Obligatory assignment 3 MVE550, autumn 2023

Petter Mostad

November 27, 2023

1. A biologist is investigating the frequency with which sea-living animals of a certain species develops a certain disease, and how this frequency depends on the concentration of a certain pollutant and the temperature. Based on experience from similar contexts, she uses a model where an animal exposed to the pollutant concentration x and the temperature y has a probability $p = f(x, y, \theta_1, \theta_2, \theta_3)$ of developing the disease, where

$$f(x, y, \theta_1, \theta_2, \theta_3) = \frac{\exp\left(e^{\theta_1}x + e^{\theta_2}(y - \theta_3)^2\right) - 1}{\exp\left(e^{\theta_1}x + e^{\theta_2}(y - \theta_3)^2\right) + 1}.$$

   Here, $\theta = (\theta_1, \theta_2, \theta_3)$ are the parameters of the model. Each of them can take on any real value. A flat prior is assumed for all of them.

   The data is given in the file "dataAssignment3.txt". It can be read into `R` with the command `read.table("dataAssignment3.txt", header=TRUE)` and consists of a matrix where each row i contains observed values (x_i, y_i, z_i) for an animal i: x_i is the pollutant concentration the animal was exposed to, y_i the temperature it was exposed to, while $z_i = 1$ indicates that the animal had the disease and $z_i = 0$ indicates it did not.

   (a) Using the model above and the function f, write down the likelihood of the data (i.e., a formula for the probability of the data given the parameters of the model). Also, write down a function that is proportional to the posterior density for the parameters.

   (b) Write an R function that takes as input values for the parameters $\theta = (\theta_1, \theta_2, \theta_3)$ and computes a function that is equal to the logarithm of the function proportional to the posterior density found in (a).

   (c) Implement an MCMC algorithm that generates a Markov chain of length 10000 with limiting distribution equal to the posterior for θ. Use a proposal distribution which adds to each parameter a normally distributed variable with expectation zero and standard deviation 0.4. Find a starting value for the chain by studying what values for θ might be reasonable for the given data. Produce trace plots (plots mapping simulated values for θ_i against its index i) for the parameters θ_1, θ_2, and θ_3.

(d) Compute numerically the predicted probability that an animal at pollutant concentration $x = 3$ and temperature $y = 13$ will develop the disease. Also, compute the predicted probability that if 10 animals are exposed to this temperature and this concentration, 9 will develop the disease.

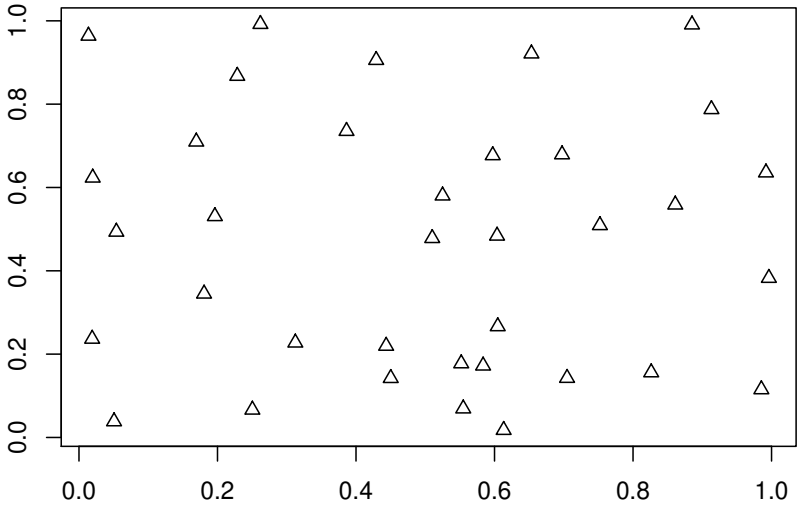

Figure 1: The data used in Question 1 below

2. We consider the positions of trees in a square area. Our model is that the trees are placed in the square $[0, 1] \times [0, 1]$ according to a spatial Poisson process with parameter λ.

   (a) Assume that $\lambda = 36$. Compute the probability that there are 6 trees or more in the area $[0.2, 0.6] \times [0.2, 0.6]$.

   (b) Assume that $\lambda = 36$. Compute the probability that there are exactly 4 trees in the square $[0.2, 0.6] \times [0.2, 0.6]$ and at the same time exactly 4 trees in the square $[0.4, 0.8] \times [0.4, 0.8]$.

   (c) Assume that $\lambda = 36$. Write R code to simulate the spatial Poisson process above, so that your code can output a figure showing the placement of trees in the square $[0, 1] \times [0, 1]$. Show one such example figure.

(d) Now, assume that λ has the prior $\pi(\lambda) \propto_\lambda 1/\lambda$, and that our data are those illustrated in Figure 1, where we have observed 36 trees in a square of size 1. Derive the posterior for λ. Extend your code from (c) to a simulation which uses this posterior instead of a fixed λ.

(e) Consider the stochastic process you simulated from in (d). Let Z be the random variable representing the average over all points of the distance from this point to its nearest neighbour. In other words, if $(X_1, Y_1), (X_2, Y_2), \ldots, (X_K, Y_K)$ are the simulated points, define

$$Z = \frac{1}{K} \sum_{i=1}^{K} \min_{j=1,\ldots,i-1,i+1,\ldots,K} \sqrt{(X_i - X_j)^2 + (Y_i - Y_j)^2}.$$

Use simulation to derive and plot a histogram of a random sample from the distribution of Z.

(f) In the data shown in Figure 1, one can compute that the value of Z is 0.1358. Use this result and your results from (e) to discuss whether the Poisson model is a good model for these tree data, and if not, why not / what should be changed.