Chapter 6

The devil is in the denominator

6.1 Too many coin flips

Suppose we flip two coins. Each coin i is either fair ($Pr(H) = \theta_i = 0.5$) or biased towards heads ($Pr(H) = \theta_i = 0.9$) however, we cannot visibly detect the coin's nature. Suppose we flip both coins twice and record each result.

Problem 6.1.1. Suppose that we specify a discrete uniform prior on both θ_1 and θ_2. Find the joint distribution of the data and the coins' identity.

Denote the result of coin 1's flips by X_1 and X_2, where $X_i = 1$ is heads. Similarly for coin 2 except we use Y_i to denote the result on the ith flip. We can then write down the joint distribution as follows,

$$\begin{aligned} Pr(X_1, X_2, Y_1, Y_2, \theta_1, \theta_2) &= Pr(X_1, X_2|\theta_1)Pr(Y_1, Y_2|\theta_2)Pr(\theta_1)Pr(\theta_2) && (6.1)\\ &= \theta_1^{X_1+X_2}(1-\theta_1)^{2-X_1-X_2}\theta_2^{Y_1+Y_2}(1-\theta_2)^{2-Y_1-Y_2}0.5^2 && (6.2) \end{aligned}$$

Problem 6.1.2. Show that the above distribution is a valid probability distribution.

It trivially satisfies $Pr(.) \geq 0$ and so we only need to show that the sum of all values is 1 by summing over all the $2^4 = 16$ possible parameter combinations. This can be done by exhaustively checking every term but the symmetry of the problem means there are probably easier ways to proceed here.

Problem 6.1.3. We flip each coin twice and obtain for coin 1 $\{HH\}$ and coin 2 $\{HT\}$. Assuming that the result of each coin flip is independent of the previous result write down a likelihood function.

Since the coin flips are independent we can write down the overall likelihood by multiplying together the individual ones,

$$Pr(\{HH\},\{HT\}|\theta_1,\theta_2) = \theta_1^2\theta_2(1-\theta_2) \tag{6.3}$$

Problem 6.1.4. What are the maximum likelihood estimators of each parameter?

We can consider each parameter separately due to the independence of the problem. For coin 1, clearly $\hat{\theta}_1 = 0.9$ since this maximises θ_1^2; for coin 2 $\hat{\theta}_2 = 0.5$ since this maximises $\theta_2(1-\theta_2)$.

Problem 6.1.5. Calculate the marginal likelihood of the data (that is, the denominator of Bayes' rule).

This can be obtained by marginalising out all θ dependence in the joint distribution,

$$\begin{aligned}
Pr(HH,HT) &= \sum_{\theta_1}\sum_{\theta_2} Pr(HH,HT|\theta_1,\theta_2)Pr(\theta_1)Pr(\theta_2) &(6.4)\\
&= \sum_{\theta_1} Pr(HH|\theta_1)Pr(\theta_1)\sum_{\theta_2} Pr(HT|\theta_2)Pr(\theta_2) &(6.5)\\
&= \frac{1}{4}(\frac{1}{4}+\frac{81}{100})(\frac{1}{4}+\frac{9}{10}\frac{1}{10}) &(6.6)\\
&= \frac{901}{10000} &(6.7)
\end{aligned}$$

Problem 6.1.6. Hence calculate the posterior distribution, and demonstrate that this is a valid probability distribution.

Using Bayes' rule we have that,

$$\begin{aligned}
Pr(\theta_1,\theta_2|HH,HT) &= \frac{\theta_1^2\theta_2(1-\theta_2)\frac{1}{4}}{\frac{901}{10000}} &(6.8)\\
&= \frac{2500}{901}\theta_1^2\theta_2(1-\theta_2). &(6.9)
\end{aligned}$$

Summing the above over both θ_1 and θ_2,

$$\begin{aligned}
\sum_{\theta_1}\sum_{\theta_2}\frac{2500}{2809}\theta_1^2\theta_2(1-\theta_2) &= \frac{2500}{901}\sum_{\theta_1}\theta_1^2\sum_{\theta_2}\theta_2(1-\theta_2) &(6.10)\\
&= \frac{2500}{901}\left(0.5^2+0.9^2\right)(0.5\times 0.5+0.9\times 0.1) &(6.11)\\
&= 1. &(6.12)
\end{aligned}$$

Problem 6.1.7. Find the posterior mean of θ_1. What does this signify?

θ_1	θ_2	Z	$Pr(HT, HH, \theta_1, \theta_2 \vert Z)$
0.5	0.5	0	0.000625
0.5	0.5	1	0.04
0.5	0.9	0	0.002025
0.5	0.9	1	0.0036
0.9	0.5	0	0.018225
0.9	0.5	1	0.0324
0.9	0.9	0	0.059049
0.9	0.9	1	0.002916

Table 6.1: The likelihood function for the dependent coin flip example.

$$\begin{aligned}
\mathbb{E}(\theta_1|HH, HT) &= \mathbb{E}(\theta_1|HH) && (6.13)\\
&= \sum_{\theta_1} \theta_1 \frac{\theta_1^2}{0.5^2 + 0.9^2} && (6.14)\\
&= \frac{1}{0.5^2 + 0.9^2}(0.5^3 + 0.9^3) && (6.15)\\
&\approx 0.81 && (6.16)
\end{aligned}$$

This is a bit tricky to interpret since $\theta_1 \in \{0.5, 0.9\}$. However it basically means that there is greater weight towards $\theta_1 = 0.9$.

Problem 6.1.8. Now suppose that away from our view a third coin is flipped, and denote $Z = 1$ for a heads. The result of this coin affects the bias of the other two coins that are flipped subsequently so that,

$$Pr(\theta_i = 0.5|Z) = 0.8^Z 0.1^{1-Z} \tag{6.17}$$

Suppose we again obtain for coin 1 $\{HH\}$ and coin 2 $\{HT\}$. Find the maximum likelihood estimators (θ_1, θ_2, Z). How do the inferred biases of coin 1 and coin 2 compare to the previous estimates?

To do this we enumerate over the 8 possible combinations of θ_1, θ_2 and Z (see Table 6.1). From this it is evident that the maximum likelihood estimators are $(\hat{\theta}_1, \hat{\theta}_2, \hat{Z}) = (0.9, 0.9, 0)$. So here we have that coin 2's ML bias has changed from 0.5 to 0.9. Intuitively this is because there is a strong penalty for the coin 2 being fair if coin 1 is not, because of the dependence structure.

Problem 6.1.9. Calculate the marginal likelihood for the coin if we suppose that we specify a discrete uniform prior on Z, i.e. $Pr(Z = 1) = 0.5$.

To do this we simply multiply all the calculated likelihoods from Table 6.1 by 0.5 and sum them, to obtain, $Pr(HH, HT) = 0.0794$.

Problem 6.1.10. Suppose we believe that the independent coin flip model (where there is no third coin) and the dependent coin flip model (where the outcome of the third coin affects the biases of the two coins) are equally likely *a priori*. Which of the two models do we prefer?

Basically we want to compare,

$$\frac{Pr(\text{independent M}|HH,HT)}{Pr(\text{dependent M}|HH,HT)} = \frac{Pr(HH,HT|\text{independent M})}{Pr(HH,HT|\text{dependent M})} \times \underbrace{\frac{Pr(\text{independent M})}{Pr(\text{dependent M})}}_{1} \tag{6.18}$$

$$= \frac{0.0901}{0.0794} \tag{6.19}$$

$$\approx 1.13 \tag{6.20}$$

so using this basic test we prefer the independent flips model.

6.2 Coins combined

Suppose that we flip two coins, each of which has $Pr(H) = \theta_i$ where $i \in \{1, 2\}$, which is unknown. If their outcome is both the same then we regard this as a success; otherwise a failure. We repeatedly flip both coins (a single trial) and record whether the outcome is a success or failure. We do not record the result of flipping each coin. Suppose we model the number of failures, X, we have to undergo to attain n successes.

Problem 6.2.1. Stating any assumptions that you make specify a suitable probability model here.

The negative binomial fits this description perfectly. However we need to modify it to allow the probability of success to be a function of both coins' biases $p = \theta_1\theta_2 + (1-\theta_1)(1-\theta_2)$. This means we can write down the pmf,

$$Pr(X|n,\theta_1,\theta_2) = \begin{cases} \binom{n+X-1}{n-1}((1-\theta_1)(1-\theta_2)+\theta_1\theta_2)^n(1-(1-\theta_1)(1-\theta_2)-\theta_1-\theta_2)^X & X \geq 0 \\ 0 & \text{True} \end{cases} \tag{6.21}$$

This assumes that the flips of each coin are independent.

Problem 6.2.2. We obtain the data in `denominator_NBCoins.csv` for the number of failures to wait before 5 successes occur. Suppose that we specify the following priors $\theta_1 \sim U(0, 1)$ and $\theta_2 \sim U(0, 1)$. Calculate the denominator of Bayes' rule. (Hint: use a numerical integration routine.)

This requires us to do the following integral,

$$\int_0^1 \int_0^1 NB(X|n,\theta_1,\theta_2)\mathrm{d}\theta_1\mathrm{d}\theta_2 \approx 2.48731 \times 10^{-170}. \tag{6.22}$$

I carried out the above using Mathematica's 'NIntegrate' function which took around three seconds.

Problem 6.2.3. Draw a contour plot of the posterior. Why does the posterior have this shape?

The posterior is shown in Figure 6.1. There is a thin band of probability mass associated with the lines $\theta_1\theta_2 = const$ or $(1-\theta_1)(1-\theta_2) = const$. The mass is mostly associated in the upper left and bottom right because the data has relatively long runs before 5 successes occur, meaning that the same values for the parameters are not likely.

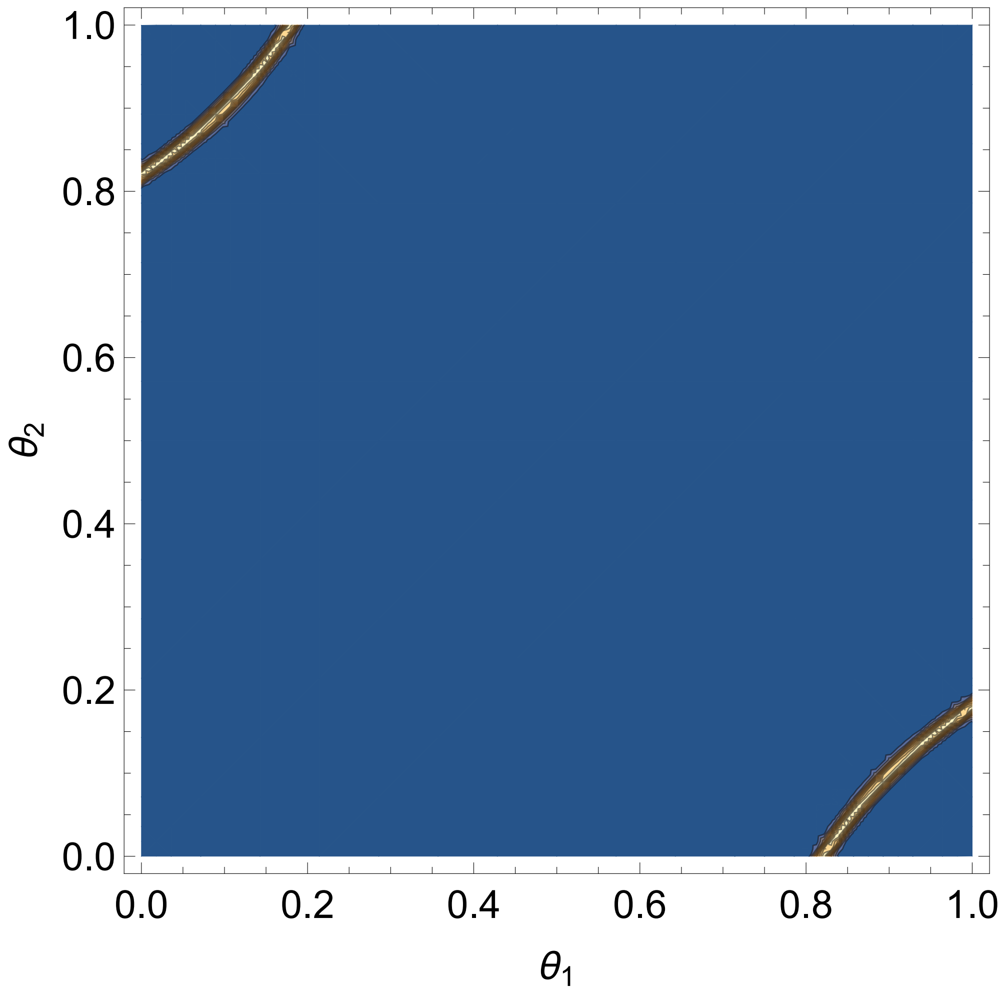

Figure 6.1: The posterior for the negative binomial coins example.

Problem 6.2.4. Comment on any issues with parameter identification for this model and how this might be rectified.

Clearly the model cannot differentiate between θ_1 and θ_2 because we have provided no further information on these. A solution would be to use a prior that assigns a strong weight to high/low values of one of the parameters. This isn't an issue that can be solved by collecting more data, unless we could see the identities of each coin.

Problem 6.2.5. Now suppose that we have three coins instead of two. Here we regard a success as all three coins showing the same result. Using the same data as before attempt to calculate the denominator term. Why is there a problem?

Even with just three dimensions even sophisticated deterministic routines struggle. After a few minutes Mathematica gave me this result,

$$\int_0^1\int_0^1\int_0^1 NB(X|5,\theta_1,\theta_2,\theta_3)\mathrm{d}\theta_1\mathrm{d}\theta_2\theta_3 \approx 3.64959\times 10^{-169} \tag{6.23}$$

Clearly with higher dimensions evaluating these integrals is going to be just too hard to attempt!

Problem 6.2.6. Assuming a denominator term equal to 3.64959×10^{-169} estimate the posterior mean of θ_1.

Again we run into problems but now with a different integral,

$$\mathbb{E}(\theta_1) = \int_0^1\int_0^1\int_0^1 \theta_1 NB(X|5,\theta_1,\theta_2,\theta_3)\frac{1}{3.64959\times 10^{-169}}\mathrm{d}\theta_1\mathrm{d}\theta_2\theta_3 \approx 0.500. \tag{6.24}$$

The above took around three minutes on Mathematica 11 on my laptop. The moral of the story is that even if we have the denominator there are still issues with integrating!

Bibliography