SF2955, Home Assignment 2, Complex MCMC

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

Victor Tingström — *vtin@kth.se*
911109-0651

Supervisor:
Jimmy Olsson — *jimmyol@kth.se*

May 14, 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.

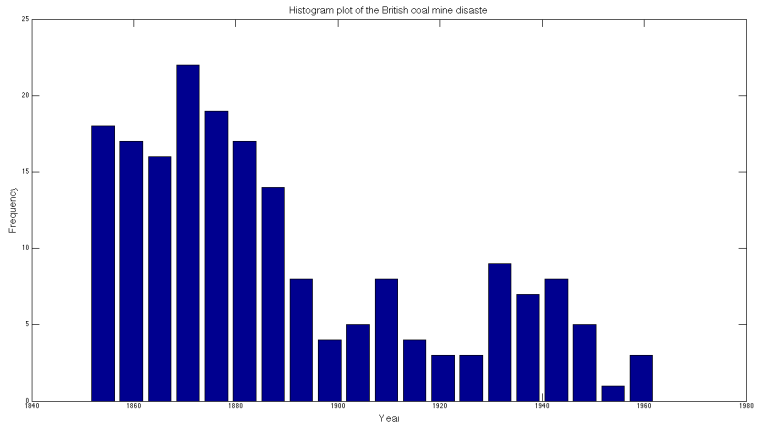


Figure 1: Histogram plot of the British coal mine disasters in the time period 1851–1962.

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.

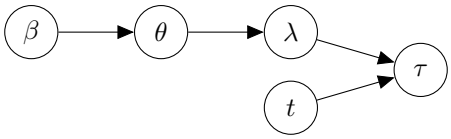
There are several ways to choose the proposal distribution for the Metropolis–Hastings, we chose to use the *Random walk proposal*, which means that we will update one breakpoint at a time and for each breakpoint t_i generate a candidate t_i^* according to

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

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 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.

c)

To start the exercise, we consider the amount of necessary samples used as burn-in, which we set to $N = 10,000$ samples. The following figure shows the draws of $\boldsymbol{t}$ for $d = 2$, i.e. one breakpoint.

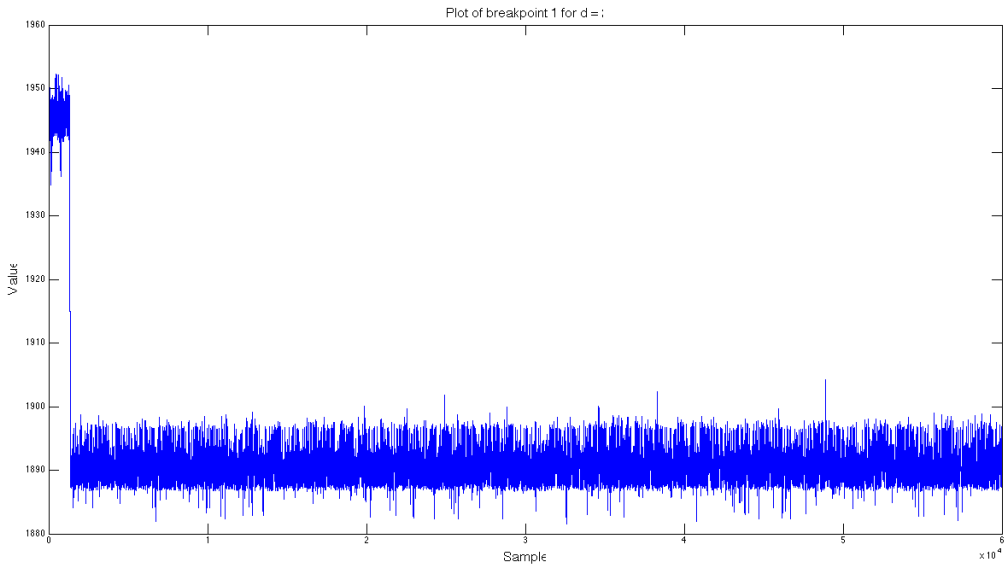


Figure 3: Figure of the draws of $\boldsymbol{t}$

As the figures shows, we have stationarity after about $M = 3,000$ samples, meaning that our burn-in is reasonable.

In assignment **c)** we will investigate the behaviour of the chain for our posterior $f(\theta, \boldsymbol{\lambda}, \boldsymbol{t}|\boldsymbol{\tau})$ for $1, 2, 3$ and 4 breakpoints in the data. We will show that the posterior of $\boldsymbol{t}$ for the intervals will be affected in various ways depending on the density for each

interval. We analyze the number of breakpoints individually with $\beta = 1$ and $\rho = 0.05$.

One Breakpoint

If we divide the time period into two intervals we get that the breakpoint is located at ≈ 1891. Figure 4 shows where the location of the breakpoint is located relative to the dataset in order to get a better overview.

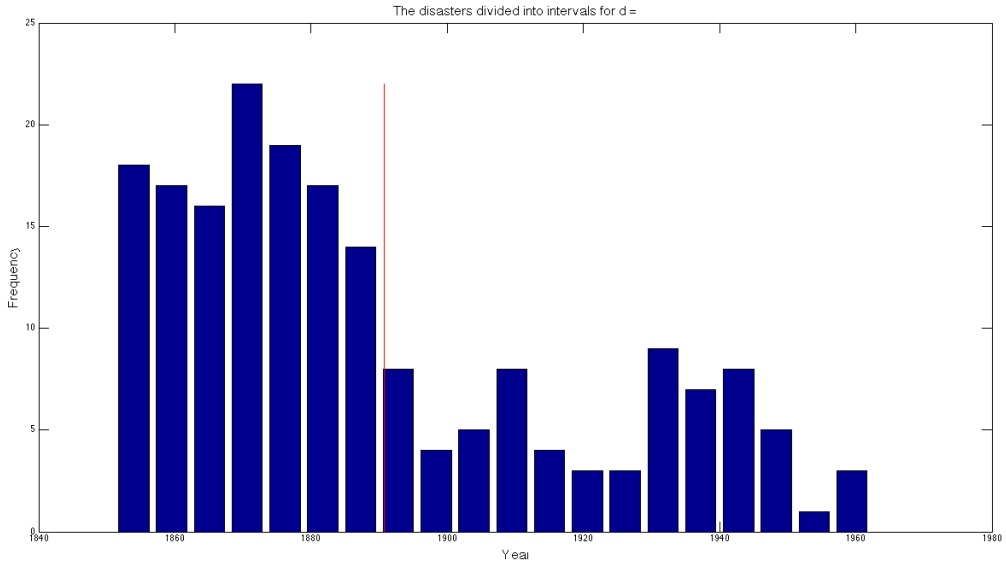


Figure 4: Plot of where the breakpoint is located.

In the intro we said that the breakpoint should lie somewhere at the turn of the century, and as is seen in figure 4 we have a very reasonable result as to where the breakpoint lies. We now look at the histogram of the breakpoint in figure 5.

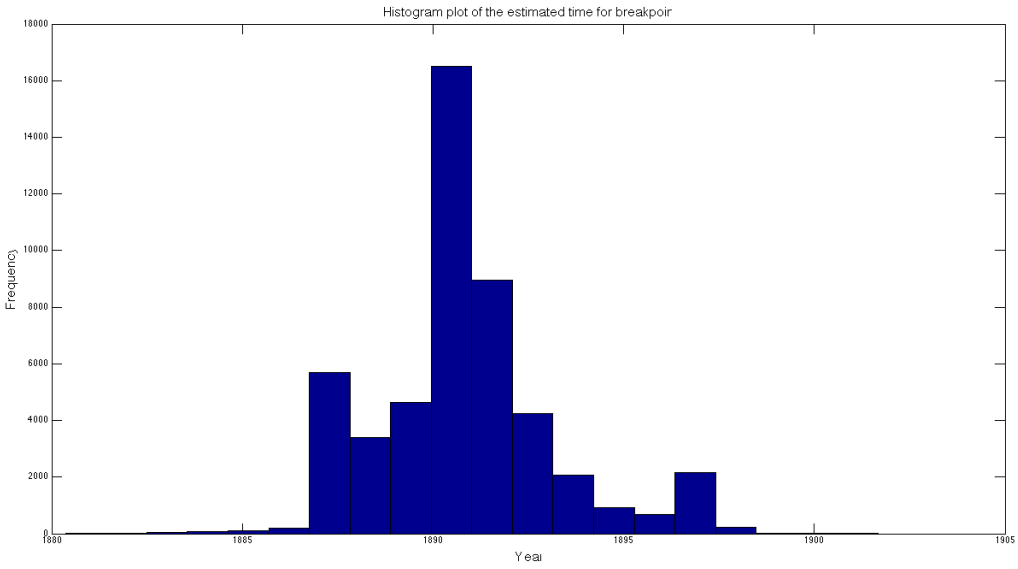


Figure 5: Histogram plot of the breakpoint.

As is seen in figure 5, the histogram is centered around 1891 with a very small tail indicating a high accuracy in the estimation of the breakpoint. We visualize the draws from the posterior in figure 6.

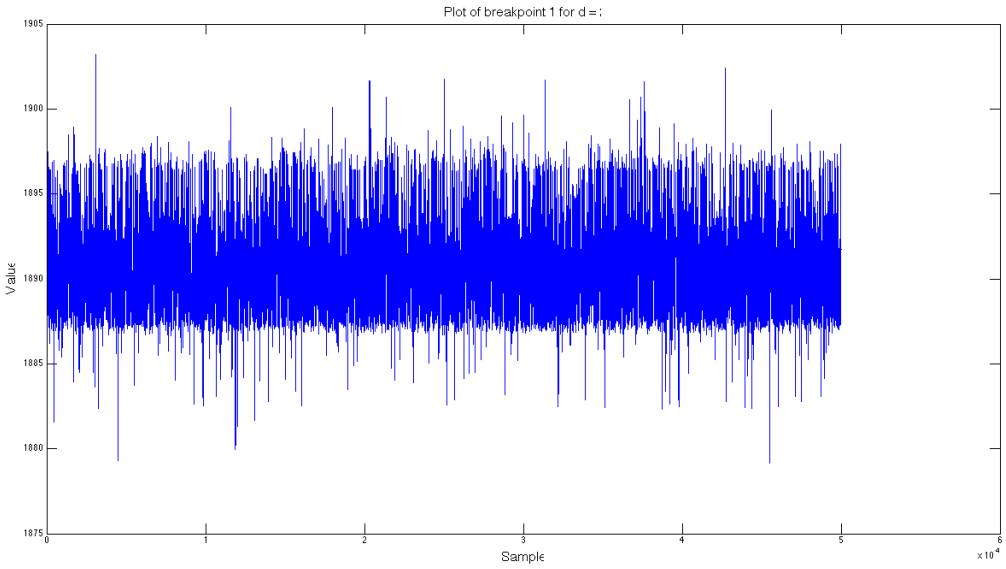


Figure 6: Visualization of the draws for the first breakpoint.

As is seen in the above figure the draws seem to be rather random and not depending on previous values. We move on to investigate the behaviour of the intensities and see figure 7.

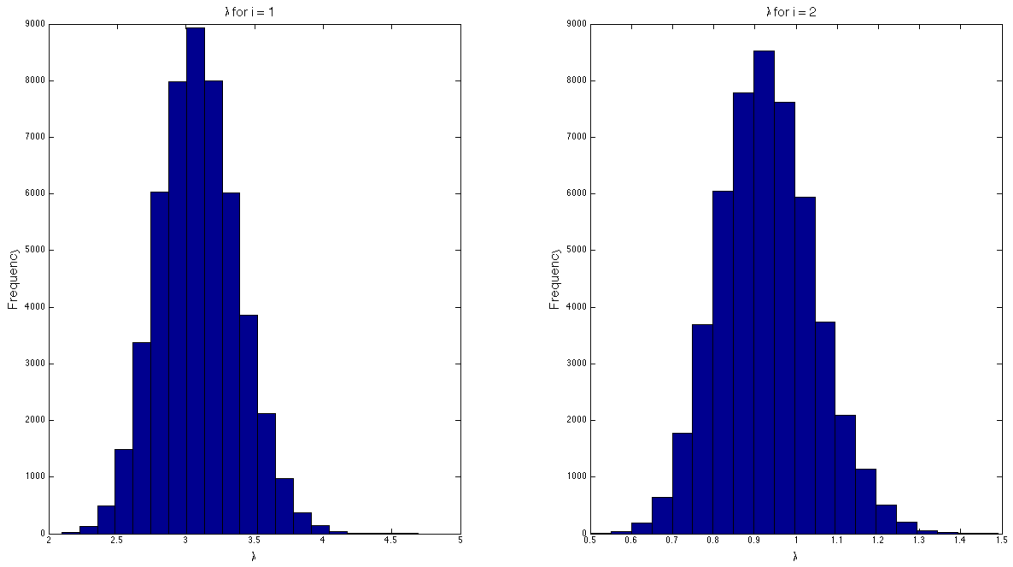


Figure 7: Histogram of the intensites of the two intervals.

The above figure suggests that the intensites both have a Γ–distribution where the first intensity is centered around ≈ 3.1 disasters/year whereas the second intensity is centered around ≈ 0.91 disasters/year, indicating a decrease of almost 70%.
Finally we look at the histogram of θ, which is seen in figure 8.

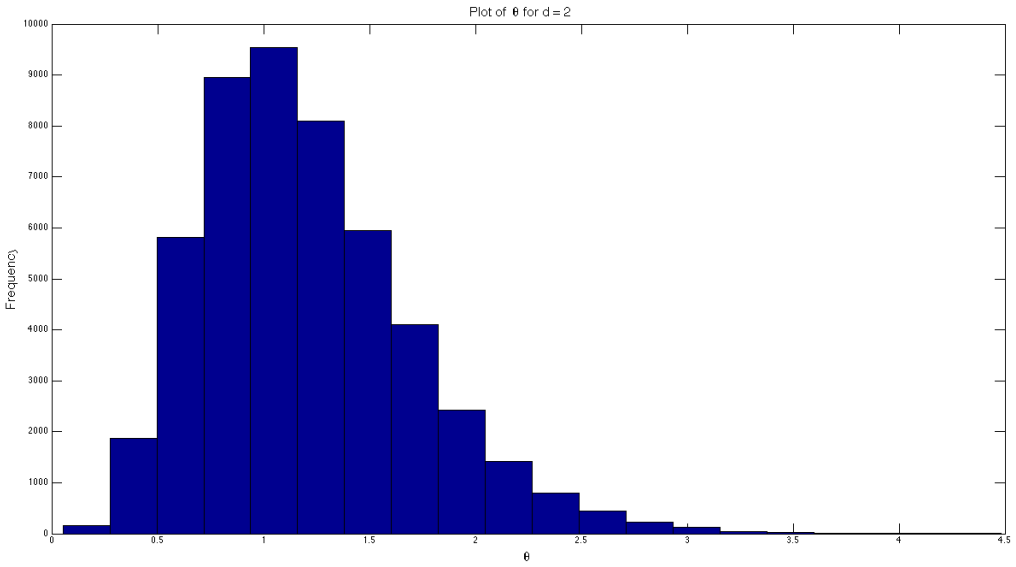


Figure 8: Histogram plot of θ.

Figure 8 is suggestive of a Γ–distribution centered around ≈ 1.2.

Two Breakpoints

We now move on to investigate the behaviour of the chain for two breakpoints. As before, we begin by looking at the location of the breakpoints, which is seen in figure 9.

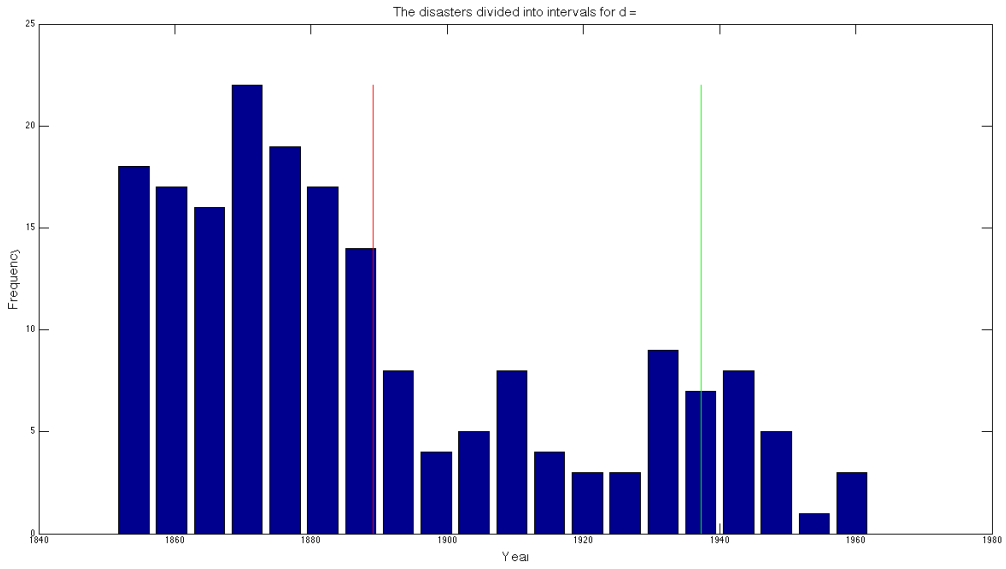


Figure 9: Plot of where the breakpoints are located.

As is seen in the above figure, the first breakpoint is almost located at the same time as before. This time it is ≈ 1890 and the second breakpoint is located at ≈ 1937. By looking at the figure, a more natural placement of the second breakpoint should be somewhere around 1930 since the change in intensity seems to occur there rather than at 1937. We move on to investigate the behaviour of the histogram of the breakpoints and look at figure 10.

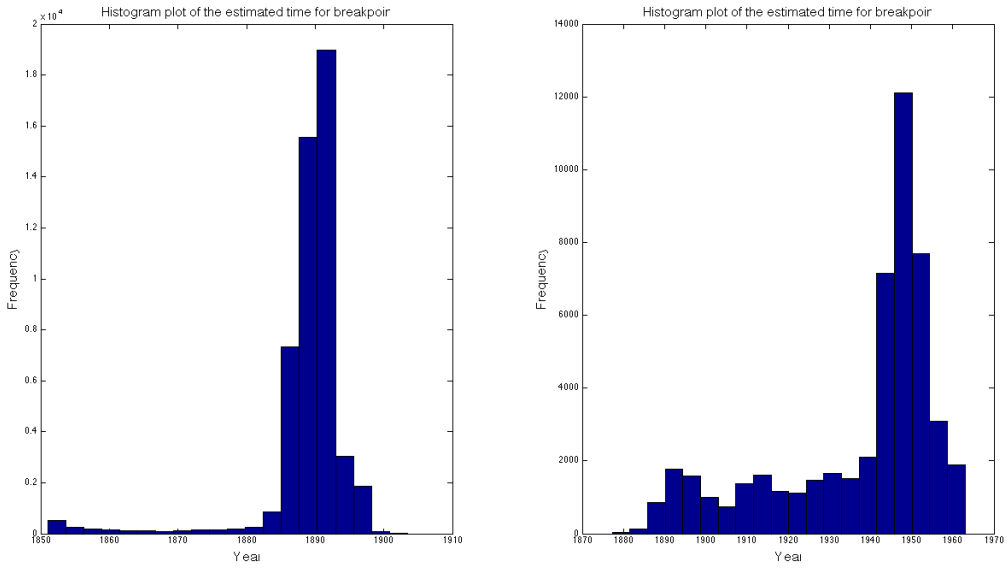


Figure 10: Histogram plot of the breakpoints.

As is seen in the above figure we have almost the same density as in figure 5 for the first breakpoint, indicating a high accuracy of the results. Whereas we for the second breakpoint have rather large tails spanning almost the entire time period. This result is rather disappointing since we can expect to have a bad estimation. The samples that are drawn from the posterior are represented in figure 11.

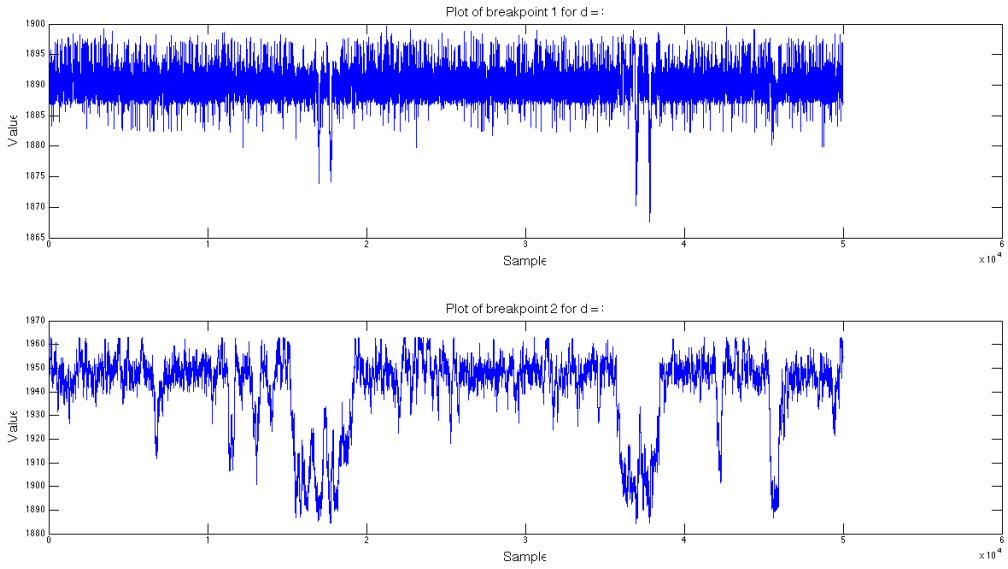


Figure 11: Visualization of the draws for both of the breakpoints.

As is seen in the figure above there is a high dependence of the draws which is very negative for our results.
We turn our eyes toward the histograms of the intensties in figure 12.

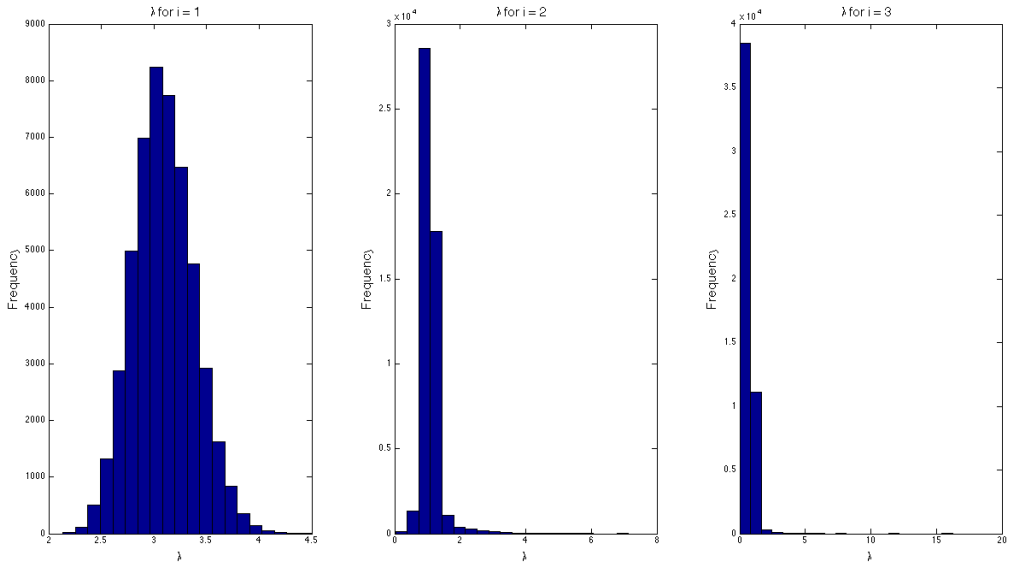


Figure 12: Histogram plot of the intensities.

By looking at the above figures we see that all of the intensities exhibit a Γ–distribution where they are centered around the values given by $\boldsymbol{\lambda} \approx (3.1, 1.2, 0.6)$. Since the first intensity and the first breakpoint are approximately the same as for one breakpoint we can assume that the "correct" breakpoint is somewhere between 1890–1891.
We finally look at the histogram of θ in figure 13.

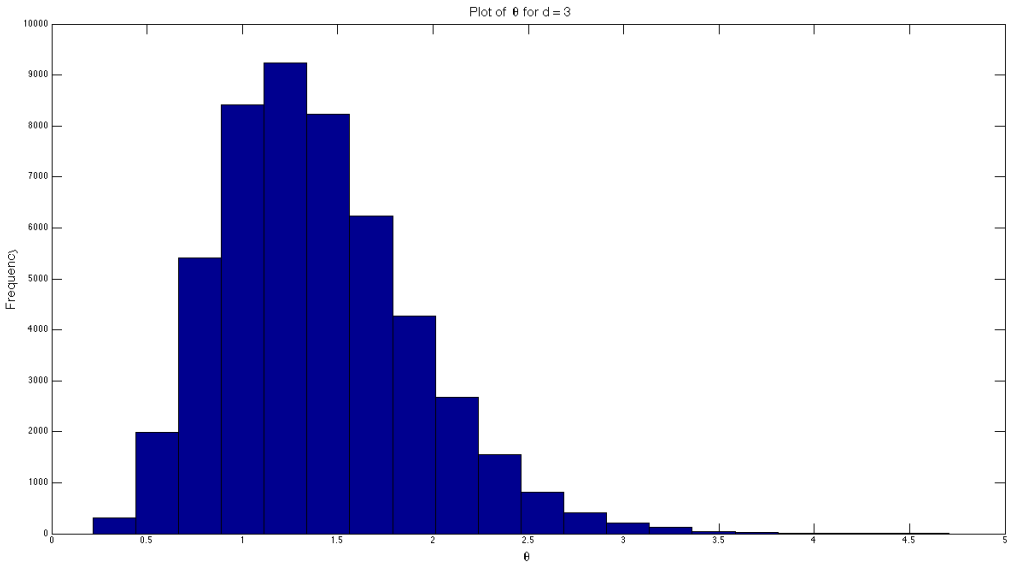


Figure 13: Histogram plot of θ.

The histogram is highly indicative of Γ–distribution centered around ≈ 1.38.

Three breakpoints

As in the previous sections we begin with investigating where the breakpoints are located, which is seen in figure 15.

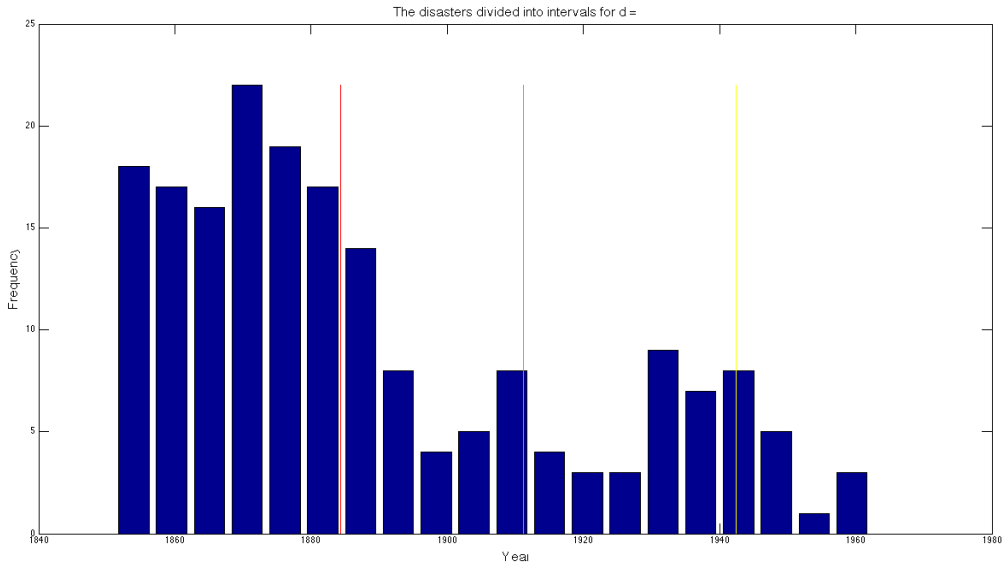


Figure 14: Plot of where the breakpoint is located.

From the above figure we see that the first breakpoint has moved to 1884 and the two other breakpoints are located around 1911 and 1942. It is difficult to comment on these locations since more than two breakpoints seem superflous. We thus look at the histogram plot of the breakpoints in figure **??**.

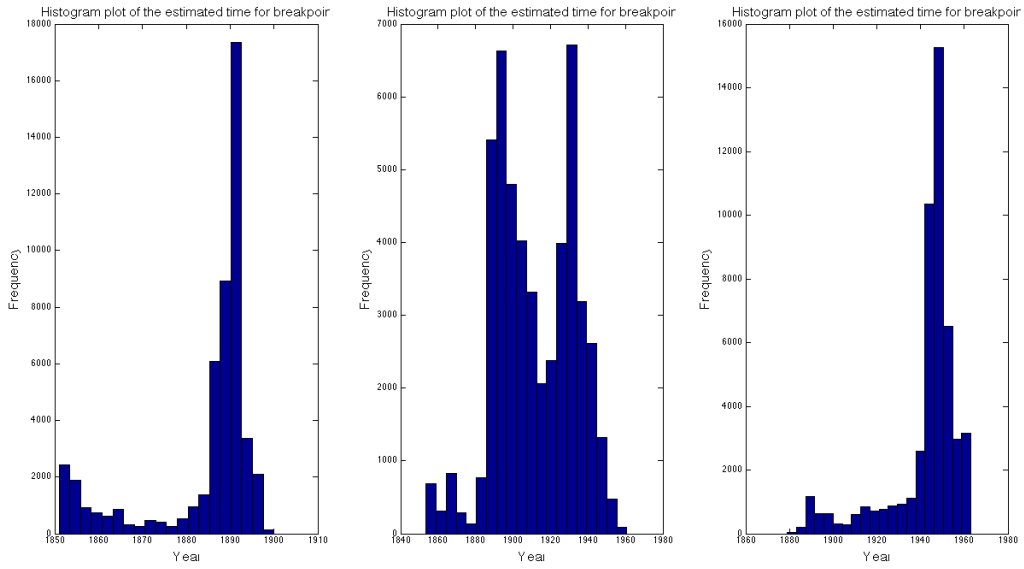


Figure 15: Histogram plot the breakpoints.

From the above figure we see that the histogram of the first breakpoint has a heavy tail to the left of the center suggesting a bad estimation. The histogram of the second breakpoint is really bad since we cannot distinguish a center of the density, which suggests that we will have a bad estimation in the breakpoint. The third histogram has almost the same density as that of the first density. However, it seems to be a little better since the tail isn't as heavy as for the first.
We now investigate the behaviour of the draws from this posterior, which are visualized in figure 16.

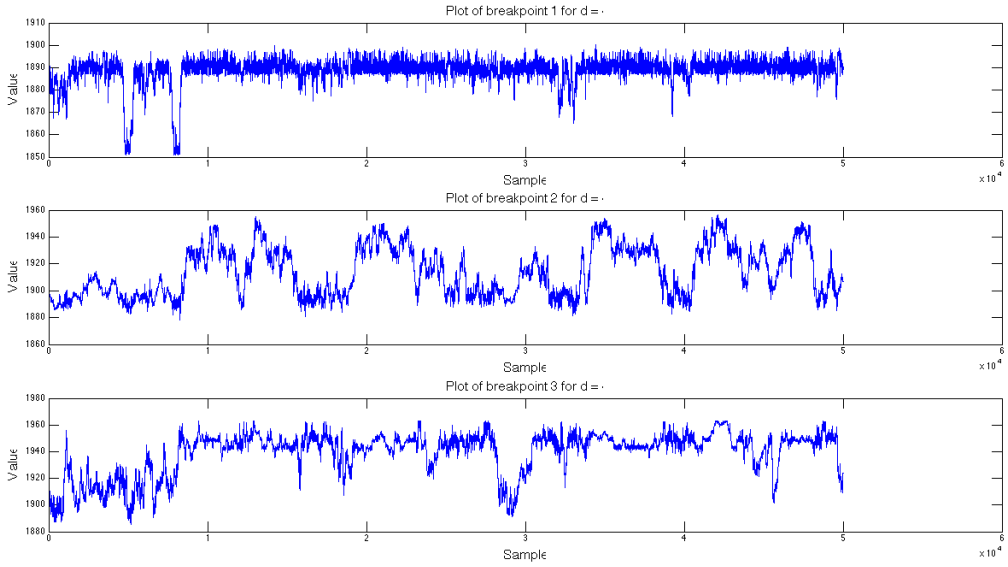


Figure 16: Visualization of the draws for both of the breakpoints.

As we can see in the above figure we have a high dependence between the different draws, especially for the second and third breakpoint.
We now move on to look at the histograms of the intensities.

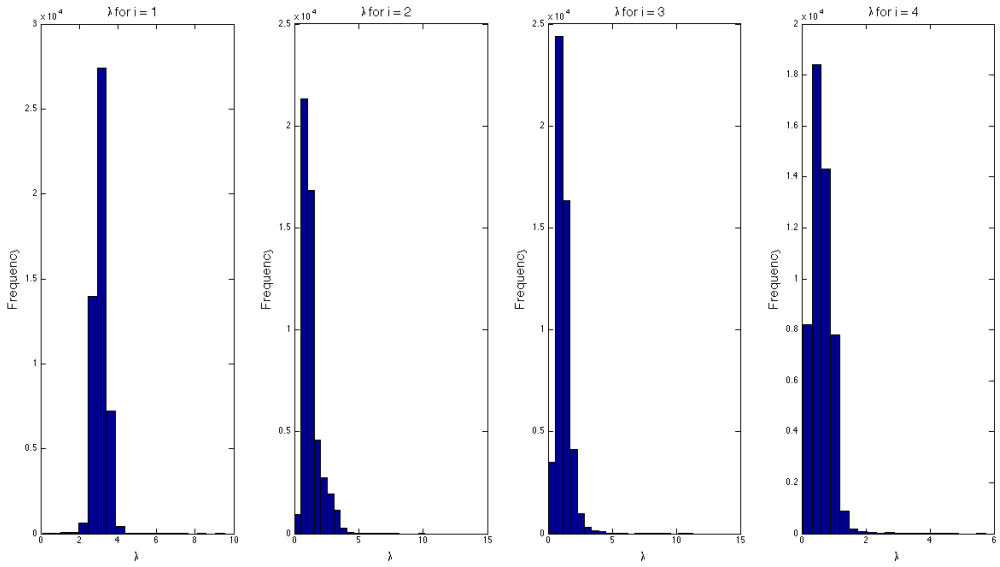


Figure 17: Histogram plots of the intensities.

From the above figure it is possible to discern a Γ–distribution for the distributions. The intensity for the first interval is still ≈ 3.1. Lastly, we look at the histogram plot of θ.

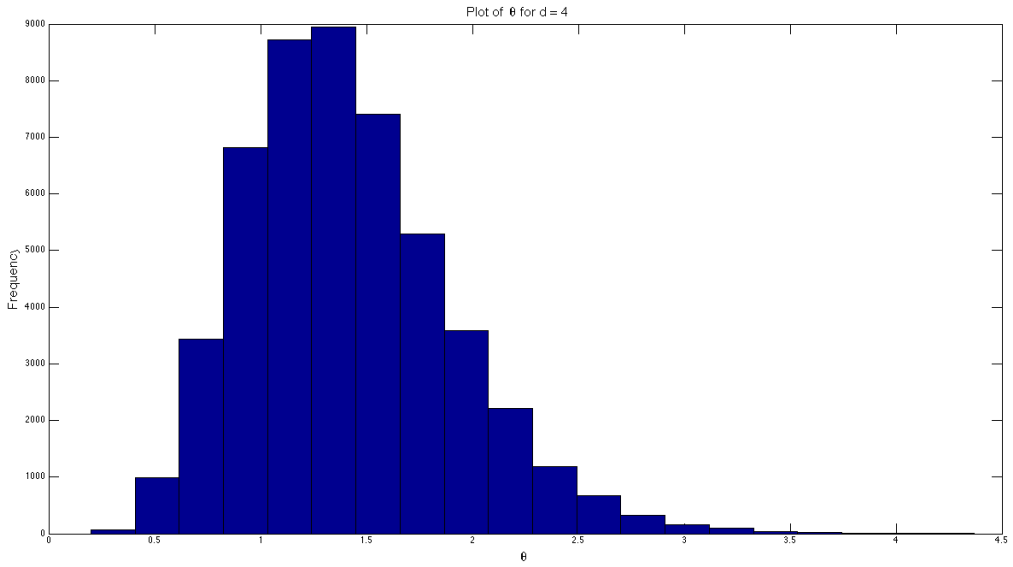


Figure 18: Histogram plot of θ.

Breakpoints	1	2	3	4
$\boldsymbol{t}$	1891	(1889, 1937)	(1884, 1911, 1942)	(1889, 1915, 1933, 1950)
$\boldsymbol{\lambda}$	(3.1, 0.9)	(3.1, 1.2, 0.6)	(3.1, 1.5, 1.3, 0.6)	(3.1, 1.3, 1.1, 1.2, 0.6)

These plots illustrate the expected breakpoints and therefore the associated intervals. As such we can evaluate if the interval data exhibit that of a gamma-distribution. We have illustrated the histograms in figures **??** of the $\boldsymbol{t}$-posteriors for each breakpoint, in order to investigate the probability of our breakpoint estimates. To also ensure that our prior θ has been established correctly, we here plot the histogram of θ indicating that it is of Γ-distributed character. The obtained θs and $\boldsymbol{\lambda}$ from our hyperprior is plotted in the figures **??** and **??** respectively

When sampling from the $f(\theta|\boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{t})$ the histogram plot of the different θs. The same distribution occured for each case, as seen in the plots above. Hence we have only plotted two cases.

If the figures show a strong convergence towards an estimated value, we can assume that our approximation is good. Whenever a shift occurs we have a new Po-distribution in the data. Looking at the left figure in figure 9 we can easily see that the breakpoints is centered around 1891 with a small tail, thus indicating that the approximation is valid. By also looking at figure 4 we can easily see that there is a change in intensity at 1891.

When the data has been divided with two breakpoints, as seen in the right figure of figure 10. It is seen that the first breakpoint basically has the same distribution as only using one breakpoint, i.e. small tails and centered around 1891. However, the second breakpoint is centered around 1945 where most of the mass is allocated around this value. It is however also noted that there is a heavy tail of the distribution spanning across 50 years.

The left figure of figure **??** shows the histograms of the three breakpoints. It is seen the the first breakpoint corresponds to almost the same previous values, i.e. centered around 1890, but one can note that there is a tail to the left. The second breakpoint is the most interesting, since there seems to be two centers, one in 1890 and one in 1940. Indicating an uncertainty in the estimation as the Monte Carlo approximation will be highly affected. The third breakpoint has the same distribution as the first and is centered around 1944.

When implementing four breakpoints in the data, we observe yet again that the first breakpoint is of good approximation, centered around 1890. The second breakpoint has almost the shape of a uniform distribution, meaning a high uncertainty. The third and fourth breakpoints have tails and no real center, also indicating an uncertainty.

Summary

From the previous analysis we can deduce that if we would divide the disasters into blocks with different intensities, we would use one breakpoint at the year 1891. Giving the intensities of the disasters as the figures show in 7. The intensity between the years 1851–1891 would be $\lambda_1 \approx 3.1$ disasters per year and the intensity between 1891–1963 would be $\lambda \approx 0.9$.

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 $T = 3 \cdot 14 \cdot 100$.

To 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 $\Delta_b^* = \hat{\theta}(Y_b^*) - \hat{\theta}, b \in \{1, 2, \ldots, B\}$.