lower.tail=FALSE)` produces 0.74431.

6.4 Exercise from Chapter 4

1. (a) Beta (α, β) is the conjugate family: If $y \sim$ Geometric (p) and $p \sim$ Beta (α, β), then it is easy to show that $p \mid y \sim$ Beta $(\alpha + 1, \beta + y)$.

   (b) The Uniform$(0, 1)$ prior is the same as the Beta $(1, 1)$ prior. In the data, there is a total of 342 cases and a total of $S = 347$ offspring. Thus the posterior becomes Beta $(1 + 342, 1 + 347) =$ Beta $(343, 348)$.

   (c) The Geometric distribution with parameter p has expectation $(1 - p)/p$ and $(1 - p)/p \geq 1$ is the same as $p < 1/2$. Using the R code `pbeta(0.5, 343, 348)` results in 0.5754726.

Chapter 7

Appendix: List of some probability distributions

The Bernoulli distribution

If $x \in \{0, 1\}$ has a Bernoulli(p) distribution, with $0 \leq p \leq 1$, then the probability mass function is

$$\pi(x) = p^x (1-p)^{1-x}.$$

R: Use the Binomial with sample size 1.

The Beta distribution

If $x \geq 0$ has a Beta(α, β) distribution with $\alpha > 0$ and $\beta > 0$ then the density is

$$\pi(x \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}.$$

`R: dbeta, pbeta, qbeta, rbeta`

The Beta-binomial distribution

If $x \in \{0, 1, 2 \ldots, n\}$ has a Beta-binomial(n, α, β) distribution with n a positive integer, $\alpha > 0$, and $\beta > 0$, then the probability mass function is

$$\pi(x \mid n, \alpha, \beta) = \binom{n}{x} \frac{B(x+\alpha, n-x+\beta)}{B(\alpha, \beta)}$$

where B is the Beta function.

The Binomial distribution

If $x \in \{0, 1, 2, \ldots, n\}$ has a Binomial(n, p) distribution, with n a positive integer and $0 \leq p \leq 1$, then the probability mass function is

$$\pi(x \mid n, p) = \binom{n}{x} p^x (1-p)^{n-x}.$$

`R: dbinom, pbinom, qbinom, rbinom`

The Cauchy distribution

If $x \geq 0$ has a Cauchy(μ, γ) distribution, with $\gamma > 0$, then the probability density is

$$\pi(x \mid \mu, \gamma) = \frac{1}{\pi\gamma\left(1 + \left(\frac{x-\mu}{\gamma}\right)^2\right)}.$$

The standard Cauchy distribution with $\mu = 0$ is the t-distribution with $\nu = 1$.

The Dirichlet distribution

A vector $\theta = (\theta_1, \ldots, \theta_k)$ of non-negative real numbers satisfying $\sum_{i=1}^k \theta_i = 1$ has a Dirichlet $(\alpha_1, \ldots, \alpha_k)$ distribution with parameter vector $\alpha = (\alpha_1, \ldots, \alpha_k)$, with each $\alpha_i > 0$, if it has probability density function

$$\pi(\theta \mid \alpha) = \frac{\Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_k)}{\Gamma(\alpha_1)\Gamma(\alpha_2) \cdot \Gamma(\alpha_k)} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \cdots \theta_k^{\alpha_k - 1}.$$

If θ has the distribution above, the expectation is the vector $\frac{\alpha}{\sum_{i=1}^k \alpha_i}$.

The Exponential distribution

If $x \geq 0$ has an Exponential(λ) distribution with $\lambda > 0$ as parameter, then the density is

$$\pi(x \mid \lambda) = \lambda \exp(-\lambda x)$$

and the cumulative distribution function is

$$F(x) = 1 - \exp(-\lambda x).$$

`R: dexp, pexp, qexp, rexp`

The Gamma distribution

If $x > 0$ has a Gamma(α, β) distribution, with $\alpha > 0$ and $\beta > 0$, then the density is

$$\pi(x \mid \alpha\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha - 1} \exp(-\beta x).$$

The expectation and variance are α/β and α/β^2, respectively, while the mode is $(\alpha - 1)/\beta$ (when $\alpha \geq 1$). `R: dgamma, pgamma, qgamma, rgamma`

The Geometric distribution

If the non-negative integer x has a Geometric distribution with parameter $p \in [0, 1]$, its probability mass function is given by

$$\pi(x \mid p) = (1-p)^x p.$$

`R: dgeom, pgeom, qgeom, rgeom`

The Multinomial distribution

A vector $x = (x_1, \ldots, x_k)$ of non-negative integers satisfying $\sum_{i=1}^k x_i = n$ has a Multinomial $(n, p_1, \ldots, p_k)$ distribution with parameters n and $p = (p_1, \ldots, p_k)$, where $n > 0$ is an integer and $p_i \geq 0$ and $\sum_{i=1}^k p_i = 1$, if the probability mass function is given by

$$\pi(x \mid n, p) = \frac{n!}{x_1! x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}.$$

The Negative Binomial distribution

A stochastic variable x taking on as possible values any positive integer has a Negative Binomial distribution if its probability mass function is given by

$$\pi(x \mid r, p) = \binom{x+r-1}{x} \cdot (1-p)^r p^x = \frac{\Gamma(x+r)}{\Gamma(x+1)\Gamma(r)} (1-p)^x p^r$$

where $r > 0$ and $p \in (0, 1)$ are parameters. (NOTE: The definition has now been updated to conform with the definition used for example in R) `R: dnbinom, pnbinom, qnbinom, rnbinom`

The Normal distribution

If the real x has a Normal distribution with parameters μ and σ^2, its density is given by

$$\pi(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right).$$

`R: dnorm, pnorm, qnorm, rnorm`

The Pareto distribution

If the real number $x \in [M, \infty)$ has a Pareto(M, α) distribution with parameters $M > 0$ and $\alpha > 0$, its density on this interval is given by

$$\pi(x \mid M, \alpha) = \alpha M^\alpha x^{-(\alpha+1)}$$

The Poisson distribution

If the nonnegative integer x has a Poisson(λ) distribution with parameter $\lambda > 0$, its probability mass function is given by

$$\pi(x \mid \lambda) = e^{-\lambda} \frac{\lambda^x}{x!}$$

`R: dpois, ppois, qpois, rpois`

The t-distribution

If the real number x has a t(ν) distribution with parameter $\nu > 0$, its density is

$$\pi(x \mid \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}.$$

`R: dt, pt, qt, rt`

The Uniform distribution

If $x \in [a, b]$ has a Uniform(a, b) distribution with $b > a$, then the density is given by

$$\pi(x \mid a, b) = \frac{1}{b-a}.$$

Chapter 8

Appendix: List of some conjugacies

Note: More conjucacies can be found on the Wikipedia page "Conjugate priors".

The Beta Binomial conjugacy

Likelihood: $x \sim \text{Binomial}\,(n, \theta)$
Prior: $\theta \sim \text{Beta}\,(\alpha, \beta)$
Posterior: $\theta \mid x \sim \text{Beta}\,(\alpha + x, \beta + n - x)$
Prior predictive: $x \sim \text{Beta-binomial}(n, \alpha, \beta)$

The Exponential Gamma conjugacy

Likelihood: $x \sim \text{Exponential}\,(\theta)$
Prior: $\theta \sim \text{Gamma}(\alpha, \beta)$
Posterior: $\theta \mid x \sim \text{Gamma}(\alpha + 1, \beta + x)$

The Multinomial Dirichlet conjugacy

Likelihood: $x = (x_1, \ldots, x_k) \sim \text{Multinomial}\,(n, \theta_1, \ldots, \theta_k)$
Prior: $\theta = (\theta_1, \ldots, \theta_k) \sim \text{Dirichlet}\,(\alpha_1, \ldots, \alpha_k)$
Posterior: $\theta \mid x \sim \text{Dirichlet}\,(\alpha_1 + x_1, \ldots, \alpha_k + x_k)$

The Poisson Gamma conjucacy

Likelihood: $x \sim \text{Poisson}\,(\theta)$
Prior: $\theta \sim \text{Gamma}(\alpha, \beta)$
Posterior: $\theta \mid x \sim \text{Gamma}(\alpha + x, \beta + 1)$
Prior predictive: $x \sim \text{Negative-Binomial}\,(\alpha, \beta/(1 + \beta))$ (NOTE: This formula has now been updated to conform with the definition of the Negative Binomial used for example in R).

The Normal-Gamma conjucacy

Likelihood: $x \sim \text{Normal}\left(\mu, \theta^{-1}\right)$
Prior: $\theta \sim \text{Gamma}(\alpha, \beta)$
Posterior: $\theta \mid x \sim \text{Gamma}\left(\alpha + \frac{1}{2}, \beta + \frac{1}{2}(x-\mu)^2\right)$

The Normal-Normal conjugacy

Likelihood: $x \sim \text{Normal}\left(\theta, \tau^{-1}\right)$
Prior: $\theta \sim \text{Normal}\left(\mu, \tau_0^{-1}\right)$
Posterior: $\theta \mid x \sim \text{Normal}\left(\frac{\tau x + \tau_0 \mu}{\tau + \tau_0}, \frac{1}{\tau + \tau_0}\right)$
Prior predictive: $x \sim \text{Normal}\left(\mu, \tau^{-1} + \tau_0^{-1}\right)$
Computations:

$$
\begin{aligned}
\pi(\theta \mid x) &\propto_\theta \pi(x \mid \theta)\pi(\theta) \\
&\propto_\theta \exp\left(-\frac{\tau}{2}(x-\theta)^2\right)\exp\left(-\frac{\tau_0}{2}(\theta-\mu)^2\right) \\
&= \exp\left(-\frac{1}{2}\left[\tau x^2 - 2\tau x\theta + \tau\theta^2 + \tau_0\theta^2 - 2\tau_0\theta\mu + \tau_0\mu^2\right]\right) \\
&\propto_\theta \exp\left(-\frac{1}{2}\left[(\tau+\tau_0)\theta^2 - 2(\tau x + \tau_0\mu)\theta\right]\right) \\
&\propto_\theta \exp\left(-\frac{1}{2}(\tau+\tau_0)\left(\theta - \frac{\tau x + \tau_0\mu}{\tau+\tau_0}\right)^2\right) \\
&\propto_\theta \text{Normal}\left(\theta; \frac{\tau x + \tau_0\mu}{\tau+\tau_0}, \frac{1}{\tau+\tau_0}\right)
\end{aligned}
$$