AUSTRALIAN NATIONAL UNIVERSITY

RESEARCH SCHOOL OF FINANCE ACTUARIAL STUDIES, AND APPLIED STATISTICS

INTRODUCTION TO BAYESIAN DATA ANALYSIS (STAT3016/4116/7016)
SEMESTER 2 2017

ASSIGNMENT 2

DUE DATE: Friday 1 September 2017, by 3pm
(12.5% of total course grade)

INSTRUCTIONS:

1. All students must hand in an assignment of their own writing.
2. The assignment should be handed in to the assignment box for STAT3016/4116/7016 available on level 4 of the ANUCBE Building 26C. There will be no online submission facility.
3. Ensure you also complete and attach a cover sheet to your assignment (available on the course website)
4. Begin each question on a new page.
5. Where required, provide sufficient computer output to support your answers. Provide enough intermediate numerical calculations to justify working for your final answer.
6. Computer output must be interpreted in written format. A solution solely highlighting the computer output is not acceptable.
7. No late assignments will be accepted.

COLLABORATION POLICY (as stated in the course outline)

University policies on plagiarism will be **strictly** enforced. You are encouraged to (orally) discuss your assignments with your classmates, but each student must write up solutions separately. Be sure that you have worked through each problem yourself and that all answers you submit are the results of your own efforts. This includes all computer code and output.

Problem 1 Jim Albert, Chapter 5 Problem 1

Suppose y has a binomial distribution with parameters n and p, and we are interested in the log-odds value $\theta = \log(p/(1-p)$. Our prior for θ is that $\theta \sim N(mean = \mu, sd = \sigma)$.

Suppose we are interested in learning about the probability that a special coin lands heads when tossed. A priori we believe that the coin is fair, so we assign θ a $N(mean = 0, sd = 0.25)$ prior. We toss the coin $n = 5$ times and obtain $y = 5$ heads.

(a) What is the posterior distribution of θ? (up to a proportionality constant)

(b) For a univariate and roughly symmetric posterior distribution $p(\theta|y)$, it can be convenient to approximate it by a normal distribution. A Taylor series expansion of $\log p(\theta|y)$ centred at the posterior mode $\hat{\theta}$ gives

$$\log p(\theta|y) = \log p(\hat{\theta}|y) + \frac{1}{2}(\theta - \hat{\theta})^T \left[\frac{d^2}{d\theta^2} \log p(\theta|y)\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + ...$$

where the linear term in the expansion is zero because the log-posterior density has zero derivative at its mode. The remainder terms of higher order fade in importance relative to the quadratic term when θ is close to $\hat{\theta}$ and n is large. From the above expansion, we obtain the following normal approximation:

$$p(\theta|y) \approx N(\hat{\theta}, [I(\hat{\theta})]^{-1})$$

where $I(\theta)$ is the observed information ($I(\theta) = -\frac{d^2}{d\theta^2} \log p(\theta|y)$). The function `laplace` in the `LearnBayes` package in R finds the posterior mode using the function `optim`. The output of `laplace` is a list. The component `mode` gives the value of the posterior mode $\hat{\theta}$ and the component `var` is the associated variance estimate.

Use the `laplace` function to find a normal approximation to the posterior density of θ and compute the posterior probability that the coin is biased towards heads. Note: as inputs into `laplace` , you will need to write a function in R to evaluate the log posterior density and provide an initial guess for the value of $\hat{\theta}$.

(c) Using the prior density as a proposal density, design a rejection algorithm for sampling from the posterior distribution. Using simulated draws from your algorithm, approximate the probability that the coin is biased towards heads.

Hoff, Chapter 4 Problem 7

Problem 2

After a posterior analysis on data from a population of squash plants, it was determined that the total vegetable weight of a given plant could be modelled with the following distribution:

$$p(y|\theta, \sigma^2) = 0.31 \times \mathrm{dnorm}(y, \theta, \sigma) + 0.46 \times \mathrm{dnorm}(y, 2\theta, 2\sigma) + 0.23 \times \mathrm{dnorm}(y, 3\theta, 3\sigma)$$

where $\mathrm{dnorm}(y, \theta, \sigma)$ refers to the density of a normal distribution with mean θ and standard deviation σ. You are told that the posterior distributions of the parameters have been calculated as $1/\sigma^2|\mathbf{y} \sim \mathrm{gamma}(10, 2.5)$, and $\theta|\sigma^2, \mathbf{y} \sim \mathrm{normal}(4.1, \sigma^2/20)$.

(a) Sample at least 5,000 y values from the posterior predictive distribution.

(b) Form a 75% quantile-based confidence interval for a new value of Y.

(c) Form a 75% HPD region for a new Y as follows:

(i.) Compute estimates of the posterior density of Y using the `density` command in R, and then normalise the density values so they sum to 1.

(ii.) Sort these discrete probabilities in decreasing order.

(iii.) Find the first probability value such that the cumulative sum of the sorted values exceeds 0.75. Your HPD region includes all values of y which have a discretised probability greater than this cutoff. Describe your HPD region, and compare it to your quantile based region.

(d) Can you think of a physical justification for the mixture sampling distribution of Y?

Gelman Chapter 3 Problem 5

Problem 3

It is a common problem for measurements to be observed in rounded form. For a simple example, suppose we weigh an object five times and measure weights, rounded to the nearest kilogram, of 10, 10, 11, 10, 9. Assume the unrounded measurements are normally distributed with a noninformative prior distribution on the mean μ and variance σ^2.

(a) Give the posterior distribution for (μ, σ^2) obtained by pretending that the observations are exact unrounded measurements.

(b) Give the correct posterior distribution (up to a proportionality constant) for (μ, σ^2) treating the measurements as rounded.

(c) How do the incorrect and correct posterior distributions differ? Compare means, variances and contour plots.

(d) Let $z = (z_1, ..., z_5)$ be the original, unrounded measurements corresponding to the given observations above. Draw simulations from the posterior distribution of z. Compute the posterior mean of $(z_1 - z_3)^2$.

Problem 4

Jim Albert 4.4 A Bioassay Experiment

In the development of a new drug, forty animals were tested, ten at each of four dose levels. The data are provided in the table below:

Dose x_i (log g/ml)	Number of animals, n_i	Number of deaths, y_i
-0.86	10	0
-0.30	10	3
-0.05	10	7
0.73	10	8

we assume the following sampling model

$$y_i|\theta_i \sim \text{Bin}(n_i, \theta_i)$$

where θ_i is the probability of death for animals given dose x_i. To model the dose-response relationship, we use the logistic model:

$$\text{logit}(\theta_i) = \alpha + \beta x_i$$

where $\text{logit}(\theta_i) = \log(\theta_i/(1-\theta_i))$. Assume a uniform prior distribution for (α, β); that is, $p(\alpha, \beta) \propto 1$

(a) Assume a normal approximation to the true posterior density $p(\alpha, \beta|y)$, and sample $S = 10000$ draws, $(\alpha^{(1)}, \beta^{(1)}), ..., (\alpha^{(S)}, \beta^{(S)})$ from the approximate distribution. Call this approximate distribution $g(\alpha^{(s)}, \beta^{(s)})$. Then resample $k = 1000$ samples from the set $\left\{(\alpha^{(1)}, \beta^{(1)}), ..., (\alpha^{(S)}, \beta^{(S)})\right\}$ without replacement, where the probability of sampling each $(\alpha^{(s)}, \beta^{(s)})$ is proportional to the importance ratio $\frac{q(\alpha^{(s)}, \beta^{(s)}|y)}{g(\alpha^{(s)}, \beta^{(s)})}$ (where $q(\alpha, \beta|y)$ is the unnormalised density of $p(\alpha, \beta|y)$). Draw a scatter plot of your posterior draws.

(b) Repeat part (a) but sample with replacement. Discuss how your results differ, if at all.

Problem 5 Jim Albert Chapter 3 Problem 7

Suppose you are interested in estimating the mortality rate λ (per unit of exposure) of heart transplant surgeries for a particular hospital. To construct your prior, you talk to two experts. The first expert's beliefs about λ are described by a gamma(1.5,1000) distribution and the second expert's beliefs are described by a gamma(7,1000) distribution. You place equal credence in both experts.

(a) Construct a graph of the prior density of λ.

(b) Suppose the hospital of interest experiences y_{obs}=4 deaths with an exposure of 1767 patients. Compute the posterior distribution of λ.

(c) Plot the prior and posterior densisties of λ on the same graph.

(d) Find the probability that the mortality rate exceeds 0.005.

(e) Based on the mixing probabilities, were the data more consistent with the beliefs of the first expert or the beliefs of the second expert? Explain.

Problem 6 [STAT4116/7016 only] Hoff, Chapter 5 Problem 3

Given observations $Y_1, ..., Y_n \overset{\text{iid}}{\sim} Normal(\theta, \sigma^2)$ and using the conjugate prior distribution for θ and σ^2 (that is, $\frac{1}{\sigma^2} \sim Gamma(\nu_0/2, \sigma_0^2\nu_0/2)$ and $\theta|\sigma^2 \sim Normal(\mu_0, \sigma^2/\kappa_0)$), derive the formula for $p(\theta|y_1, ..., y_n)$, the marginal distribution of θ, conditional on the data but marginal over σ^2. Check your work by comparing your formula to a Monte Carlo estimate of the marginal distribution, using some values of $Y_1, ..., Y_n, \mu_0, \sigma_0^2, \nu_0$ and κ_0 that you choose.