The Chinese University of Hong Kong, Shenzhen
SDS · School of Data Science

· Fall Semester 2022

DDA 4010 – Bayesian Statistics

Exercise Sheet 3

This exercise is due on **Nov 11st, 5:00 pm.**

Assignment A3.1 (5.1 in Textbook):
(You can directly access the links like http://www2.stat.duke.edu/ pdh10/FCBS/Exercises/school1.dat or `school1 <- scan('http://www2.stat.duke.edu/~pdh10/FCBS/Exercises/school1.dat')` in R to access the data.) Studying: The files `school1.dat`, `school2.dat` and `school3.dat` contain data on the amount of time students from three high schools spent on studying or homework during an exam period. Analyze data from each of these schools separately, using the normal model with a conjugate prior distribution, in which $\{\mu_0 = 5, \sigma_0^2 = 4, \kappa_0 = 1, \nu_0 = 2\}$ and compute or approximate the following:

- posterior means and 95% confidence intervals for the mean θ and standard deviation σ from each school;
- the posterior probability that $\theta_i < \theta_j < \theta_k$ for all six permutations $\{i, j, k\}$ of $\{1, 2, 3\}$
- the posterior probability that $\tilde{Y}_i < \tilde{Y}_j < \tilde{Y}_k$ for all six permutations $\{i, j, k\}$ of $\{1, 2, 3\}$, where $\tilde{Y}_i$ is a sample from the posterior predictive distribution of school i.
- Compute the posterior probability that θ_1 is bigger than both θ_2 and θ_3, and the posterior probability that $\tilde{Y}_1$ is bigger than both $\tilde{Y}_2$ and $\tilde{Y}_3$.

Assignment A3.2 (5.4 in Textbook):
Jeffreys' prior: For sampling models expressed in terms of a p-dimensional vector $\boldsymbol{\psi}$, Jeffreys' prior (Exercise 3.11) is defined as $p_J(\boldsymbol{\psi}) \propto \sqrt{|I(\boldsymbol{\psi})|}$, where $|I(\boldsymbol{\psi})|$ is the determinant of the $p \times p$ matrix $I(\boldsymbol{\psi})$ having entries $I(\psi)_{k,l} = -\mathrm{E}\left[\partial^2 \log p(Y \mid \psi)/\partial\psi_k\partial\psi_l\right]$

- Show that Jeffreys' prior for the normal model is $p_J\left(\theta, \sigma^2\right) \propto \left(\sigma^2\right)^{-3/2}$.
- Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ be the observed values of an i.i.d. sample from a normal $\left(\theta, \sigma^2\right)$ population. Find a probability density $p_J\left(\theta, \sigma^2 \mid \boldsymbol{y}\right)$ such that $p_J\left(\theta, \sigma^2 \mid \boldsymbol{y}\right) \propto p_J\left(\theta, \sigma^2\right) p\left(\boldsymbol{y} \mid \theta, \sigma^2\right)$. It may be convenient to write this joint density as $p_J\left(\theta \mid \sigma^2, \boldsymbol{y}\right) \times p_J\left(\sigma^2 \mid \boldsymbol{y}\right)$. Can this joint density be considered a posterior density?

Assignment A3.3 (6.2 in Textbook):
Mixture model: The file glucose. dat contains the plasma glucose concentration of 532 females from a study on diabetes (see Exercise 7.6).

- Make a histogram or kernel density estimate of the data. Describe how this empirical distribution deviates from the shape of a normal distribution.

- Consider the following mixture model for these data: For each study participant there is an unobserved group membership variable X_i which is equal to 1 or 2 with probability p and $1-p$. If $X_i = 1$ then $Y_i \sim \text{normal}\left(\theta_1, \sigma_1^2\right)$, and if $X_i = 2$ then $Y_i \sim \text{normal}\left(\theta_2, \sigma_2^2\right)$. Let $p \sim \text{beta}(a, b), \theta_j \sim \text{normal}\left(\mu_0, \tau_0^2\right)$ and $1/\sigma_j \sim \text{gamma}\left(\nu_0/2, \nu_0\sigma_0^2/2\right)$ for both $j = 1$ and $j = 2$. Obtain the full conditional distributions of $\left(X_1, \ldots, X_n\right), p, \theta_1, \theta_2, \sigma_1^2$ and σ_2^2.

- Setting $a = b = 1, \mu_0 = 120, \tau_0^2 = 200, \sigma_0^2 = 1000$ and $\nu_0 = 10$, implement the Gibbs sampler for at least 10,000 iterations. Let $\theta_{(1)}^{(s)} = \min\left\{\theta_1^{(s)}, \theta_2^{(s)}\right\}$ and $\theta_{(2)}^{(s)} = \max\left\{\theta_1^{(s)}, \theta_2^{(s)}\right\}$. Compute and plot the autocorrelation functions of $\theta_{(1)}^{(s)}$ and $\theta_{(2)}^{(s)}$, as well as their effective sample sizes.

- For each iteration s of the Gibbs sampler, sample a value $x \sim \text{binary}\left(p^{(s)}\right)$, then sample $\tilde{Y}^{(s)} \sim \text{normal}\left(\theta_x^{(s)}, \sigma_x^{2(s)}\right)$. Plot a histogram or kernel density estimate for the empirical distribution of $\tilde{Y}^{(1)}, \ldots, \tilde{Y}^{(S)}$, and compare to the distribution in part a). Discuss the adequacy of this two-component mixture model for the glucose data.

Sheet 3 is due on **Nov 11st**. Submit your solutions before **Nov 11st**, **5:00 pm**.