NATIONAL UNIVERSITY OF SINGAPORE

ST4234 Bayesian Statistics

Take-Home Final Assignment

(Semester 2: AY 2019/2020)

Instructions:

1. This take-home final assignment contains **FOUR (4)** problems and comprises **SEVEN (7)** printed pages. The total marks of this assignment is **70**.

2. Copying of solutions is **strictly prohibited**.

3. Please upload your work in **a single pdf file** to the LumiNUS folder **"Final assignment submission"**. The deadline for submission is

   Singapore Time (GMT+8) 2pm, Saturday, 25 April, 2020

   For those students who are in different time zones, please convert this deadline to your current time zone and strictly follow this deadline.

4. Please write clearly **your student number and your name** in your submission.

5. If your student number is XXX, please name your file

   XXX.pdf

   For example, if your matriculation number is A0012345R, your file should be named **A0012345R.pdf**.

6. If the question requires R codes, you must include your R codes and outputs in your submitted file. **Your codes must be executable in R or RStudio and produce the outputs in your file. Otherwise, they will be subject to mark deduction.**

7. You are allowed to scan your handwritings and combine files. Your file should be well organized. **Unrecognizable handwriting will be subject to mark deduction.**

Ideally, you should compress the final pdf file to no more than **5Mb**, without severely compromising the quality of images in your file.

8. **No hard copy** will be accepted. **No late submission** will be accepted (i.e. marks for your take-home final assignment = zero).

Question 1. (20 marks) Let $\boldsymbol{y} = \{y_1, \dots, y_n\}$ be independent and identically distributed (i.i.d.) random variables, each with the following density function

$$p(y|\alpha, \beta) = \frac{\alpha\beta^\alpha}{(y+\beta)^{\alpha+1}}, \quad \text{for } y > 0, \quad \alpha > 0, \ \beta > 0. \tag{1}$$

(a) (2 marks) Write down the mathematical expression of the likelihood function $p(\boldsymbol{y}|\alpha, \beta)$ and simplify your answer.

(b) (8 marks) Suppose that $\underline{\beta = 1}$. The model has α as the only unknown parameter.

(i) (3 marks) Find a class of conjugate priors for α. Clearly specify the name of the family and the parameters. Find the posterior distribution $p(\alpha|\boldsymbol{y})$ under this conjugate prior.

(ii) (5 marks) Find the Jeffreys prior for α. Is this a proper prior for α? Under this Jeffreys prior, is the posterior a proper probability distribution of α? Briefly explain your answers.

(Hint: You can assume that $n \geq 1$, $y_i > 0$ for all $i = 1, \dots, n$. You can use the fact that $\int_0^\infty x^a b^x \mathrm{d}x < \infty$, for any $a \geq 0$ and $0 < b < 1$.)

(c) (6 marks) Suppose that $\underline{\alpha = 1}$. The model has β as the only unknown parameter. The prior distribution on β has the density $p(\beta) = \frac{1}{(1+\beta)^2}$ for $\beta > 0$. Consider rejection sampling from the posterior $p(\beta|\boldsymbol{y})$. We consider two candidate proposal densities. The first candidate is $g(\beta) = \lambda \exp(-\lambda\beta)$ for $\beta > 0$, which is the density of the exponential distribution with rate parameter $\lambda > 0$. The second candidate is the prior density $p(\beta) = \frac{1}{(1+\beta)^2}$ for $\beta > 0$. For each of these two densities, determine whether they are appropriate proposal densities for rejection sampling from $p(\beta|\boldsymbol{y})$. If so, write down the detailed steps to draw S samples from $p(\beta|\boldsymbol{y})$, where S is a large integer. If not, explain why.

(Hint: You may assume that you can draw random samples from both proposal densities. You only need to write the steps. Please do not write R codes in your answer.)

(d) (4 marks) Consider the following hierarchical model:

$$y|\lambda \sim \text{Exp}(\lambda), \qquad \lambda|\alpha, \beta \sim \text{Gamma}(\alpha, \beta), \tag{2}$$

where $\text{Exp}(\lambda)$ is the exponential distribution with rate parameter $\lambda > 0$, and $\text{Gamma}(\alpha, \beta)$ is the gamma distribution with shape parameter α and rate parameter β. The variable λ is a latent variable in the hierarchical model. Show that the model specified in Equation (2) is equivalent to the model specified in Equation (1), i.e. from Equation (2), you can derive the density $p(y|\alpha, \beta)$ given in Equation (1).

Question 2. (18 marks) Let $\boldsymbol{y} = \{y_1, \ldots, y_n\}$ be an i.i.d. sample from the $\text{Geometric}(\theta)$ distribution, with the probability density function $p(y|\theta) = \theta(1-\theta)^y$ for $y = 0, 1, 2, \ldots$ and $\theta \in (0, 1)$. Suppose that θ is assigned a $\text{Beta}(a, b)$ prior, i.e. the prior density of θ is $p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$ for $\theta \in (0, 1)$, where $a > 0$ and $b > 0$ are fixed hyperparameters.

(a) (6 marks) Find the normal approximation to the posterior density $p(\theta|\boldsymbol{y})$.

(b) (6 marks) Find the Laplace approximation (the second method in Chapter 5) to the posterior mean $\text{E}(\theta|\boldsymbol{y})$ and simplify your answer.

(c) (6 marks) Consider the importance sampling estimation of the posterior mean $\text{E}(\theta|\boldsymbol{y})$. Suppose that

$$n = 5, \quad \sum_{i=1}^{n} y_i = 18, \quad a = 2, \quad b = 2.$$

Draw $S = 100$ samples of θ from the $\text{Uniform}(0, 1)$ proposal density, and generate their values using the following R codes:

```
set.seed(...)
theta.samples <- runif(100)
```

In the R codes above, you need to replace "..." by your **student number** without the letters. For example, if the number is A012345R, then your first line of codes will be `set.seed(0012345)`. After you obtain `theta.samples`, use this sample to find the importance sampling estimate of $\text{E}(\theta|\boldsymbol{y})$ and its standard error. Provide complete R codes and outputs in your answer.

Question 3. (15 marks) In an extension of the genetic linkage model, n ($n = y_1 + y_2 + y_3 + y_4 + y_5$) animals are distributed into 5 categories as follows:

$$\boldsymbol{y} = (y_1, y_2, y_3, y_4, y_5)$$

with cell probabilities

$$\left(\frac{\theta_1}{4} + \frac{1}{8}, \frac{\theta_1}{4}, \frac{\theta_2}{4}, \frac{\theta_2}{4} + \frac{3}{8}, \frac{1 - \theta_1 - \theta_2}{2}\right),$$

where the parameter $\theta = (\theta_1, \theta_2)$ satisfies $0 \leq \theta_1 \leq 1$, $0 \leq \theta_2 \leq 1$, and $0 \leq \theta_1 + \theta_2 \leq 1$. Suppose that $\theta = (\theta_1, \theta_2)$ is assigned a noninformative prior $p(\theta_1, \theta_2) \propto 1$.

(a) (3 marks) Write down the unnormalized posterior density of (θ_1, θ_2) and simplify your answer.

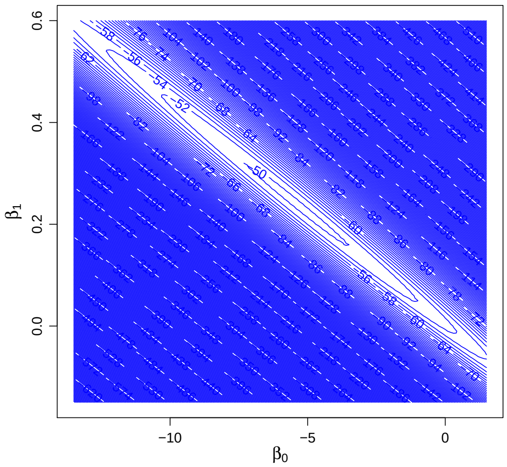
(b) (12 marks) Derive a Gibbs sampler to draw T samples from the posterior $p(\theta_1, \theta_2|\boldsymbol{y})$, where T is a large integer. You can introduce some additional latent variables. You need to show how to derive all the conditional posterior distributions, clearly specify the names and parameters of these conditional posterior distributions, and write down all steps of the Gibbs sampling algorithm. All your conditional posterior distributions must be familiar distributions which can be sampled directly in R, such as normal, gamma, inverse gamma, beta, bionomial, Poisson, etc. Do not write R codes in your answer.

Question 4. (17 marks) A math department is interested in exploring the relationship between students' scores on the ACT test, a standard college entrance exam, and their success (getting an A or a B) in a business calculus class. Data were obtained and summarized fom a sample of students. The summarized dataset consists of triplets (y_i, n_i, x_i) for $i = 1, \ldots, 8$, where n_i and y_i are the total number and the number of successful students with ACT score x_i. We treat all n_i's and x_i's as known constants and we model y_i's. We assume that the y_i's are independently distributed as Binomial(n_i, p_i), where the probabilities p_i satisfy the logistic model

$$\log \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 x_i.$$

We are interested in the Bayesian estimation of the regression coefficients (β_0, β_1). Suppose (β_0, β_1) is assigned an improper uniform prior $p(\beta_0, \beta_1) \propto 1$. We define the log posterior function $\log p(\beta_0, \beta_1|\boldsymbol{y})$ in R as `logpost(beta,data)`, where `beta` is the parameter vector (β_0, β_1) and `data` is a list that includes the data vector of $\{y_1, \ldots, y_8\}$, $\{n_1, \ldots, n_8\}$, and $\{x_1, \ldots, x_8\}$. We optimize `logpost(beta,data)` over `beta` to find the posterior mode. We obtain the following R output and contour plot.

```
> (out <- optim(par=c(0,0),fn=logpost,hessian=TRUE,
                control=list(fnscale=-1),data=data))
$par
[1] -5.366289  0.238423
$value
[1] -49.52815
$counts
function gradient
      77       NA
$convergence
[1] 0
$message
NULL
$hessian
          [,1]       [,2]
[1,]  -17.4066  -393.5803
[2,] -393.5803 -9017.5273
```



(a) (4 marks) Write down the mathematical expression of the log posterior density function $\log p(\beta_0, \beta_1|\boldsymbol{y})$ and simplify your answer.

(b) (4 marks) Based on the R output above, find the normal approximation to the posterior $p(\beta_0, \beta_1|\boldsymbol{y})$.

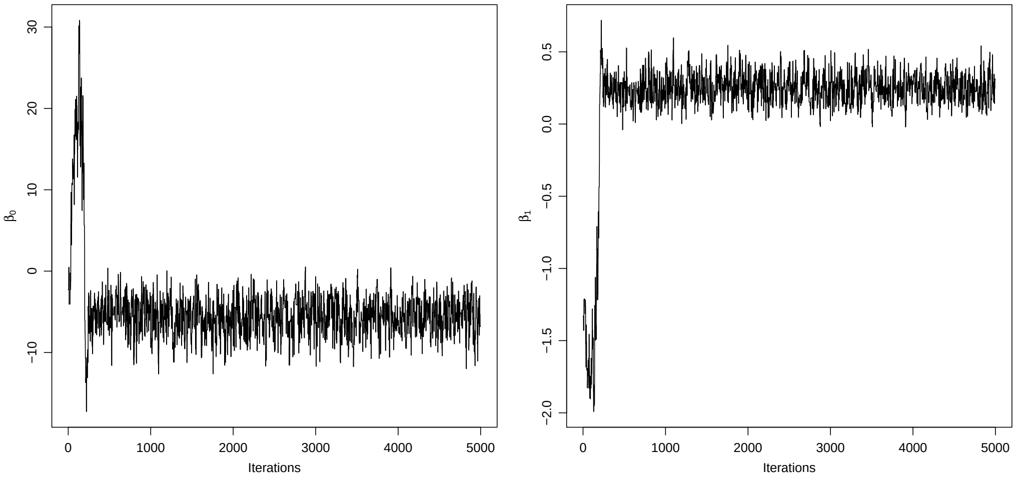
(c) (5 marks) We consider the random walk Metropolis algorithm to sample from $p(\beta_0, \beta_1|\boldsymbol{y})$ with two different proposal densities:

Algorithm 1: In each iteration, the proposal density is the normal density with the mean equal to the current parameter value and the covariance matrix $\begin{pmatrix} 0.01 & 0 \\ 0 & 0.01 \end{pmatrix}$;

Algorithm 2: In each iteration, the proposal density is the normal density with the mean equal to the current parameter value and the covariance matrix equal to 4 times the covariance matrix in normal approximation.

Give at least two reasons why Algorithm 1 will be less efficient than Algorithm 2, and describe what could happen if we use Algorithm 1 to sample from the posterior $p(\beta_0, \beta_1|\boldsymbol{y})$.

(d) (4 marks) Suppose that we use a random walk Metropolis algorithm to draw 5000 samples from the posterior $p(\beta_0, \beta_1|\boldsymbol{y})$. We observe the following traceplots for β_0 and β_1, respectively. Discuss what problem(s) can be identified from these traceplots and what are the possible solutions.



End of Final Assignment