Bayensian Learning

Homework 3: Bayesian Prediction using
Markov Chain Monte Carlo Approach

Emilie Krutnes Engen

Master in Big Data Analytics
Carlos III University of Madrid

Universidad
Carlos III de Madrid

Bayesian Prediction of Antarctic Glacier Discharge

In this exercise we use Bayesian inference and prediction for Antarctic glacier discharge using a non-conjugate prior model.

The Data

The dataset is provided by GLACKMA (2009) and contains hourly glacier discharge records from 2009 in King George Island. Glacier discharge is recorded liquid loss through surface melting and run off, or melting at the base as a result of global warming. The summary statistics for the discharge data is presented in Table 1. The distribution of the discharge data is presented in a histogram in Figure 1. From the plot we observe that the data is clearly right skewed with a mean of 0.2120.

Min.	Q1	Median	Mean	Q3	Max.
0.0610	0.1440	0.2120	0.2354	0.3000	0.7400

Table 1: Summary statistics for the glacier discharge data

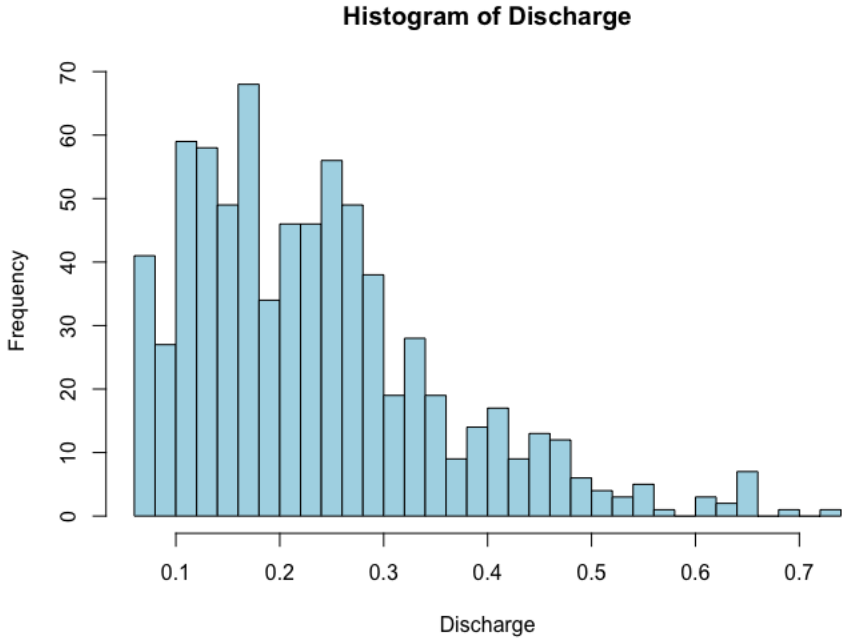

Figure 1: Histogram showing the distribution of the observed discharge data

We assume that the discharge, Y, follows a Weibull distribution, $Y|k,\theta \sim W(k,\theta)$. The density function is given by

$$f(y|\kappa,\theta) = \kappa\theta y^{\kappa-1} exp(-\theta y^{\kappa}), \quad y > 0 \tag{1}$$

The Prior

The Weibull distribution have no conjugate prior. As a result we assume that the prior distribution is given by

$$\kappa \sim Uniform(\kappa_{min}, \kappa_{max}) \tag{2}$$

$$\theta \sim Gamma(a, b) \tag{3}$$

The Posterior

Given a sample of the glacier discharge data $\boldsymbol{y} = y_1, \ldots, y_n$ the joint posterior distribution is an unknown distribution. We further have that the conditional posterior of the shape, κ, is proportional to

$$f(\kappa|\theta, \boldsymbol{y}) \propto \kappa^n \prod_{i=1}^{n} y_i^{\kappa-1} exp(-\theta \sum_{i=1}^{n} y_i^{\kappa}) \tag{4}$$

which is an unknown distribution. However, it can be proven that the conditional posterior distribution for the scale θ is a semi-conjugate to the prior. Therefore we have that the conditional posterior for θ follows a gamma distribution.

$$\theta|\kappa, \boldsymbol{y} \sim Gamma(a + n, b + \sum_{i=1}^{n} y_i^{\kappa}) \tag{5}$$

Because the structure of the posterior for κ is unknown, we implement a Metropolis within Gibbs algorithm step to sample from the joint posterior of $(\kappa, \theta|\boldsymbol{y})$.

Sampling from the Markov Chain Monte Carlo Algorithm

The Metropolis within the Gibbs algorithm is given by the following steps.

For $t = 1, \ldots, T$ repeat the following:

1. Construct a sample for θ:

   $\theta_t \sim Gamma\Big(a + n, b + \sum_{i=1}^{n} y_i^{\kappa_{t-1}}\Big)$

2. Generate a candidate $\tilde{\kappa} \sim N(\kappa_{t-1}, \sigma)$

   - Reject if $\tilde{\kappa} < \kappa_{min}$ or $\tilde{\kappa} > \kappa_{max}$

- Sample $u \sim U(0,1)$ and accept $\tilde{\kappa}$ if:

$$u < \frac{\tilde{\kappa}^n \prod_{i=1}^n y_i^{\tilde{\kappa}-1} exp(-\theta_t \sum_{i=1}^n y_i^{\tilde{\kappa}})}{\kappa_{t-1}^n \prod_{i=1}^n y_i^{\kappa_{t-1}-1} exp(-\theta_t \sum_{i=1}^n y_i^{\kappa_{t-1}})}$$

From the simulated Markov Chain we can obtain plots of the traces for κ and θ.

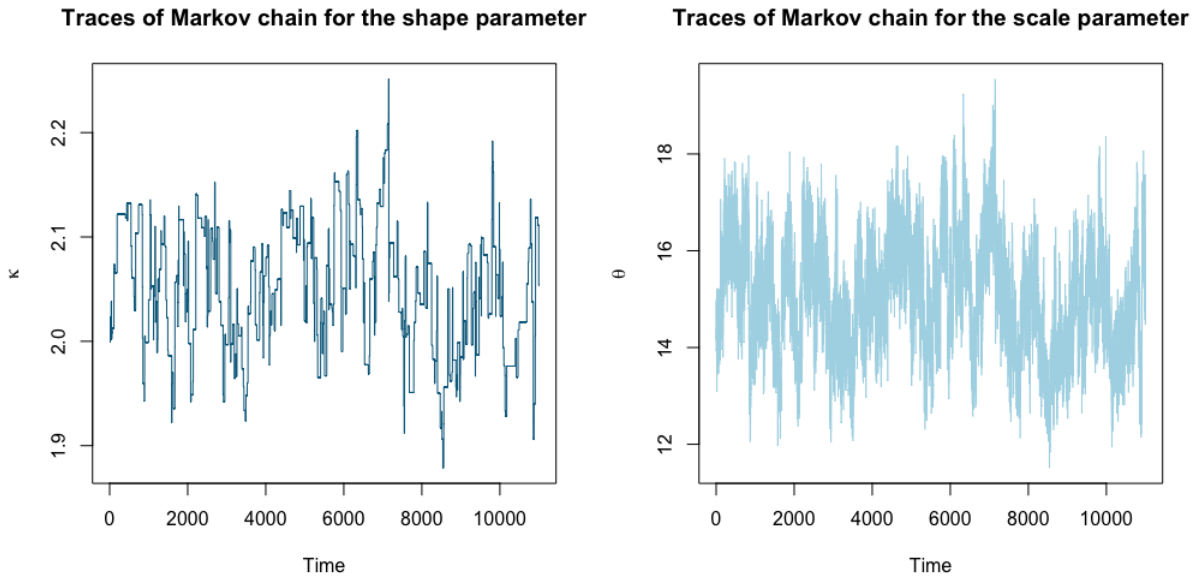

Figure 2: Traces of the shape parameter κ and scale θ from the simulated Markov Chain

We further obtain an approximation of the posterior distribution from the parameters κ and θ. The distributions are given in the histograms presented in Figure 3. Both parameters show a posterior distribution close to normal.

In Table 2 the 95 % credible intervals, the mean and standard deviation are given for the two parameters κ and θ.

Parameter	Mean	St. Dev.	Cred. Int.
κ	2.0475	0.0613	[1.9377, 2.1604]
θ	14.9751	1.1436	[12.9052, 17.2514]

Table 2: Summary statistics for the posterior parameters

Posterior Predictive Distribution

We now want to estimate the predictive probability that a future glacier discharge Y_{n+1} is larger than $1\text{m}^3/\text{sec} \cdot \text{km}^2$, given the recorded discharge data, $\boldsymbol{y}$.

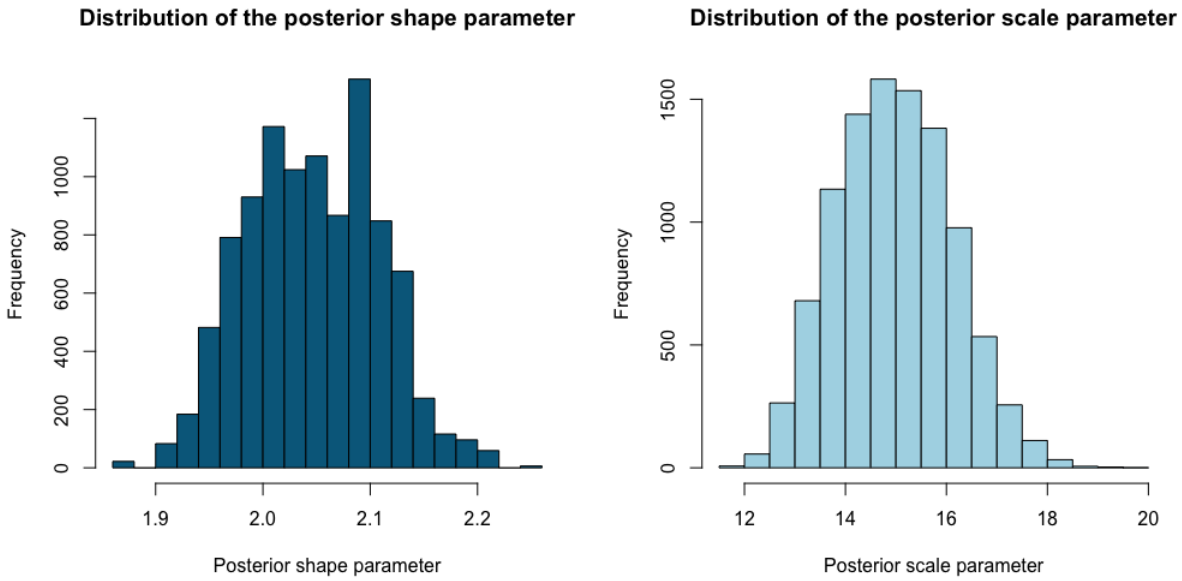

Figure 3: Traces of the shape parameter κ and scale θ from the simulated Markov Chain

The predictive density can be given by

$$f(y_{n+1}|\boldsymbol{y}) = \int f(y|\kappa,\theta) f(\kappa,\theta|\boldsymbol{y}) d\kappa d\theta \tag{6}$$

This is not considered a good expression for the predicted density as it cannot be solved analytically. Instead we use the Markov Chain Monte Carlo posterior sample to approximate the predictive probabilities. By using a sample size $M = 10,000$ from the predictive distribution we obtain the histogram presented in Figure 4.

We further compare the recorded data with the estimated predictive density, presented in Figure 5. The histogram represent the observed data, while the solid line represent the predicted density. From the plot we observe that the predicted density put a higher emphasis on lower discharge values than the observed data.

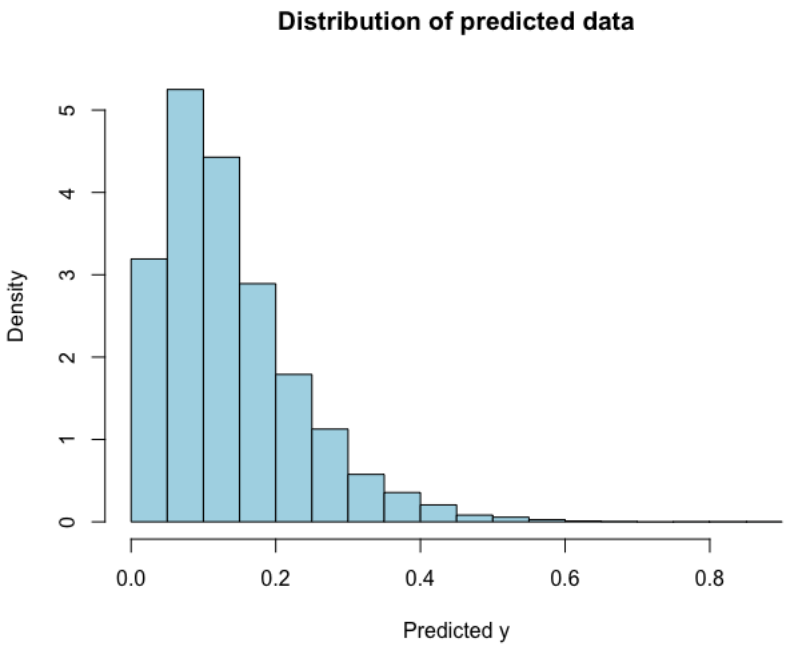

Figure 4: Histogram showing the predictive distribution obtained from the MCMC posterior sample

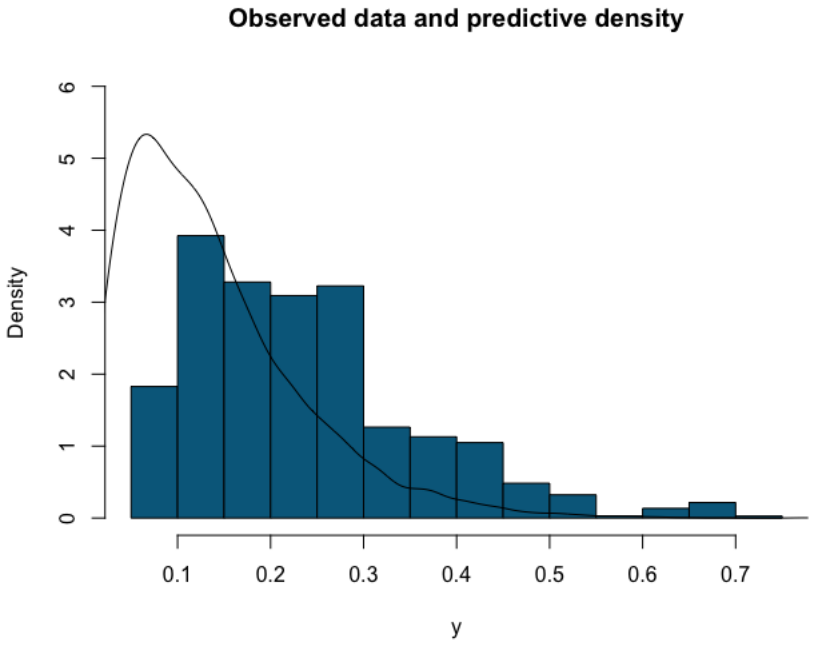

Figure 5: Histogram of the observed data and predicted density

Then we approximate the predictive probability of a discharge larger than $1\text{m}^3/\text{sec} \cdot \text{km}^2$, $p(Y_{n+1} > 1|\boldsymbol{y})$, by calculating the mean of the sampled values larger than 1. From this we get that the predictive probabilities of an extreme discharge is 0. From the observed data we have that the maximum discharge is 0.74. As the data used for this study only includes data from January 2009, the dataset does not reflect seasonal variations. Thus it is not likely that we will observe any extreme discharge values during the winter season. To perform a more accurate prediction more data should be considered. We further notice that assuming Weibull distributed observations for the discharge in Antarctica might not be realistic. One should therefore consider approaches based on time series models. In addition, it might be appropriate to investigate the inclusion of other variables, such as temperature. It is reasonable to believe that these variables are correlated and that the discharge is affected by the recorded temperatures.

References

GLACKMA, G. C. y. M. A. (2009). Antarctica glacier discharge data. http://www.glackma.org/.