Assignment 3, MVE550, autumn 2022

Lisa Samuelsson - *lisasam@student.chalmers.se*
Isak Palenius - *palenius@student.chalmers.se*
Oskar Giljegård - *oskargi@student.chalmers.se*

December 12, 2022

Question 1

A biologist is investigating the frequency with which 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

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

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 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.

All code for this question can be found in the submitted file `A3Q1.R`.

a. Using the model above, 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.

The likelihood for a single data point is the following

$$P(x, y, z|\theta) = P(z|x, y, \theta)P(x, y|\theta)$$

Since θ is just a vector of parameters provided to the function f and x, y are points in our dataset, we know that they are independent, meaning that $P(x, y|\theta) = P(x, y)$. This means that

$$P(x, y, z|\theta) \propto_\theta P(z|x, y, \theta) = \begin{cases} f(x, y, \theta_1, \theta_2, \theta_3) & \text{if } z = 1 \\ 1 - f(x, y, \theta_1, \theta_2, \theta_3) & \text{if } z = 0 \end{cases}$$

By assuming that all data points are independent we can find an expression for the likelihood of all data

$$\begin{aligned} P(data|\theta) &= P(x_1, y_1, z_1, x_2, y_2, z_2, \cdots|\theta) \\ &= P(x_1, y_1, z_1|\theta)P(x_2, y_2, z_2|\theta)\cdots \\ &= \prod_{(x,y,z)\in data} P(x, y, z|\theta) \end{aligned}$$

Using this we can find a function proportional to the posterior

$$\begin{aligned} P(\theta|data) &= \frac{P(data|\theta)P(\theta)}{P(data)} \\ &\propto_\theta P(data|\theta)P(\theta) \\ &\propto_\theta P(data|\theta) \\ &\propto_\theta \prod_{(x,y)\in data} P(x,y,z|\theta) \end{aligned}$$

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 mentioned in (a).

See the code for this in the submitted file `A3Q1.R`.

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 $\theta_1, \theta_2, \theta_3$.

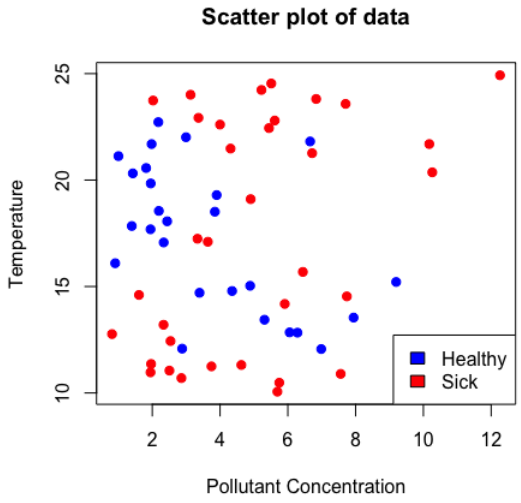


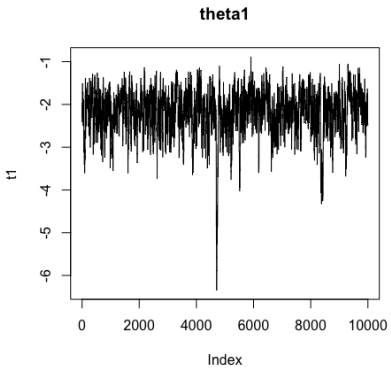
Figure 1: Scatter plot of the data given in the assignment. A red point indicates that the animal is sick, and a blue point indicates that the animal isn't sick.

From studying the data (see Figure 1) we can see that healthy animals tend to

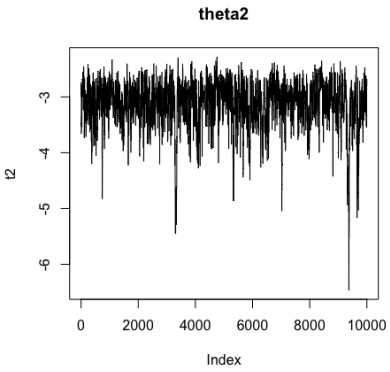
be near $y = 18$. This implies that we should set $\theta_3 = 18$. We also see a slight bias towards more sick animals as the x value increases which means that θ_1 should not be too large of a negative number. The θ_2 parameter only scales the $y = 18$ bias so we do not know exactly what that should be. Therefore we tried starting at $\theta = (-10, -5, 18)$. By running the MCMC algorithm multiple times and seeing where the θ tended to get a high likelihood for the data we eventually decided to use $\theta = (-2, -3, 18)$ as initial value.

See the code for this in the submitted file `A3Q1.R`.

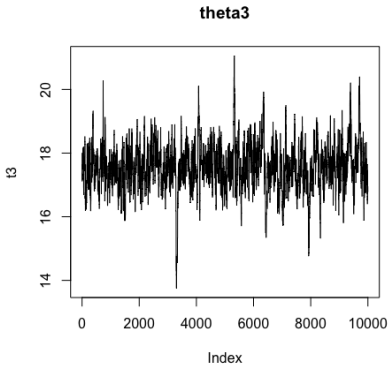
The trace plots for simulated θ_1, θ_2 and θ_3 can be seen in Figure 2.



(a) Trace plot for θ_1.



(b) Trace plot for θ_2.



(c) Trace plot for θ_3.

Figure 2: The three trace plots generated.

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.

Computing $f(x, y, \theta_1, \theta_2, \theta_3)$, where $x = 3$ and $y = 13$ for each of the sampled θ values and taking the mean of those results gave 0.5822 as the predicted probability that the animal would develop the disease.

To compute the predicted probability that 9 out of 10 animals would develop the disease we can use a binomial distribution. Running the R code `dbinom(9, 10, theta)` for each sampled θ and then taking the mean of this gave us the probability 0.0554 that 9 out of 10 animals develop the disease.

Question 2

All code for this question can be found in the submitted file `A3Q2.R`.

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

Let A be the area $[0.2, 0.6] \times [0.2, 0.6]$.

Then $|A| = (0.6 - 0.2)(0.6 - 0.2) = 0.16$

$N_A \sim Poisson(\lambda|A|) = Poisson(36 \cdot 0.16) = Poisson(5.76)$

$P(N_A \geq 6) = 1 - P(N_A < 5) = 1 - ppois(5, 5.76) \approx 0.5150434$

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 $[0.4, 0.8] \times [0.4, 0.8]$.

Let

$$\begin{aligned} A &= [0.2, 0.6] \times [0.2, 0.6] \\ B &= [0.4, 0.8] \times [0.4, 0.8] \\ AB &= [0.4, 0.6] \times [0.4, 0.6] \end{aligned}$$

We then have no overlap between $A - AB$ and $B - AB$.

We know

$$\begin{aligned} |AB| &= 0.04 \\ |A - AB| &= 0.12 \\ |B - AB| &= 0.12 \end{aligned}$$

Then the probability of 4 trees in each can be broken down into

$$\begin{aligned} &P(N_A = 4, N_B = 4) \\ &= P(N_{A-AB} + N_{AB} = 4, N_{B-AB} + N_{AB} = 4) \\ &= \sum_{k=0}^{4} P(N_{A-AB} + N_{AB} = 4, N_{B-AB} + N_{AB} = 4 | N_{AB} = i) P(N_{AB} = i) \\ &= \sum_{k=0}^{4} P(N_{A-AB} = 4 - i, N_{B-AB} = 4 - i) P(N_{AB} = i) \\ &= \sum_{k=0}^{4} P(N_{A-AB} = 4 - i) P(N_{B-AB} = 4 - i) P(N_{AB} = i) \end{aligned}$$

Implementing this in R (see `A3Q2.R`), we found that the probability was approximately 0.0239.

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.

See Figure 3 below.

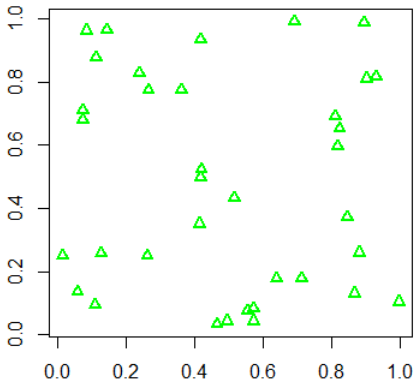


Figure 3: A possible placement of trees in the square [0,1] $\times$ [0,1].

d. Now, assume that λ has the prior $\pi(\lambda) \propto_\lambda \frac{1}{\lambda}$, 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.

$$
\begin{aligned}
\pi(\lambda) &\propto_\lambda \frac{1}{\lambda} \\
\pi(data|\lambda) &= P(N_{ALL} = 36) \\
&= Poisson(36; \lambda) \\
&= \frac{e^{-\lambda}\lambda^{36}}{36!} \\
\pi(\lambda|data) &\propto_\lambda P(data|\lambda)\pi(\lambda) \\
&\propto_\lambda e^{-\lambda}\lambda^{35} \\
&\propto_\lambda Gamma(36, 1)
\end{aligned}
$$

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 $(X_1, Y_1), (X_2, Y_2), ..., (X_K, Y_K)$ are the simulated points, define

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

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

See Figure 4 below.

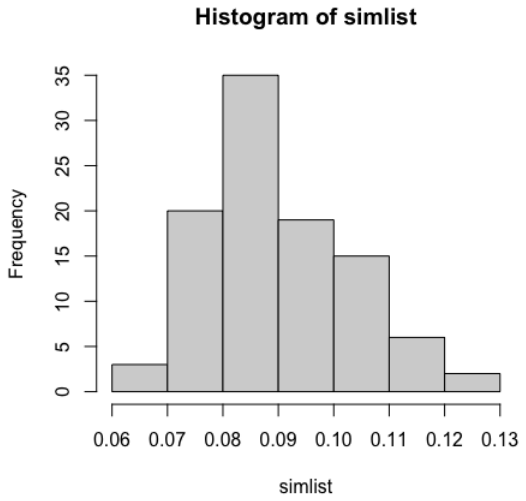


Figure 4: A histogram of a random sample from the distribution of Z.

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.

The main issue with the Poisson distribution is that it does not take into consideration the placement of the already existing trees. In nature, we would not expect multiple trees to grow on top of eachother, however, this could happen with our model. If a group of trees are to close to eachother they would likely compete for resources and some of them would die out. If the trees are not manually planted we might also expect trees to somewhat close to other trees (since threes do not appear randomly). Since we have no distance unit model

it is impossible to determine if the distances between the trees are a couple of meters or a couple of kilometers (which would be unrealistic). This however, could be solved by just increasing the value for lambda. Although if we want to simulate an area very sparse in trees we would expect the trees to be more grouped up.

g. Determine the name and the parameters of the distribution of N, the number of points that is simulated by the process in (d).

By the Gamma-Poisson conjugacy we know that

$$\pi(\lambda|N, data) = Gamma(36 + N, 2)$$

We can then calculate

$$\begin{aligned}
\pi(N|data) &= \frac{\pi(N|\lambda, data)\pi(\lambda|data)}{\pi(\lambda|N, data)} \\
&= \frac{Poisson(N;\lambda)Gamma(\lambda;36,1)}{Gamma(\lambda;36+N,2)} \\
&= \frac{\frac{e^{-\lambda}\lambda^N}{N!} \quad \frac{1^{36}}{\Gamma(36)}\lambda^{35}e^{-\lambda}}{\frac{2^{36+N}}{\Gamma(36+N)}\lambda^{35+N}e^{-2\lambda}} \\
&= \frac{\Gamma(36+N)}{N!\Gamma(36)2^{36+N}} \\
&= NegativeBinomial(36, 1/(1+1)) \\
&= NegativeBinomial(36, 1/2)
\end{aligned}$$