University of Cape Town

Department of Statistical Sciences

Introduction to Bayes Assignment - 2025

Instructions:

- This assignment is to be completed in groups of up to three members.
- The assignment **will contain some concepts** that we have not formally covered in class. You will be expected to do some research on these aspects of the assignment - give yourself sufficient time to complete the assignment. Do not start a day or two before the hand in date.
- Late hand-ins will be penalized.
- An electronic copy of the assignment as well as your (working) R code should be submitted by 11.55pm on the 14th of April.
- This electronic copy must be a LaTeX-based document. In other words, the write up should be in Markdown or LaTeX. Any handwritten components or Microsoft Word documents converted to pdf will be penalised.
- There may be bonus marks available, but the maximum you may earn is 100%.
- Include your student number on the hand-in and remember to sign a plagiarism declaration.

Preamble

Use the code chuck below to simulate the data and install the necessary packages required for this assignment.

```
# Packages required
require(MASS)
require(cubature)

#Lets simulate some data
set.seed(2021)
n = 150    # Number of data points
X.c = data.frame(matrix(rnorm(5*n), ncol=5))
colnames(X.c) = c("X1", "X2", "X3", "X4", "X5")

X = as.matrix(cbind(1, X.c))    # Design matrix
e = matrix(rnorm(n), ncol=1)    # Errors

beta.true = matrix(c(1, 0, 10, 0, 2, -3), ncol=1)
Y = X%*%beta.true + e   # Observations
```

Q1) Linear regression with Gibbs sampling [28 marks]

Consider the linear regression model, $y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i$ where $e_i \sim \mathcal{N}(0, \sigma^2)$ for $i = 1, \ldots, n$.

(a) Assume that we have the following priors:

$$\begin{aligned} \left[\boldsymbol{\beta} | \sigma^2\right] &\sim \mathcal{N}_{k+1}\left(\tilde{\boldsymbol{\beta}}, \sigma^2 \boldsymbol{M}\right) \text{ and} \\ \left[\sigma^2\right] &\sim \mathcal{IG}\left(a, b\right), \end{aligned} \tag{A1}$$

where $\tilde{\boldsymbol{\beta}} = \mathbf{0}$, $\boldsymbol{M} = \boldsymbol{I}_{k+1}$ and $a = b = 1$. Using the assumptions from eq. (A1), (clearly) **show that** the conditional posterior distributions are given as:

$$\begin{aligned} \left[\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}, \boldsymbol{X}\right] &\sim \mathcal{N}_{k+1}\left(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \sigma^2\left(\boldsymbol{M} + \boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right) \\ \left[\sigma^2|\boldsymbol{y}, \boldsymbol{X}\right] &\sim \mathcal{IG}\left(a + \frac{n}{2}, b + \frac{A_2}{2}\right) \end{aligned} \tag{1}$$

where

$$\mu_{\boldsymbol{\beta}} = \left(\boldsymbol{M} + \boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\left(\left(\boldsymbol{X}^T\boldsymbol{X}\right)\hat{\boldsymbol{\beta}} + \boldsymbol{M}\tilde{\boldsymbol{\beta}}\right)$$

$$A_2 = \boldsymbol{y}^T\boldsymbol{y} + \tilde{\boldsymbol{\beta}}^T\boldsymbol{M}\tilde{\boldsymbol{\beta}} - \mu_{\boldsymbol{\beta}}^T\left(\boldsymbol{M} + \boldsymbol{X}^T\boldsymbol{X}\right)\mu_{\boldsymbol{\beta}}.$$

Note that $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$.

Hints for $\left[\sigma^2|\boldsymbol{y}, \boldsymbol{X}\right]$

- $\left[\sigma^2|\boldsymbol{y}, \boldsymbol{X}\right] = \int_{\boldsymbol{\beta}}\left[\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \boldsymbol{X}\right]d\boldsymbol{\beta}$
- $|aA| = a^k|A|$, where A is a k by k matrix.

[5+6 marks]

(b) Using the results from above, write R code to generate posterior samples. [5 marks]

(c) Using the function above, obtain **at least** 50 000 posterior samples.

(i) Create trace plots for each of the regression coefficients. Briefly discuss the results and what they mean. [2+1 marks]

(ii) Create density plots for each of the regression coefficients.

i. Show where the sample average and true coefficients lie on the density plots. Briefly discuss the results. [2+1 marks]

ii. Show where the cutoffs are for the 95% credibility interval on the density plots. Briefly discuss the difference between credibility intervals and confidence intervals. Discuss the computed credibility intervals. [2+2 marks]

iii. Using the credibility intervals, briefly discuss which variables are worth selecting. [2 marks]

Q2) A Bayesian Search [10 marks]

The use of Bayesian methods became extremely popular and instrumental in operations research during the Second World War. In particular, the method of searching for German Submarines grew to rely almost solely on the field of Bayesian Search Theory. We will briefly explore this topic through a question based on a true story.

One evening, there were two fisherman out at sea. They left the boat on autopilot and whilst one was sleeping, the other fisherman fell overboard. When the fisherman still on the boat awoke to find his peer missing, he contacted the coast guard to initiate a search and rescue operation. We will employ a Bayesian search method to find the lost fisherman.

We begin by creating a search grid of the possible areas which the fisherman could be lost. We will use a simple 20×20 search grid where we can search a single cell at a time. Let us define the following:

$$\begin{aligned} i &= \text{indices of the grid cells} \\ Y_i &= \text{Binary variable to determine the true state of whether the fisherman is in the } i^{th} \text{ cell,} \\ &\quad \text{where 1 represents that the fisherman is indeed in that cell.} \\ \theta_i &= \text{probability of occurrence} = \mathbb{P}(Y_i = 1) \end{aligned}$$

The main problem arises from the difficulty of finding a fisherman floating in the ocean. In particular, there is a *probability of detection* which represents the probability that we find the fisherman given that he is in the cell we are searching. Hence, let

$$\begin{aligned} Z_i &= \text{search result of the } i^{th} \text{ cell, 1 meaning detection} \\ p_i &= \text{probability of detection} = \mathbb{P}(Z_i = 1 | Y_i = 1) \end{aligned}$$

This means that the two probabilities θ_i and p_i are prior values obtained from experts before the search and rescue begins. We will use Bayes theorem to update these probabilities after each cell search that we conduct.

Let us assume that there is independence between the grid cells and that both the occurrence and detection of the fisherman in each cell follows a Bernoulli distribution such that:

$$Z_i|Y_i \sim Ber(Y_i p_i) = (Y_i p_i)^z (1 - Y_i p_i)^{1-z},$$
$$Y_i \sim Ber(\theta_i) = (\theta_i)^y (1 - \theta_i)^{1-y}$$

Hence, we are obviously interested in the case where the i^{th} cell contains the fisherman but we fail to detect him. The posterior probability for this case is:

$$\pi(Y_i = 1|Z_i = 0) = \frac{(1 - p_i)\theta_i}{1 - p_i\theta_i} \tag{2}$$

We will use this posterior distribution to update the occurrence probability, θ_i, once we search cell i and have no detection such that:

$$\theta_{i,new} = \frac{(1 - p_i)\theta_{i,old}}{1 - p_i\theta_{i,old}} \tag{3}$$

Similarly, the probability of occurrence in any other cell j, $j \neq i$, has the updating equation:

$$\theta_{j,new} = \frac{\theta_{j,old}}{1 - p_i\theta_{i,old}} \tag{4}$$

What have we achieved? We now have a representation for our posterior distribution to update the probability of the fisherman's location. We can now easily carry out a search and rescue process as such:

1. Gain a prior distribution for the fisherman's location from experts
2. Gain a detection distribution given the location
3. Merge these two distributions to find the distribution of detecting the fisherman in a given cell
4. Perform a search of the cell with the highest probability of a successful detection
5. If the fisherman is not detected, then update the probability distribution via Bayes' Theorem
6. Continue until the fisherman is found

We will simulate this process. Use the provided R code template to carry out this process. The template includes the initial setup of the problem, as well as the prior and detection probability distributions obtained from the expert.

a) **Clearly** derive the equations 2, 4 from the given Bernoulli distributions. [2 marks]

b) Explain why the occurrence probability updating equation, given by 3, follows from the posterior distribution, given by 2.

Hint: Think in terms of priors and posteriors over time. [1 mark]

c) Using the template, simulate the search and rescue process. When you search a cell, simulate the search process by drawing from a Bernoulli trial. The clock is ticking. Try and find him in 48 hours (one cell search per hour). Show the change in heatmap at step one and the final search step. Secondly, track the posterior probability of occurrence *for the cell in which the fisherman is truly in.* [6 marks]

d) Currently, the detection probability p_i varies across the grid. Suppose instead that p_i is constant for all cells (i.e., the detection probability does not depend on location). How does this simplification affect the search strategy? Explain your reasoning. [1 mark]

Q3) A Twist on Linear Regression [12 marks]

We are often faced with the problems of outliers. In many cases, statisticians suggest removing outliers from a dataset; however, this is not the only option.

Consider the linear regression model where the data is generated as

$$Y_i = \begin{cases} \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i, & e_i \sim \mathcal{N}(0, \sigma_1^2), i \in \mathcal{I}_1 \\ \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i, & e_i \sim \mathcal{N}(0, \sigma_2^2), i \in \mathcal{I}_2 \end{cases}$$

where

$\mathcal{I}$ is an index set

$\sigma_1^2 < \sigma_2^2$.

This implies that observations of the second index set have a higher variance than those of the first, potentially outliers in the data set.

As usual, let $\tau_i = \frac{1}{\sigma_i^2}$. If we assume that $\mathcal{I}_1$ is known, then w.l.o.g., we can let

$$\mathcal{I}_1 = \{1, \dots, n_1\},$$

$$\mathcal{I}_2 = \{n_1 + 1, \dots, n\}.$$

where the first n_1 is the number of 'standard' observations and the rest of the observations are outliers.

a) Derive the following conditional posterior distributions: $\pi(\beta|\tau_1, \tau_2, X, Y)$, $\pi(\tau_1|\beta, \tau_2, X, Y)$, and $\pi(\tau_2|\tau_1, \beta, X, Y)$. For the derivation of the posteriors, assume the following prior distributions: $\beta \sim \mathcal{N}(0, T_0)$, $\tau_1 \sim \mathcal{G}(a, b)$, and $\tau_2|\tau_1 \sim \mathcal{G}(a, b)I_{\tau_1 > \tau_2}$. [9 Marks]

b) Write R code to obtain samples from $\pi(\beta, \tau_1, \tau_2|X, Y)$. [3 Marks]