ST3457 Statistical Inference I (Bayesian Statistics)

Lab 1 - 16/11/17

Bernardo Nipoti

Lab 1: Monte Carlo method

Download the file `LAB_Montecarlo.R` from the course webpage

`http://www.bernardonipoti.com/teaching/st3457_17_18`

and use it to address the points below. You can work individually or in small groups. If you do not complete the instructions today, you can work on it during the week and compare your work with the solutions, that I will put online. If any doubt arises, ask next Thursday.

Part 1. Poisson model: Monte Carlo analysis *birthrate data*

Your first task is to go through and understand the code for the analysis of birthrate data. To this end read the following summary and go through PART 1 of the `R` script.

From Section 3.2.2 of the textbook:

DATA. In this example we consider a group of women with college degree and analyse their number of children. Let $Y_1, \ldots, Y_n$ be the observed data (Y_i is the number of children for the ith woman of the group). We have $n = 44$, $\sum_{i=1}^n Y_i = 66$, $\bar{Y} = 1.50$.

MODEL. We assume a Poisson sampling model with conjugate gamma prior, that is

$$\begin{aligned} Y_1, \ldots, Y_n \mid \theta &\overset{\text{iid}}{\sim} \text{Poisson}(\theta) \\ \theta &\sim \text{gamma(a=2,b=1)}. \end{aligned}$$

POSTERIOR DISTRIBUTIONS. The posterior distribution of θ is given by

$$\theta \mid n = 44, \sum_{i=1}^{n} Y_i = 66 \sim \text{gamma}(2 + 66, 1 + 44) = \text{gamma}(68, 45).$$

Part 2. Normal model with fixed variance: Monte Carlo analysis for *Iris data*

The second task consists in analysing the Iris dataset. To this end, read the model description below and follow the instructions that you find in the `R` script. Commands used in Part 1 must be adapted to the different model considered here.

DATA. We analyse the Iris dataset, available in the `Datasets` package in `R`, which collects different measurements of three species of iris flowers. Specifically we focus on the petal length of these flowers. We have three species: *versicolor*, *setosa*, *virginica*. We let $Y_{1,1}, \dots, Y_{n_1,1}$ be the observations for the first group (versicolor), that is $Y_{i,1}$ is the petal length in centimeters of the ith observed flower of species versicolor. Similarly, $Y_{1,2}, \dots, Y_{n_2,2}$ and $Y_{1,3}, \dots, Y_{n_3,3}$ are the observations for second and third group (setosa and virginica, respectively).

MODEL. We assume a Normal sampling model for each group, with fixed common variance $\sigma^2 = 0.5$ and with conjugate normal prior for the mean. The priors for the three groups are assumed independent and identical. That is

$$\begin{aligned} Y_{1,1}, \dots, Y_{n_1,1} \mid \theta_1, \sigma^2 = 0.5 &\overset{\text{iid}}{\sim} \text{Normal}(\theta_1, \sigma^2) \\ Y_{1,2}, \dots, Y_{n_2,2} \mid \theta_2, \sigma^2 = 0.5 &\overset{\text{iid}}{\sim} \text{Normal}(\theta_2, \sigma^2) \\ Y_{1,3}, \dots, Y_{n_3,3} \mid \theta_3, \sigma^2 = 0.5 &\overset{\text{iid}}{\sim} \text{Normal}(\theta_3, \sigma^2) \\ \theta_1, \theta_2, \theta_3 &\overset{\text{iid}}{\sim} \text{Normal}(\mu_0, \sigma^2/k_0). \end{aligned}$$

Based on previous studies, we have limited prior information about the length of the petals, which is known to be on average 3.5 cm. These studies made no distinction between species and therefore we set $\mu_0 = 3.5$ for each one of the three groups and set $k_0 = 1$ (prior sample size) so to assign only little weight to our prior guess.

POSTERIOR DISTRIBUTION. The posterior distribution of θ_i, for each $i = 1, 2, 3$, and conditionally on $\sigma^2 = 0.5$, is given by

$$\theta_i \mid n_i, \bar{Y}_i \sim \text{Normal}(\mu_{n_i,i}, \sigma^2_{n_i,i}),$$

where, for each $i = 1, 2, 3$,

$$\mu_{n_i,i} = \frac{k_0}{k_0 + n_i}\mu_0 + \frac{n_i}{k_0 + n_i}\bar{y}_i, \qquad \text{and} \qquad \sigma^2_{n_i,i} = \frac{\sigma^2}{k_0 + n_i}.$$