Bayesian course - problem set 4 (lecture 5)

Ben Lambert

November 17, 2016

1 Cerebral malaria: coding up samplers

Suppose you work for the WHO where you research malaria. In particular it is your job to come up with a model for the number of cases of cerebral malaria in a large country. Cerebral malaria is one of the most severe complications resulting from infection with *Plasmodium falciparum* malaria, and without treatment invariably causes death. However, even for patients receiving treatment there is still a significant chance of permanent cognitive impairment.

You decide to model the number of cases of cerebral malaria ($X = 5$) as being from a joint normal distribution along with the number of all malaria cases ($Y = 20$). The mean number of cases of cerebral malaria is μ_c, and the mean cases of all malaria is μ_t. If we assume an (improper) uniform prior distribution on these quantities and assume that the correlation between cerebral and total cases is known ($\rho = 0.8$) the posterior turns out to be:

$$\begin{pmatrix} \mu_t \\ \mu_c \end{pmatrix} \sim N\left[\begin{pmatrix} 20 \\ 5 \end{pmatrix}, \quad \begin{pmatrix} 2 & 0.8 \\ 0.8 & 0.5 \end{pmatrix}\right]$$

where all quantities are measured in units of "000s".

Note: this example does not test your ability to do Bayesian inference (because we have already provided the exact form of the posterior distribution). Rather its purpose is allow you to compare the performance of a number of different sampling algorithms.

Problem 1.1 *Use your statistical software to generate 100 independent samples of (μ_t, μ_c). Draw a line plot of your (μ_t, μ_c) samples. How close are the sample-estimated means to the true means? (Hint: to do this in R you will need to use the MASS package:*

```
library(MASS)

Sigma <- matrix(c(2,0.8,0.8,0.5),2,2)
mvrnorm(n = 100, c(20,5), Sigma)

)
```

Problem 1.2 *Code up a Random Walk Metropolis sampler for this example. This is composed of the following steps:*

1. *Create a proposal function that takes a current value of $\theta = (\mu_t, \mu_c)$ and outputs a proposed value of these using a multivariate normal centred on the current estimates. (Here use a multivariate normal proposal with an identity covariance matrix.)*

2. *Create a function which takes as inputs $\theta^{current}$ and $\theta^{proposed}$, and outputs the ratio of the posteriors of the proposed value to the current one (Hint: to do this in R you will need to use the following to calculate the value of the posterior at (x, y):*

   ```
   library(mvtnorm)

   Sigma <- matrix(c(2,0.8,0.8,0.5),2,2)
   dmvnorm(c(x,y),c(20,5), Sigma)
   ```

).

3. *Create an accept/reject function which takes as inputs $\theta^{current}$ and $\theta^{proposed}$, then uses the above ratio function to find: $r = \frac{\theta^{proposed}}{\theta^{current}}$; then compares r with a uniformly-distributed random number u between 0 and 1. If $r > u \implies$ output $\theta^{proposed}$; otherwise output $\theta^{current}$.*

4. *Combine the proposal function along with the accept/reject function to make a function that takes as input $\theta^{current}$, proposes a new value of θ, then based on r moves to that new point or stays in the current position.*

5. *Create a function called "RWMetropolis" that takes a starting value of θ and runs for n steps.*

Use your "RWMetropolis" function to generate 100 samples from the posterior starting from $(\mu_t, \mu_c) = (10, 5)$. Draw a line plot of your (μ_t, μ_c) samples. How do your estimates of the posterior mean from Random Walk Metropolis compare with the true values? Why is there a bias in your estimates, and how could this be corrected?

Problem 1.3 *For your 100 samples using Random Walk Metropolis calculate the percentage of accepted steps.*

Problem 1.4 *Create a function that calculates Gelman's $\hat{R}$ for each of (μ_t, μ_c) using:*

$$\hat{R}(t) = \sqrt{\frac{W(t) + \frac{1}{T}(B(t) - W(t))}{W(t)}} \tag{1}$$

where:

$$W(t) = \frac{1}{m}\sum_{j=1}^{m} s(t)_j^2 \tag{2}$$

measures the within-chain variance at time t averaged over m chains, and $s(t)_j^2$ *is the sample variance of chain j. And:*

$$B(t) = \frac{t}{m-1}\sum_{j=1}^{m}(\overline{\theta(t)}_j - \overline{\theta(t)})^2 \tag{3}$$

measures the between-chain variance at time t. Here $\overline{\theta(t)}_j$ *is the average value of a parameter in chain j, and* $\overline{\theta(t)}$ *is the average value of a parameter across all chains. (Hint 1: first create two separate functions that calculate the within and between chain variance. Hint 2: you will obtain a value of* $\hat{R}$ *for each of* (μ_t, μ_c)*.)*

Problem 1.5 *Start all 8 chains at* $(\mu_t, \mu_c) = (20, 5)$ *and calculate* $\hat{R}$ *for a per chain sample size of 5. Does this mean we have reached convergence?*

Problem 1.6 *Using 8 chains calculate* $\hat{R}$ *for each of* (μ_t, μ_c) *for a sample size of 100. This time make sure to start your chains in overdispersed positions in parameter space. Use a random number from a multivariate normal centred on the posterior means with a covariance matrix of 40 times the identity matrix.*

Problem 1.7 *After about how many iterations does Random Walk Metropolis reach* $\hat{R} < 1.1$*?*

Problem 1.8 *The conditional distributions of each variable are given by:*

$$\mu_t \sim N(20 + 1.6(\mu_c - 5), (1 - 0.8^2)2)$$
$$\mu_c \sim N(5 + 0.4(\mu_t - 20), (1 - 0.8^2)0.5)$$

Use this information to code up a Gibbs sampler, again starting at $(\mu_t, \mu_c) = (10, 5)$*. (Hint: use "rnorm", (in R), or equivalent to create two functions: one that produces draws of* μ_t *given* μ_c*; and the other that produces draws of* μ_c *given* μ_t*. Then create a function that cycles between these updates. Make sure to always draw samples using the most recent values of* (μ_t, μ_c)*).*

Problem 1.9 *Use your Gibbs sampler to draw 100 samples. Draw a line plot of your* (μ_t, μ_c) *samples. Discarding the first 50 observations, how do the estimates of the mean of each parameter compare with their true values?*

Problem 1.10 *Generate 200 samples from each of your Random Walk Metropolis and Gibbs samplers. Discard the first 100 observations of each as warm-up. For each calculate the error in estimating the posterior mean of* μ_t*. Repeat this exercise 40 times; each time recording the error. How does their error compare to the independent sampler? (Hint: use 40 replicates of each sampler to estimate its error.)*

Problem 1.11 *Repeat the same exercise as the previous problem to obtain the average error in estimating the posterior mean of μ_t across a range of sample sizes $n = 5 : 200$.*

Problem 1.12 *Using the results from the previous question estimate the effective sample size for 150 observations of the Random Walk Metropolis and Gibbs samplers.*

Problem 1.13 *What do the above results tell you about the relative efficiency of each of the three samplers?*

Problem 1.14 *Code up a Hamiltonian Monte Carlo sampler for this problem. (Alternatively, use the functions provided in the R file "HMC_scripts.R".) Use a standard deviation of the momentum proposal distribution (normal) of 0.18, along with a step size $\epsilon = 0.18$ and $L = 10$ individual steps per iteration to simulate 100 samples from the posterior. How does the estimate of the mean compare with that from the Independent, Random Walk Metropolis and Gibbs samplers?*

Problem 1.15 *What is the acceptance rate for HMC? How does this compare with RWM?*

Problem 1.16 *Gibbs sampling has an acceptance rate of 100%. How can HMC be more efficient than Gibbs give that its acceptance rate is less than 100%?*

Problem 1.17 *You receive new data that results in a change in the posterior to:*

$$\begin{pmatrix}\mu_t \\ \mu_c\end{pmatrix} \sim N\left[\begin{pmatrix}20 \\ 5\end{pmatrix}, \quad \begin{pmatrix}2 & 0.99 \\ 0.99 & 0.5\end{pmatrix}\right]$$

Using your Random Walk Metropolis sampler calculate $\hat{R}$ for 8 chains; each generating 100 samples for each.

Problem 1.18 *Estimate the value of $\hat{R}$ for HMC on the posterior from the new data, for a sample size of 100. How does it compare to Random Walk Metropolis?*

2 The sensitivity and specificity of a test for a disease

Suppose that for a particular tropical disease no gold standard clinical test exists. Instead we have a test that is known to be imperfect; not always identifying a disease if the patient has the disease, and sometimes yielding false positives (patients that do not have the disease but test positive). However, by using this test in a clinical trial it is hoped that we can obtain better estimates for the disease sensitivity (S; the proportion of disease positive individuals who test

		Truth		
		+	-	
Test	+	Y_1	$a - Y_1$	a
	-	Y_2	$b - Y_2$	b
		$Y_1 + Y_2$	$N - (Y_1 + Y_2)$	N

Table 1: Test outcomes versus true outcomes. a and b are the number of observed positive and negative test results respectively. Y_1 and Y_2 are latent variables that represent the gold standard – the true number of positive individuals out of a and b respectively.

positive) and specificity (C; the proportion of individuals who don't have the disease who test negative).

To do this we can construct a table of the observed and latent data for the test outcomes (see table 1). In the table a and b are the number of observed positive and negative results respectively. Y_1 and Y_2 are latent variables that represent the gold standard – the true number of positive individuals out of a and b respectively.

Problem 2.1 *Write down an expression for the likelihood, supposing that the prevalence for the disease is π. Hint: multiply together the likelihoods corresponding to each of the cells in table 1.*

Problem 2.2 *Assuming priors of the form: $\pi \sim beta(\alpha_\psi, \beta_\pi)$, $S \sim beta(\alpha_S, \beta_S)$ and $C \sim beta(\alpha_C, \beta_C)$, it is possible to code up a Gibbs sampler for this problem [1] of the form,*

$$Y_1|a, \pi, S, C \sim binomial\left(a, \frac{\pi S}{\pi S + (1-\pi)(1-C)}\right) \tag{4}$$

$$Y_2|b, \pi, S, C \sim binomial\left(b, \frac{\pi(1-S)}{\pi(1-S) + (1-\pi)C}\right) \tag{5}$$

$$\pi|a, b, Y_1, Y_2 \sim beta\left(Y_1 + Y_2 + \alpha_\pi, a + b - Y_1 - Y_2 + \beta_\pi\right) \tag{6}$$

$$S|Y_1, Y_2 \sim beta\left(Y_1 + \alpha_S, Y_2 + \beta_S\right) \tag{7}$$

$$C|a, b, Y_1, Y_2 \sim beta\left(b - Y_2 + \alpha_C, a - Y_1 + \beta_C\right) \tag{8}$$

Using the above expressions code up a working Gibbs sampler.

Problem 2.3 *Suppose that out of a sample of 100 people, 20 of those tested negative and 80 positive. Assuming uniform priors on π, S and C, use Gibbs sampling to generate posterior samples for π. What do you conclude?*

Problem 2.4 *Suppose that a previous study that compare the clinical test with a laboratory gold standard concludes that $S \sim beta(10, 1)$ and $C \sim beta(10, 1)$. Use Gibbs sampling to estimate the new posterior for π. Why does this look different to your previously-estimated distribution?*

Problem 2.5 *Suppose a previous analysis concluded that $\pi \sim beta(1, 10)$. Using this distribution as a prior, together with uniform priors on S and C, determine the posterior distributions for the test sensitivity and specificity respectively. Why does the test appear to be quite specific, although it is unclear how sensitive it is?*

Problem 2.6 *Suppose that based on lab results you suppose that the test specificity $C \sim beta(10, 1)$, and $\pi \sim beta(1, 10)$, but the prior for S is still uniform. Explain the shape of the posterior for S now.*

Problem 2.7 *Now suppose that the sample size was 1000 people of which 200 tested positive. Using the same priors as the previous question, determine the posterior for S. What do you conclude about your test's sensitivity?*

Problem 2.8 *What do the previous results suggest is necessary to assess the sensitivity of a clinical test for a disease?*

3 Inference example: Poisson regression

Suppose that you work for the WHO where it is your job to research the behaviour of malaria-carrying mosquitoes. In particular, an important part of your research remit is to estimate adult mosquito lifespan. The lifespan of an adult mosquito is a critical determinant of the severity of malaria, since the longer a mosquito lives the greater the chance it has of a. becoming infected by biting an infected human; b. surviving the period where the malarial parasite undergoes a metamorphosis in the mosquito gut and migrates to the salivary glands; and c. passing on the disease by biting an uninfected host.

Suppose you estimate the lifespan of mosquitoes by analysing the results of a mark-release-recapture field experiment. The experiment begins with the release of 1,000 young adult mosquitoes (assumed to have an adult age of zero); each of which has been marked with a fluorescent die. On each day (t) you attempt to collect mosquitoes using a large number of traps, and count the number of marked mosquitoes that you capture (X_t). The mosquitoes caught each day are then re-released unharmed. The experiment goes on for 15 days in total.

Since X_t is a count variable and you assume that the recapture of an individual marked mosquito is i.i.d., then you choose to use a Poisson model (as an approximation to the Binomial since n is large):

$$X_t \sim Poisson(\lambda_t)$$
$$\lambda_t = 1000 \times exp(-\mu t)\psi$$

where μ is the mortality hazard rate (assumed to be constant) and ψ is the daily recapture probability. Initially you choose to use a Gamma(2,20) prior for μ (which has a mean of 0.1), and a Beta(2,40) prior for ψ.

The data for the experiment is contained in the file "prob4_mosquito.csv".

Problem 3.1 *Using the data in "prob4_mosquito.csv" create a function that returns the likelihood. (Hint: it is easiest to first write a function that accepts* (μ, ψ) *as an input, and outputs the mean on a day* t*.)*

Problem 3.2 *Find the maximum likelihood estimates of* (μ, ψ)*. Hint 1: this may be easier if you create a function that returns the log-likelihood, and maximise this instead. Hint 2: use R's optim function.*

Problem 3.3 *Construct 95% confidence intervals for the parameters. (Hint: find the information matrix, and use it to find the Cramer-Rao lower bound. Then approximately find confidence intervals by assuming normality.)*

Problem 3.4 *Write a function for the prior, and use this to create an expression for the un-normalised posterior.*

Problem 3.5 *Create a function that proposes a new point in parameter space using a log-normal proposal with mean at the current* μ *value, and a* $Beta(2+\psi, 40-\psi)$ *proposal for* ψ*. (Hint: use a log-normal*$(0.5(-\sigma^2+2log(\mu)), \sigma)$*, where* μ *is the current value of the parameter.)*

Problem 3.6 *Create a function that returns the ratio of the un-normalised posterior at the proposed step location, and compares it to the current position.*

Problem 3.7 *Create an accept/reject function using the Metropolis-Hastings accept-reject function.*

Problem 3.8 *Create a Metropolis-Hastings sampler by combining your proposal and accept/reject functions.*

Problem 3.9 *Use your sampler to estimate the posterior mean of* μ *and* ψ *for a sample size of 4000 (discard the first 50 observations.) (Hint: if possible, do this by running 4 chains in parallel.)*

Problem 3.10 *By numeric integration compute numerical estimates of the posterior means of* μ *and* ψ*. How does your sampler's estimates compare with the actual values? How do these compare to the MLEs?*

Problem 3.11 *Carry out appropriate posterior predictive checks to test the fit of the model. What do these suggest might be a more appropriate sampling distribution? (Hint: generate a single sample of recaptures for each value of* (μ, ψ) *using the Poisson sampling distribution. You only need to do this for about 200 sets of parameter values to get a good idea.)*

Problem 3.12 *An alternative model that incorporates age-dependent mortality is proposed where:*

$$\lambda_t = 1000 \times exp(-\mu t^{\beta+1})\psi \tag{9}$$

where $\beta \geq 0$*. Assume that the prior for this parameter is given by* $\beta \sim exp(5)$*. Using the same log-normal proposal distribution as for* μ *create a Random Walk Metropolis sampler for this new model. Use this sampler to find 80% credible intervals for the* (μ, ψ, β) *parameters.*

Problem 3.13 *Look at a scatter plot of μ against β. What does this tell you about parameter identification in this model?*

References

[1] Lawrence Joseph, Theresa W Gyorkos, and Louis Coupal. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology*, 141(3):263–272, 1995.