Bayesian course - problem set 3 (lecture 4)

Ben Lambert

November 14, 2016

1 Ticked off

Imagine once again that you are investigating the occurrence of Lyme disease in the UK. This is a vector-borne disease caused by bacteria of species *Borrelia* which is carried by ticks. (The ticks pick up the infection by blood-feeding on animals/humans that are infected with *Borrelia*.) As such, you decide to estimate the prevalence of this bacteria in ticks you collect from the grasslands and woodlands around Oxford.

As previously, you decide to use sample sizes of 100 ticks, out of which you count the number of ticks testing positive for *Borrelia*. You decide to use a binomial likelihood since you assume that the presence of *Borrelia* in one tick is independent of that in other ticks. Also because you sample a relatively small area you assume that the presence of *Borrelia* can be assumed to be identically-distributed across ticks.

Problem 1.1 *Suppose you choose a prior to be a Beta(1,1) distribution. Use sampling to estimate the prior predictive distribution. What is the mean of this distribution?*

Problem 1.2 *In a single sample you find that there are 6 ticks that test positive for Borrelia. Assuming a Beta(1,1) prior analytically calculate the posterior distribution. (Hint: by analytically here I mean look up the result on Google/in the lecture notes.) Graph this distribution.*

Problem 1.3 *Generate 100 independent samples from this distribution using your software's inbuilt (pseudo-)random number generator. Graph this distribution. How does it compare to the pdf of the exact posterior? (Hint: in R the command is "rbeta"; in Matlab it is "betarnd"; in Mathematica it is "RandomVariate[BetaDistribution...]"; in Python it is "numpy.random.beta".)*

Problem 1.4 *Evaluate the effect of increasing the sample size for your independent sampler on the estimate of the mean of the distribution. (Hint: for each sample you are essentially comparing the sample mean with the true mean of the posterior.)*

Problem 1.5 *Estimate the variance of the posterior using independent sampling for a sample size of 100. How does your sample estimate compare with the exact solution?*

Problem 1.6 *Create a proposal function for this problem that takes as input a current value of θ, along with a step size, and outputs a proposed value. For a proposal distribution here we use a normal distribution centred on the current θ value with a standard deviation (step size) of 0.1. This means you will need to generate a random θ from a normal distribution using your statistical software's inbuilt random number generator. (Hint: the only slight modification you need to make here is to ensure that we don't get $\theta < 0$ or $\theta > 1$ is to use periodic boundary conditions. To do this we use modular arithmetic. In particular we set $\theta_{proposed} = mod(\theta_{proposed}, 1)$. The command for this in R is $x\%\%1$; in Matlab the command is $mod(x, 1)$; in Mathematica it is $Mod[x, 1]$; in Python it is $x\%1$.)*

Problem 1.7 *Create the "accept/reject" function of Random Walk Metropolis that accepts as input $\theta_{current}$ and $\theta_{proposed}$ and outputs the next value of θ. This is done based on a ratio:*

$$r = \frac{p(X|\theta_{proposed}p(\theta_{proposed})}{p(X|\theta_{current}p(\theta_{current})} \tag{1}$$

and a uniformly-distributed random number between 0 and 1, which we call a. If $r > a$ then we update our current value of $\theta_{current} \rightarrow \theta_{proposed}$; alternatively we remain at $\theta_{current}$.

Problem 1.8 *Create a function that is a combined version of the previous two functions; so it takes as input a current value of $\theta_{current}$, generates a proposed $\theta_{proposed}$, and updates $\theta_{current}$ in accordance with the Metropolis accept/reject rule.*

Problem 1.9 *Create a full-working Random Walk Metropolis sampler! (Hint: you will need to iterate the last function repeatedly. As such, you will need to decide on a starting position for θ. I would recommend that you use a uniformly-distributed random number between 0 and 1.)*

Problem 1.10 *For a sample size of 100 from your Metropolis sampler compare the sampling distribution to the exact posterior. How does the estimated posterior compare with that obtained via independent sampling using the same sample size?*

Problem 1.11 *Run 1000 iterations, where in each iteration you run a single chain for 100 iterations. Store the results in a 1000 x 100 matrix. For each iterate calculate the sample mean. Graph the resultant distribution of sample means. How does MCMC do at estimating the posterior mean?*

Problem 1.12 *Graph the distribution of the sample mean estimates of the for the second 50 observations of each chain. How does this result compare with that of the previous question? Why is there a difference?*

Problem 1.13 *Decrease the standard deviation (step size) of the proposal distribution to 0.01. For a sample size of 200, how the posterior for a step size of 0.01 compare to that obtained for 0.1?*

Problem 1.14 *Increase the standard deviation (step size) of the proposal distribution to 1. For a sample size of 200, how the posterior for a step size of 1 compare to that obtained for 0.1?*

Problem 1.15 *Suppose we collect data for a number of such samples (each of size 100), and find the following numbers of ticks that test positive for Borrelia: (3,2,8,25). Either calculate the new posterior exactly, or use sampling to estimate it. (Hint: in both cases make sure you include the original sample of 6!)*

Problem 1.16 *Generate samples from the posterior predictive distribution, and use these to test your model. What do these suggest about your model's assumptions?*

Problem 1.17 *A colleague suggests as an alternative you use a Beta-binomial likelihood, instead of the existent Binomial likelihood. This distribution has two uncertain parameters $\alpha > 0$ and $\beta > 0$ (the other parameter is the sample size; $n = 100$ in this case), where the mean of the distribution is $\frac{n\alpha}{\alpha+\beta}$ Your colleague and you decide to use weakly informative priors of the form: $\alpha \sim Gamma(1, \frac{1}{8})$ and $\beta \sim Gamma(10, 1)$. (Here I am assuming the mean of $Gamma(a, b) = \frac{a}{b}$.) Visualise the joint prior in this case.*

Problem 1.18 *For this situation your colleague tells you, there are unfortunately no conjugate priors. As such, three possible solutions (of many) open to you are: 1. you use numerical integration to find the posterior parameters, or 2. use the Random Walk Metropolis-Hastings algorithm, or 3. you transform each of (α, β) so that they lie between $-\infty < \theta < \infty$. Why can't you use vanilla Random Walk Metropolis for (α, β) here?*

Problem 1.19 *By using one of the three methods above estimate the joint posterior distribution. Visualise the pdf of the joint posterior. How are α and β correlated here?*

Problem 1.20 *Construct 80% credible intervals for the parameters of the Beta-binomial distribution.*

Problem 1.21 *Carry out appropriate posterior predictive checks using the new model. How does it fare?*

2 The fairground revisited

You again find yourself in a fairground, and where there is a stall offering the chance to win money if you participate in a game. Before you participate you watch a few other plays of the game (by other people in the crowd) to try to determine whether you want to play.

Problem 2.1 *In the most-boring version of the game, a woman flips a coin and you bet on its outcome. If the coin lands heads-up, you win; if tails, you lose. Based on your knowledge of similar games (and knowledge that the game must be rigged for the woman to make a profit!) you assume that the coin must be biased towards tails. As such you decide to specify a prior* $\theta \sim beta(2, 5)$*. Graph this function, and – using your knowledge of the beta distribution – determine the mean parameter value specified by this prior.*

Problem 2.2 *You watch the last 10 plays of the game, and the outcome is heads 3/10 times. Assuming a binomial likelihood, create a function that determines the likelihood for a given value of the heads-probability,* θ*. Hence or otherwise, determine the maximum likelihood estimate of* θ*.*

Problem 2.3 *Graph the likelihood* $\times$ *prior. From the graph approximately determine the MAP* θ *estimate value.*

Problem 2.4 *By using R's* `integrate` *function find the denominator, and hence graph the posterior pdf.*

Problem 2.5 *Use your posterior to determine your fair price for participating in the game, assuming that you win £1 if the coin comes up heads, and zero otherwise.*

Problem 2.6 *(Optional) Another variant of the game is as follows: the woman flips a first coin – if it is tails you lose (*$Y_i = 0$*), and if it is heads you proceed to the next step. In this step, the woman flips another coin ten times, and records the number of heads,* Y_i*, which determine your winnings. Explain why a reasonable choice for the likelihood might be,*

$$L(\theta, \phi|Y_i) = \begin{cases} (1-\theta) + \theta(1-\phi)^{10}, & \text{if } Y_i = 0 \\ \theta\binom{10}{Y_i}\phi^{Y_i}(1-\phi)^{10-Y_i}, & \text{if } Y_i > 0 \end{cases}$$

where θ *and* ϕ *are the probabilities of the first and second coins falling heads-up, and* Y_i *is the score on the game.*

Problem 2.7 *(Optional) Using the above formula, write down the overall log-likelihood for a series of* N *observations for* $Y_i = (Y_1, Y_2, ..., Y_N)$*.*

Problem 2.8 *(Optional) Using R's* `optim` *function determine the maximum likelihood estimate of the parameters for* $Y_i = (3, 0, 4, 2, 1, 2, 0, 0, 5, 1)$*.*

Hint 1: Since R's `optim` *function does minimisation by default, you will need to put a minus sign in front of the function to maximise it.*

Problem 2.9 *(Very very optional) Determine confidence intervals on your parameter estimates. Hint 1: use the second derivative of the log-likelihood to estimate the Fischer Information matrix, and hence determine the Cramer-Rao lower bound. Hint 2: use Mathematica!*

Problem 2.10 *Assuming uniform priors for both θ and ϕ create a function in R that calculates the unnormalised posterior (the numerator of Bayes' rule).*

Problem 2.11 *By implementing the Metropolis algorithm, estimate the posterior means of each parameter. Hint 1: use a normal proposal distribution. Hint 2: use periodic boundary conditions on each parameter, so that a proposal off one side of the domain maps onto the other side.*

Problem 2.12 *Find the 95% credible intervals for each parameter.*

Problem 2.13 *Using your posterior samples determine the fair price of the game. Hint: find the mean of the posterior predictive distribution.*