SF2955, Home Assignment 2, Complex MCMC

Authors:
Eric Leijonmarck
870820-0095
ericle@kth.se

Victor Tingström
911109-0651
vtin@kth.se

Supervisor:
Jimmy Olsson
jimmyol@kth.se

May 15, 2014

Coal mine disasters - Constructing a complex MCMC algorithm

Introduction

For this exercise, we analyze a time series containing the British coal mining disasters under the time period 1851–1962. The difference between this exercise and the one in the course literature, is that we have a continuous time series as well as more than 1 **breakpoint**. A breakpoint is thus the year at which the intensity of the disasters change. We will use $d-1$ breakpoints, where d then corresponds to the number of intervals we will use. In order to get a feel for the time series we will analyze, a figure containing a histogram of the disasters is found in figure 1.

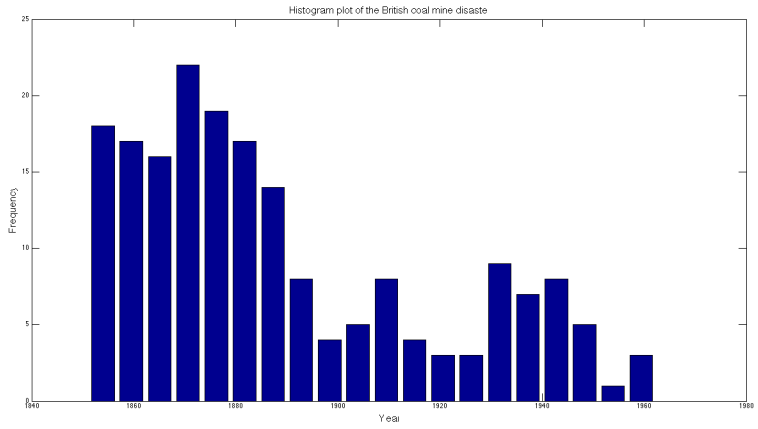


Figure 1: Histogram plot of the British coal mine disasters in the time period 1851–1962.

As is seen in figure 1 there is a change in the disaster intensity at the turn of the century, perhaps due to some legislation regarding work safety or some technological advancement.

To begin the exercise, we define the vector $\boldsymbol{t}$ to be the vector containing all of the breakpoints t_i, $i = 2, \dots d$ aswell as the start and end point $t_1 = 1851$ and $t_{d+1} = 1963$. We wish to model the disasters using an inhomogenous Poisson process with an intensity λ_i for each of the intervals $[t_i, t_{i+1})$, $i = 1, \dots, d$. Where all of the λ_i's are collected in a vector $\boldsymbol{\lambda}$.
We will denote the time series containing the year at which disaster struck as $\boldsymbol{\tau} = (\tau_1, \dots, \tau_n)$ for $n = 191$, where the subscript denotes accident i.
We then define the number of accidents under the interval $[t_i, t_{i+1})$ to be

$$n_i(\boldsymbol{\tau}) = \sum_{j=1}^{n} \mathbb{1}\{[t_i, t_{i+1})\} \cdot \tau_j$$

We set a $\Gamma(2, \theta)$ prior on the intensities, λ_i, and a $\Gamma(2, \beta)$ hyperprior on θ. Where β is a hyperparameter that needs to be specified. Furthermore, we put the prior

$$f(\boldsymbol{t}) \propto \begin{cases} \prod_{i=1}^{d}(t_{i+1} - t_i), & \text{for } t_1 < t_2 < \cdots < t_{d+1} \\ 0, & \text{else} \end{cases}$$

This prior prevents the breakpoints from being located to closely. All of these prior assumptions then imply that

$$f(\boldsymbol{\tau}|\boldsymbol{\lambda}, \boldsymbol{t}) \propto \prod_{i=1}^{d} \lambda_i^{n_i(\boldsymbol{\tau})} \cdot \exp\left\{ -\sum_{i=1}^{d} \lambda_i (t_{i+1} - t_i) \right\}$$

In order to sample from the posterior $f(\theta, \boldsymbol{t}, \boldsymbol{\lambda}|\boldsymbol{\tau})$ we will construct a hybrid **M**arkov **C**hain **M**onte **C**arlo algorithm, where the hybrid comes from the fact that we will need to sample $\boldsymbol{t}$ using a Metropolis–Hastings step whereas the other components can be updated using a Gibbs sampler.

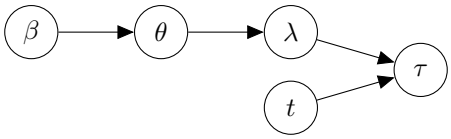
There are several ways to choose the proposal distribution for the Metropolis–Hastings, we chose to use the *Random walk proposal*, which means that we will update one breakpoint at a time and for each breakpoint t_i generate a candidate t_i^* according to

$$t_i^* = t_i + \epsilon, \quad \epsilon \sim \text{Unif}(-R, R)$$

Where $R = \rho(t_{i+1} - t_{i-1})$ and ρ is a tuning parameter.

a)

For this exercise, we are supposed to find the marginal posteriors for $f(\theta|\boldsymbol{\tau}, \boldsymbol{t}, \boldsymbol{\lambda})$, $f(\boldsymbol{\lambda}|\theta, \boldsymbol{t}, \boldsymbol{\tau})$ and $f(\boldsymbol{t}|\theta, \boldsymbol{\lambda}, \boldsymbol{\tau})$. We began this exercise by identifying the different dependences , which are found in the following figure.



Now that we've identified all of the dependences, we begin by analyzing the first posterior.

I. We begin by rewriting the expression using *Bayes' Theorem*

$$f(\theta|\boldsymbol{\tau}, \boldsymbol{t}, \boldsymbol{\lambda}) \propto f(\theta) \cdot f(\boldsymbol{\tau}, \boldsymbol{t}, \boldsymbol{\lambda}|\theta)$$

The second term is rewritten as

$$f(\boldsymbol{\tau}, \boldsymbol{t}, \boldsymbol{\lambda}|\theta) = f(\boldsymbol{\tau}|\boldsymbol{\lambda}, \boldsymbol{t}, \theta) \cdot f(\boldsymbol{t}, \boldsymbol{\lambda}|\theta)$$

We then notice that there's independence between some of the variables, which then means that the expression is rewritten into

$$f(\boldsymbol{\tau}|\boldsymbol{\lambda},\boldsymbol{t})\cdot f(\boldsymbol{t})\cdot f(\boldsymbol{\lambda}|\theta)\propto f(\boldsymbol{\tau}|\boldsymbol{\lambda},\boldsymbol{t})\cdot f(\boldsymbol{\lambda}|\theta)$$

Inserting this expression then yields

$$f(\theta|\boldsymbol{\tau},\boldsymbol{t},\boldsymbol{\lambda})\propto f(\theta)\cdot f(\boldsymbol{\lambda}|\theta)\cdot f(\boldsymbol{\tau}|\boldsymbol{\lambda},\boldsymbol{t})$$

The explicit expression for the distribution then becomes

$$f(\theta|\boldsymbol{\lambda},\boldsymbol{t},\boldsymbol{\tau})\propto\theta\cdot\exp\{-\beta\cdot\theta\}\cdot\theta^{2d}\prod_{i=1}^{d}\lambda_i\cdot\exp\left\{-\theta\sum_{i=1}^{d}\lambda_i\right\}\prod_{i=1}^{d}\lambda_i^{n_i(\boldsymbol{\tau})}\cdot\exp\left\{-\sum_{i=1}^{d}\lambda_i(t_{i+1}-t_i)\right\}$$

We then identify the terms containing θ and finally get that

$$f(\theta|\boldsymbol{\lambda},\boldsymbol{t},\boldsymbol{\tau})\propto\theta^{2d+1}\exp\left\{-\theta\cdot\left(\beta+\sum_{i=1}^{d}\lambda_i\right)\right\}\sim\Gamma\left(2(d+1),\beta+\sum_{i=1}^{d}\lambda_i\right)$$

II. For economy of text, the entire proof of the distribution will be left out, the procedure is the same as in I.

$$f(\boldsymbol{\lambda}|\boldsymbol{\tau},\boldsymbol{t},\theta)\propto f(\boldsymbol{\tau}|\boldsymbol{t},\boldsymbol{\lambda})\cdot f(\boldsymbol{\lambda}|\theta)\cdot f(\theta)$$

Whose explicit expression is

$$f(\boldsymbol{\lambda}|\boldsymbol{\tau},\boldsymbol{t},\theta)\propto\prod_{i=1}^{d}\lambda_i^{1+n_i(\boldsymbol{\tau})}\cdot\exp\left\{-\sum_{i=1}^{d}(\theta+(t_{i+1}-t_i))\lambda_i\right\}$$

Which implies that

$$\lambda_i\sim\Gamma\left(2+n_i(\boldsymbol{\tau}),\theta+(t_{i+1}-t_i)\right)$$

III. We are now supposed to calculate the posterior for

$$f(\boldsymbol{t}|\boldsymbol{\tau},\boldsymbol{\lambda},\theta)$$

If one proceeds as in the previous examples, we will see that we cannot determine the distribution of $\boldsymbol{t}$ explicitly. This then means that we need to use the Metropolis–Hastings algorithm in order to sample $\boldsymbol{t}$. If one proceeds as in the previous examples, we will get that

$$f(\boldsymbol{t}|\boldsymbol{\tau},\boldsymbol{\lambda},\theta)\propto\prod_{i=1}^{d}(t_{i+1}-t_i)\lambda^{n_i(\boldsymbol{\tau})}\cdot\exp\left\{-\sum_{i=1}^{d}\lambda_i(t_{i+1}-t_i)\right\}\sim\text{☹}$$

b)

Here we construct our hybrid MCMC algorithm to sample from the posterior $f(\theta, \boldsymbol{\lambda}, \boldsymbol{t}|\boldsymbol{\tau})$. All components except the breakpoints $\boldsymbol{t}$ have been updated using Gibbs sampling, which is explained in algorithm **??**. The Metropolis–Hastings algorithm is explained below and has been used for the posterior of $\boldsymbol{t}$. Where we, as already mentioned, used a symmetric proposal for the proposal kernel. We will now briefly discuss the two different algorithms used for sampling.

- The **Metropolis–Hastings** algorithm is an algorithm that is used for sampling from high–dimensional and/or complicated distributions that are for example only known up to a normalizing constant. The idea behind the MH algorithm is to choose a transition density q almost arbitrarily. However, this density will not give the desired asymptotic distribution f, i.e. the distribution from which you wish to sample. We correct this by introducing a new transition density $\hat{q}$ using q and a probability $\alpha(x, x^*)$ where x is the current level and x^* is a proposal. The proposal x^* is rejected with probability $1 - \alpha(x, x^*)$. In order for the global balance equation to be satisfied, we must have that $\alpha(\cdot, \cdot)$ is defined as

  $$\alpha(x, x^*) = \min\left(1, \frac{f(x^*)q(x|x^*)}{f(x)q(x^*|x)}\right).$$

  Since we now have a quotient, we can easily sample from distributions known only up to normalizing constants. In our case we have, as already mentioned, a symmetric proposal. Which practically means that $q(x^*|x) = q(x|x^*)$ and the two right terms of the quotient cancel out. The MH algorithm is explained briefly in figure 2.

1. Choose a starting value $x^{(0)}$.
2. Repeat for $i = 1, \ldots, N$:
   - i.1 Draw y^* from $q(x^{(i-1)}, \cdot)$.
   - i.2 Draw u from $U(0, 1)$.
   - i.3 If $u < \alpha(x^{(i-1)}, y^*)$ set $x^{(i)} = y^*$, else set $x^{(i)} = x^{(i-1)}$.
3. $x^{(1)}, x^{(2)}, \ldots, x^{(N)}$ is now a sequence of dependent draws, approximately from f.

Figure 2: Overview of how the Metropolis–Hastings algorithm works. The figure is taken from the lecture notes by Martin Sköld.

- The **Gibbs sampler** algorithm is a Markov Chain Monte Carlo (MCMC) algorithm for obtaining a sequence of observations, which in turn can be approximated from a multivariate probability distribution. Typically the Gibbs sampler is used to approximate joint distributions, marginal distributions of one variable or computing an

integral, when some of the variables are known observations, and is not in need of sampling. As this a MCMC algorithm, Gibbs sampling generates a Markov Chain of samples, hence the beginning of the chain may not accurately describe the desired distribution and one may want to remove the burn-in of the chain. Below a brief description of the algorithm is presented

1. Assuming we want to sample from multivariate distribution f on X. Assuming that X can be divided into blocks.
2. We denote x^k by the kth component and x^{-k} by the remaining components. Assuming that it is easy to simulate from $f_k(x^k|x^{-k}), \forall k$, i.e. the conditional distribution of X^k given the other components.
3. Now, simulating a sequence X_k, forming a Markov Chain on X, with the following properties: Given X_k, draw $X_{k+1}^n \sim f_n(x^n|x^{-n})$.

The symmetric proposal yields a correct MCMC algorithm, as long as the proposal kernel specifies an irreducible aperiodic Markov Chain. The algorithm will generate a reversible Markov Chain with stationary distribution π, assuming that χ is a subset of $\mathbb{R}^k$ and the transition kernel $r(z|x)$ and $\pi(z)$ are densities on χ.

The proposed kernel yields a correct MCMC as we have introduced a symmetric proposal with transition densities shown in equation (**??**),

$$r(z|x) = r(x|z), \quad \forall (x, z) \in t \tag{1}$$

implying that we can reach each state from any state and is non-periodic, i.e. a irreducible aperiodic Markov Chain.

d)

In this exercise, we are supposed to investigate the sensitivity of the posteriors when we vary the hyperparameter β. We will investigate what happens for $\beta = 0, 1, 20$ with a constant $\rho = 0.05$. In order to investigate the sensitivity of the choice of hyperparameter we chose to simulate the chain 25 times for each value of β with a $\text{burn}_i nof 10,000 and 10,000 samples and two breakpoints. Th$

Parameters λ_3	θ	t_2	t_3	λ_1	λ_2
$\beta = 0$ 0.0063	0.0026	1.86	81.37	$2.9 \cdot 10^{-4}$	0.013
$\beta = 1$ 0.0035	$7.3 \cdot 10^{-4}$	0.25	47.9	$5.9 \cdot 10^{-5}$	0.0052
$\beta = 20$ 0.0031	$2.43 \cdot 10^{-5}$	10.2	125.4	0.0068	0.091

Parametric bootstrap method

Introduction

The problem consists of a dataset containing significant wave-heights recorded 14 times a month during several winter months in the north Atlantic. This is could for example be used to predict the probability of high waves in the north atlantic to warn oil platforms. One can estimate the extreme value of this dataset by assuming that the data has a *Gumpel Distribution* with distribution function

$$F(x;\mu,\beta) = \exp\left(-\exp\left(-\frac{x-\mu}{\beta}\right)\right), \quad x \in \mathbb{R},$$

where $\mu \in \mathbb{R}$ and $\beta > 0$. The paramaters of this distribution for an arbitrary dataset was estimated using the matlab function `est_gumbel.m`.

During a 100 year interval, one can calculate a estimate of the likelihood of an 100-year return period event. It is a statistical measurement typically based on historic data denoting the average recurrence interval over an extended period of time. The analysis assumes that the probability does not vary in time and is independant of past events.

The expected 100-year return value of the data gives the largest expected wave-height during a 100-year period. The Tth return value is denoted by $F^{-1}(1-1/T;\mu,\beta)$. Since the data has been observed during 14 times a month and assuming we have three winter months during a year, T is $T = 3 \cdot 14 \cdot 100$. The investigated dataset is shown below in figure **??**.

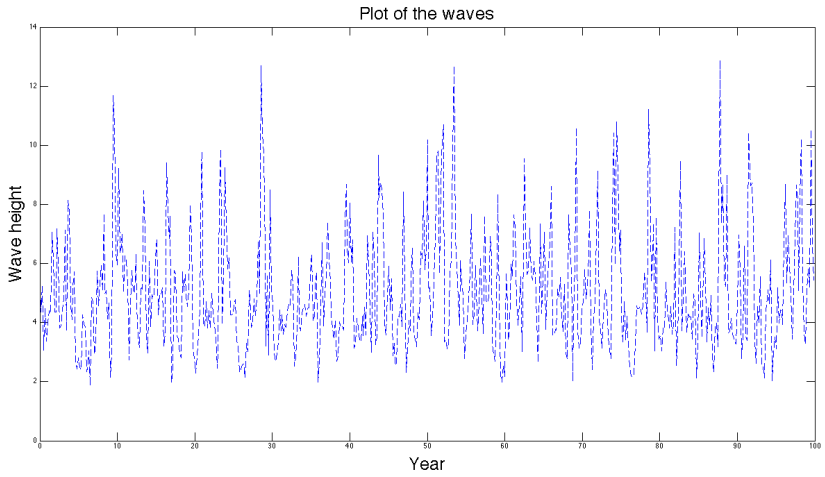


Figure 3: A figure, showing the data of the significant observed wave-heights

o perform a statistical test on the dataset, we use a parametric bootstrap approach. The bootstrap technique evaluates the uncertainty of an unknown distribution or data. The bootstrap replaces the unknown statistic by data-based approximations and analyzes the variation using MC simulation from the approximation. The approximations are done using a **empirical distribution** (ED) associated with the data and gives equal weights to each observed value. Below we will present a brief description of a general bootstrap method.

- For a given statistic y, we replace $\mathbb{P}_0$ by $\hat{\mathbb{P}}_0$.
- Approximation can be done by plugging $\hat{\mathbb{P}}_0$ into the quantity, i.e.

$$\tau = \tau(\mathbb{P}_0) \approx \hat{\tau} = \tau(\hat{\mathbb{P}}_0)$$

  .

- Uncertainty of $t(y)$ is analyzed by looking at the variation of $\Delta(Y^*) = t(Y^*) - \hat{\tau}$ by drawing repeatedly $Y^* \sim \hat{\mathbb{P}}_0$.

In our case, we have done a parametric bootstrap, where we assume that the data comes from a distribution $\mathbb{P}_0 = \mathbb{P}_{\theta_0} \in \{\mathbb{P}_\theta; \theta \in \Theta\}$ belonging to some parametric family. Instead of using the ED, we find an estimate $\hat{\theta} = \hat{\theta}(y)$ of θ_0 from the observations and

1. generate new bootstrapped samples $Y_b^*, b \in \{1, 2, \ldots, B\}$, from $\hat{\mathbb{P}}_0 = \mathbb{P}_{\hat{\theta}}$.
2. then we form bootstrap estimates $\hat{\theta}(Y_b^*)$ and errors $\Delta_b^* = \hat{\theta}(Y_b^*) - \hat{\theta}, b \in \{1, 2, \ldots, B\}$.

For this assigment we approximated the data as being that of *Gumpel distributed* data points. In the figure **??** we have plotted the histogram of the atlantic wave-heights, indicating that the data has a a *Gumpel* distribution.

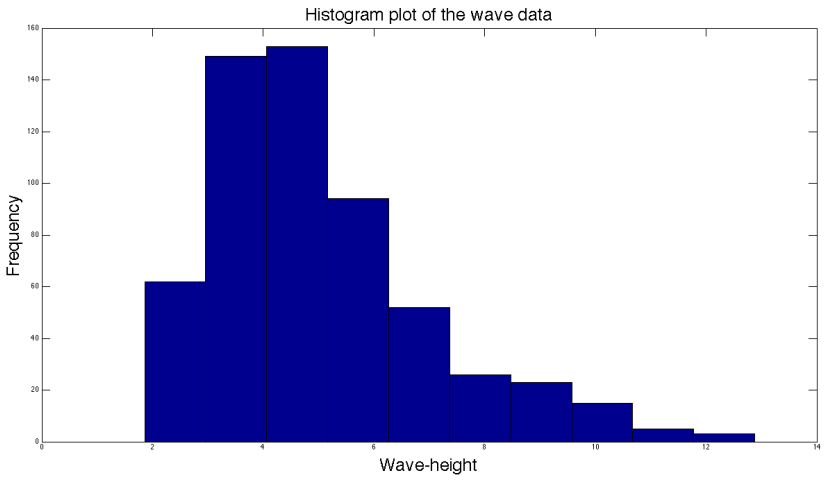


Figure 4: A figure, showing the data of the significant observed wave-heights

Parametric Bootstrap for the 100-year Atlantic wave

a)

We will here find the inverse of the atlantic wave distribution, approximated as a *Gumpel Distribution*. To begin finding an inverse we insert a new variable for the function and solve for x. This is done below

$$F(x;\mu,\beta)) = \exp\left(-\exp\left(-\frac{x-\mu}{\beta}\right)\right), x \in \mathbb{R}$$

Here we insert u for the function

$$u = \exp\left\{-\exp\left(-\frac{x-\mu}{\beta}\right)\right\} \Rightarrow \ln(u) = -\exp\left(-\frac{x-\mu}{\beta}\right) \Rightarrow$$
$$\ln\left\{\ln\left(\frac{1}{u}\right)\right\} = -\frac{x-\mu}{\beta} \Rightarrow \mu - \beta\ln\left\{\ln\left(\frac{1}{u}\right)\right\} = x.$$