STATS5084: Uncertainty Assessment and Bayesian Computation

Gabriel Ortega
2836741O@student.gla.ac.uk

Bayesian Changepoint Detection

Introduction

This project will explore a dataset of coal mining disasters in the United Kingdom recorded between the years 1851 and 1962. Several changes in the mining industry occurred during this period, most notably in 1880s when the rate of disasters decreased significantly due to several reforms and changes [1] [2] [3] within the coal mining industry. This is reflected in the data with a decreasing rate in coal mining disasters per year over the course of the observed period as seen in Figure 1.

The discussion of which series of events caused this decline in coal mining disasters is out of scope for this project. The focus will instead be put on the statistical analysis of the provided dataset [4] using methods pertaining to the domain of Bayesian statistics.

The goal of this analysis is to approximate the location(s) and possible number of changepoints within the data that best demonstrate the moments of change in the rates of coal mining disasters using just the data itself. The analysis will thus, not refer to any other sources of information regarding the coal mining industry within the observed period.

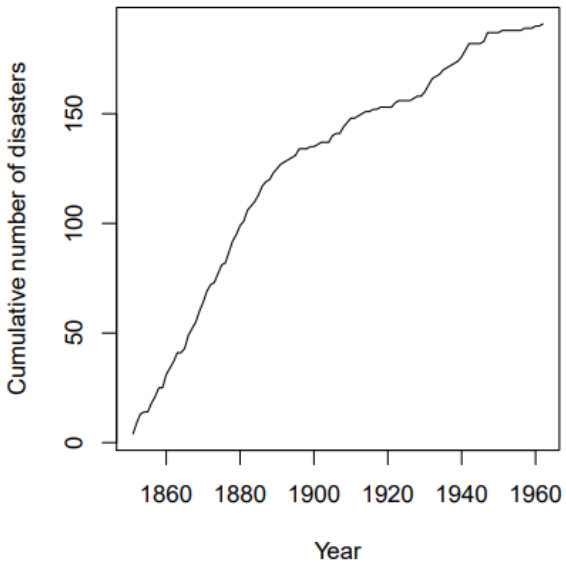


Figure 1: Cumulative number of UK coal mining disasters against time (in years).

This report will be split into 3 sections:

- The first section will look at the methodology used to infer the location(s) and number of changepoints using Bayesian statistics, generate models that put this method into practice and attempt to give a representation of the data, and finally, compare these generated models against one another to determine the optimal choice of changepoints and model parameters.
- The second section will cover the results obtained during the analysis and what can be inferred and explored further.
- The last section ends this report with a conclusion and a few closing thoughts regarding the outcome of the analysis and what could have been done to explore the dataset further if this project is to be re-visited.

Methodology

Problem specification

The observed count of annual coal mining disasters will be defined as following a Poisson distribution with an undefined rate **λ**:

$$y_i \sim Poi(\lambda), \qquad i = 1851, \dots, 1962, \qquad independently \tag{1}$$

and, in the case of one or multiple changepoints, the distribution (1) becomes:

$$\begin{cases} y_i \sim Poi(\lambda_1), & i = 1851, \dots, (T_1 - 1), \quad independently \\ y_i \sim Poi(\lambda_2), & i = T_1, \dots, (T_2 - 1), \quad independently \\ & \dots \\ y_i \sim Poi(\lambda_N), & i = T_{N-1}, \dots, (T_N - 1), \quad independently \\ y_i \sim Poi(\lambda_{N+1}), & i = T_N, \dots, 1962, \quad independently \end{cases} \tag{2}$$

where:

- N is equal to the total number of changepoints in the model.
- T_n is the n[th] changepoint of the model as follows:

$$1851, < T_1 < T_2 < \cdots < T_N \leq 1962$$

Finally, the priors are defined as follows:

$$\Pr(T_1, T_2, \dots, T_N) = \begin{cases} \binom{111}{N}^{-1}, & for\ 1851 < T_1 < T_2 < \cdots < T_N \leq 1962 \\ 0, & otherwise. \end{cases} \tag{3}$$

$$p(\lambda_i) = \text{Gamma}(2, 1), \qquad i = 1, \dots, N + 1, \qquad independently \tag{4}$$

Hereinafter, for the sake of simplicity, this report will refer to the values of T_n as being inside the interval $[1, \dots, 111]$ representing the 111 possible values for the changepoints.

Constructing the sampler

Given the discrete nature of the data, the use of a STAN implementation was out of question. This same scenario is presented in section 7.2 of the STAN users guide [5] and demonstrates the inefficiency of this approach. Metropolis-Hastings sampler can be of use here; however, the performance of this algorithm will diminish greatly as the number of changepoints increases as it creates models of higher dimensions which are difficult to estimate using this method.

Another option is to use a Gibbs sampler, like the Metropolis-Hastings sampler, it creates a Markov chain with the goal of converging towards a stationary distribution. The main difference here, is that the Gibbs sampler will not perform a complete random walk and instead only generate proposals that are acceptable.

To construct the Gibbs sampler, the full conditional distribution is used as the distribution to sample from. In the case of 1 changepoint, this is defined as follows:

Starting with the joint distribution based on the distributions of (2), (3), and (4):

$$p(y, T, \lambda_1, \lambda_2) = \left[\prod_{i=1}^{T} Poi(y_i|\lambda_1)\right] * \left[\prod_{j=T+1}^{112} Poi(y_j|\lambda_2)\right] * [Gamma(\lambda_1|2,1)] * [Gamma(\lambda_2|2,1)] * \binom{111}{1}^{-1}$$

$$= \left[\frac{\lambda_1^{\sum_i^T y_i} e^{-\lambda_1 * T}}{\prod_i^T y_i!}\right] * \left[\frac{\lambda_2^{\sum_j^{112} y_j} e^{-\lambda_2 * (112-T)}}{\prod_j^{112} y_j!}\right] * \left[\lambda_1 * e^{-\lambda_1}\right] * \left[\lambda_2 * e^{-\lambda_2}\right] * 111^{-1}$$

$$= \left[\frac{1}{\prod_i^T y_i! * \prod_j^{112} y_j!}\right] * \lambda_1^{1+\sum_i^T y_i} e^{-\lambda_1 * (1+T)} * \lambda_2^{1+\sum_j^{112} y_j} e^{-\lambda_2 * (112-T+1)} * 111^{-1} \quad (5)$$

Deriving the full conditional distributions from (5):

$$p(T|\lambda_1, \lambda_2, y) \propto \lambda_1^{\sum_i^T y_i} e^{-\lambda_1 * T} * \lambda_2^{\sum_j^{112} y_j} e^{-\lambda_2 * (112-T)}$$

$$p(\lambda_1|T, \lambda_2, y) \propto \lambda_1^{1+\sum_i^T y_i} e^{-\lambda_1 * (1+T)}$$
$$\propto Gamma\left(2 + \sum_{i=1}^T y_i, 1 + T\right)$$

$$p(\lambda_2|T, \lambda_1, y) \propto \lambda_2^{1+\sum_j^{112} y_j} e^{-\lambda_2 * (112-T+1)}$$
$$\propto Gamma\left(2 + \sum_{j=T+1}^{112} y_j, 1 + 112 - T\right)$$

This gives us the required distributions from which to sample from in our Gibbs sampler. However, considering that a sampler will have to be made for all number of changepoints, it would be better to see if it is possible to create a generic formula that can be used for the Gibbs samplers.

In the case of 2 changepoints, the joint distribution is:

$$p(y, T_1, T_2, \lambda_1, \lambda_2, \lambda_3) = \left[\prod_{i=1}^{T_1} Poi(y_i|\lambda_1)\right] * \left[\prod_{j=T_1+1}^{T_2} Poi(y_j|\lambda_2)\right] * \left[\prod_{k=T_2+1}^{112} Poi(y_k|\lambda_3)\right]$$
$$* [Gamma(\lambda_1|2,1)] * [Gamma(\lambda_2|2,1)] * [Gamma(\lambda_3|2,1)] * \binom{111}{2}^{-1}$$

$$= \left[\frac{\lambda_1^{\sum_i^{T_1} y_i} e^{-\lambda_1 * T_1}}{\prod_i^{T_1} y_i!}\right] * \left[\frac{\lambda_2^{\sum_j^{T_2} y_j} e^{-\lambda_2 * (T_2 - T_1)}}{\prod_j^{T_2} y_j!}\right] * \left[\frac{\lambda_3^{\sum_k^{112} y_k} e^{-\lambda_3 * (112-T_2)}}{\prod_k^{112} y_k!}\right]$$
$$* \left[\lambda_1 * e^{-\lambda_1}\right] * \left[\lambda_2 * e^{-\lambda_2}\right] * \left[\lambda_3 * e^{-\lambda_3}\right] * \binom{111}{2}^{-1}$$
$$= \left[\frac{1}{\prod_i^{T_1} y_i! * \prod_j^{T_2} y_j! * \prod_k^{112} y_k!}\right] * \lambda_1^{1+\sum_i^{T_1} y_i} e^{-\lambda_1 * (1+T_1)}$$

$$* \lambda_2^{1+\sum_j^{T_2} y_j} e^{-\lambda_2 * (T_2 - T_1 + 1)} * \lambda_3^{1+\sum_k^{112} y_k} e^{-\lambda_3 * (112-T_2+1)} * \binom{111}{2}^{-1} \quad (6)$$

Deriving the full conditional distributions from (6):

$$p(T_1|T_2, \lambda_1, \lambda_2, \lambda_3, y) \propto \lambda_1^{\sum_i^{T_1} y_i} e^{-\lambda_1 * T_1} * \lambda_2^{\sum_j^{T_2} y_j} e^{-\lambda_2 * (T_2 - T_1)}$$

$$p(T_2|T_1, \lambda_1, \lambda_2, \lambda_3, y) \propto \lambda_2^{\sum_j^{T_2} y_j} e^{-\lambda_2 * (T_2 - T_1)} * \lambda_3^{\sum_k^{112} y_k} e^{-\lambda_3 * (112-T_2)}$$

$$p(\lambda_1|T_1, T_2, \lambda_2, \lambda_3, y) \propto \lambda_1^{1+\sum_i^T y_i} e^{-\lambda_1 * (1+T)}$$
$$\propto Gamma\left(2 + \sum_{i=1}^{T_1} y_i, 1 + T_1\right)$$

$$p(\lambda_2|T_1, T_2, \lambda_1, \lambda_3, y) \propto \lambda_2^{1+\sum_j^{T_2} y_j} e^{-\lambda_2 * (1+T_2-T_1)}$$
$$\propto Gamma\left(2 + \sum_{j=T_1+1}^{T_2} y_j, 1 + T_2 - T_1\right)$$

$$p(\lambda_3|T_1,T_2,\lambda_1,\lambda_2,y) \propto \lambda_3^{1+\sum_k^{112} \mathrm{y_k}} e^{-\lambda_3*(1+112-T_2)}$$
$$\propto Gamma\left(2+\sum_{k=T_2+1}^{112} \mathrm{y_k}, 1+112-T_2\right)$$

Lastly, the joint distribution in the case of 3 changepoints:

$$p(y,T_1,T_2,T_3,\lambda_1,\lambda_2,\lambda_3,\lambda_4)$$

$$= \left[\prod_{i=1}^{T_1} Poi(\mathrm{y_i}|\lambda_1)\right] * \left[\prod_{j=T_1+1}^{T_2} Poi(\mathrm{y_j}|\lambda_2)\right] * \left[\prod_{k=T_2+1}^{T_3} Poi(\mathrm{y_k}|\lambda_3)\right] * \left[\prod_{k=T_3+1}^{112} Poi(\mathrm{y_l}|\lambda_4)\right]$$
$$* [Gamma(\lambda_1|2,1)] * [Gamma(\lambda_2|2,1)] * [Gamma(\lambda_3|2,1)] * [Gamma(\lambda_4|2,1)] * \binom{111}{3}^{-1}$$

$$= \left[\frac{\lambda_1^{\sum_i^{T_1} \mathrm{y_i}} e^{-\lambda_1*T_1}}{\prod_i^{T_1} \mathrm{y_i}!}\right] * \left[\frac{\lambda_2^{\sum_j^{T_2} \mathrm{y_j}} e^{-\lambda_2*(T_2-T_1)}}{\prod_j^{T_2} \mathrm{y_j}!}\right] * \left[\frac{\lambda_3^{\sum_k^{T_3} \mathrm{y_k}} e^{-\lambda_3*(T_3-T_2)}}{\prod_k^{T_3} \mathrm{y_k}!}\right] * \left[\frac{\lambda_4^{\sum_l^{112} \mathrm{y_l}} e^{-\lambda_4*(112-T_3)}}{\prod_l^{112} \mathrm{y_l}!}\right]$$
$$* \left[\lambda_1 * e^{-\lambda_1}\right] * \left[\lambda_2 * e^{-\lambda_2}\right] * \left[\lambda_3 * e^{-\lambda_3}\right] * \left[\lambda_4 * e^{-\lambda_4}\right] * \binom{111}{3}^{-1}$$
$$= \left[\frac{1}{\prod_i^{T_1} \mathrm{y_i}! * \prod_j^{T_2} \mathrm{y_j}! * \prod_k^{T_3} \mathrm{y_k}! * \prod_l^{112} \mathrm{y_l}!}\right] * \lambda_1^{1+\sum_i^{T_1} \mathrm{y_i}} e^{-\lambda_1*(1+T_1)} * \lambda_2^{1+\sum_j^{T_2} \mathrm{y_j}} e^{-\lambda_2*(T_2-T_1+1)}$$
$$* \lambda_3^{1+\sum_k^{T_3} \mathrm{y_k}} e^{-\lambda_3*(T_3-T_2+1)} * \lambda_4^{1+\sum_l^{112} \mathrm{y_l}} e^{-\lambda_4*(112-T_3+1)} * \binom{111}{3}^{-1} \quad (7)$$

Deriving the full conditional distributions from (7):

$$p(T_1|T_2,T_3,\lambda_1,\lambda_2,\lambda_3,\lambda_4,y) \propto \lambda_1^{\sum_i^{T_1} \mathrm{y_i}} e^{-\lambda_1*T_1} * \lambda_2^{\sum_j^{T_2} \mathrm{y_j}} e^{-\lambda_2*(T_2-T_1)}$$

$$p(T_2|T_1,T_3,\lambda_1,\lambda_2,\lambda_3,\lambda_4,y) \propto \lambda_2^{\sum_j^{T_2} \mathrm{y_j}} e^{-\lambda_2*(T_2-T_1)} * \lambda_3^{\sum_k^{T_3} \mathrm{y_k}} e^{-\lambda_3*(T_3-T_2)}$$

$$p(T_3|T_1,T_2,\lambda_1,\lambda_2,\lambda_3,\lambda_4,y) \propto \lambda_3^{\sum_k^{T_3} \mathrm{y_k}} e^{-\lambda_3*(T_3-T_2)} * \lambda_4^{\sum_l^{112} \mathrm{y_l}} e^{-\lambda_4*(112-T_3)}$$

$$p(\lambda_1|T_1,T_2,T_3,\lambda_2,\lambda_3,\lambda_4,y) \propto \lambda_1^{1+\sum_i^{T} \mathrm{y_i}} e^{-\lambda_1*(1+T)}$$
$$\propto Gamma\left(2+\sum_{i=1}^{T_1} \mathrm{y_i}, 1+T_1\right)$$

$$p(\lambda_2|T_1,T_2,T_3,\lambda_1,\lambda_3,\lambda_4,y) \propto \lambda_2^{1+\sum_j^{T_2} \mathrm{y_j}} e^{-\lambda_2*(1+T_2-T_1)}$$
$$\propto Gamma\left(2+\sum_{j=T_1+1}^{T_2} \mathrm{y_j}, 1+T_2-T_1\right)$$

$$p(\lambda_3|T_1,T_2,T_3,\lambda_1,\lambda_2,\lambda_4,y) \propto \lambda_3^{1+\sum_k^{T_3} \mathrm{y_k}} e^{-\lambda_3*(1+T_3-T_2)}$$
$$\propto Gamma\left(2+\sum_{k=T_2+1}^{T_3} \mathrm{y_k}, 1+T_3-T_2\right)$$

$$p(\lambda_3|T_1,T_2,T_3,\lambda_1,\lambda_2,\lambda_3,y) \propto \lambda_4^{1+\sum_l^{112} \mathrm{y_l}} e^{-\lambda_4*(1+112-T_3)}$$
$$\propto Gamma\left(2+\sum_{l=T_3+1}^{112} \mathrm{y_l}, 1+112-T_3\right)$$

From these distributions, a generic formula can be obtained for the full conditional distribution of T_n:

$$p(T_n|T_1,\dots,T_{n-1},T_{n+1},\dots,T_N,\lambda_1,\dots,\lambda_{\mathrm{N}+1},y), \qquad n=1,\dots,N$$
$$\propto \lambda_{\mathrm{n}}^{\sum_{i=T_{n-1}+1}^{T_n}\mathrm{y_i}} e^{-\lambda_{\mathrm{n}}*(T_n-T_{n-1})} * \lambda_{\mathrm{n+1}}^{\sum_{j=T_n+1}^{T_{n+1}}\mathrm{y_j}} e^{-\lambda_{\mathrm{n+1}}*(T_{n+1}-T_n)} \tag{8}$$

and the generic formula for the full conditional distribution of λ_{n}:

$$p(\lambda_{\mathrm{n}}|T_1,\dots,T_N,\lambda_1,\dots,\lambda_{\mathrm{n-1}},\lambda_{\mathrm{n+1}},\dots,\lambda_{\mathrm{N}+1},y), \qquad \mathrm{n}=1,\dots,\mathrm{N}+1$$
$$\propto \lambda_{\mathrm{n}}^{1+\sum_{i=T_{n-1}+1}^{T_n}\mathrm{y_i}} e^{-\lambda_{\mathrm{n}}*(1+T_n-T_{n-1})} \propto Gamma\left(2+\sum_{i=T_{n-1}+1}^{T_n}\mathrm{y_i}, 1+T_n-T_{n-1}\right) \tag{9}$$

In the context of these generic formulas, the extended lower and upper bounds are constant for all iterations and defined as:

$$\mathrm{T}_0 = 0$$
$$\mathrm{T}_{N+1} = 112$$

which allows the Gibbs sampler implementation to treat the sampling of the new values of T_n and λ_{n} as a series of sliding window operations applied on the extended changepoint vector.

Moreover, the sampling for T_n is bounded by $(T_{n-1},\dots,T_{n+1})$, where T_n can take any value in this interval excluding the bounds themselves. As such, the implementation does not need to consider the full range of 111 possible values and instead, only apply the generic formula to this interval. If the calculated interval is of size 1 (i.e.: there is only 1 possible changepoint value), then this value is immediately returned as the result of the formula. This small optimization greatly reduces the computational cost of sampling for T_n.

N.B.: Despite this extension of the changepoint vector, the values of T_n used in the final model will never be equal to neither T_0 nor T_{N+1}. Also, when sampling for T_n, the densities obtained are first normalized.

The last step for the Gibbs sampler is to apply formulas (8) and (9). The order of operations should be to first sample all T_n to obtain all the changepoints of the current sample, followed by sampling the values of all λ_{n}. At first glance, this means that a Gibbs sampler iteration is composed of 2 loops, however, notice that for sampling T_n, only λ_{n} and $\lambda_{\mathrm{n+1}}$ are required, and for sampling λ_{n}, only T_{n-1} and T_n are required. Since the first loop will sample all T_n, this means that the λ_{n}, $\lambda_{\mathrm{n+1}}$, values required for these operations will always be taken from the sample of the **previous** iteration and the second loop sampling all λ_{n} will take the T_n, T_{n-1} values from the sample of the **current** iteration.

Thus, the 2 loops can be combined into a single loop if the order execution of these operations is changed to:

$$T_1,\lambda_1,T_2,\lambda_2,\dots,T_n,\lambda_{\mathrm{n}},\dots,T_N,\lambda_{\mathrm{N}},\lambda_{\mathrm{N+1}}$$

with a final operation for sampling $\lambda_{\mathrm{N+1}}$ outside of this loop, which leads to an increase of performance per sampler iteration.

The final implementation of the Gibbs sampler algorithm can be summarized as follows:

Initialize the chain as $\mathrm{T}_1^{(0)}, \dots, \mathrm{T}_N^{(0)}, \lambda_1^{(0)}, \dots, \lambda_{\mathrm{N}+1}^{(0)}$
for $\mathrm{i} = 1, \dots, max\_iterations$ **do**
  for $\mathrm{n} = 1, \dots, \mathrm{N}$ **do**
    $T_n^{(i)} \sim p\left(T_n | \lambda_1^{(i)}, \dots, \lambda_{\mathrm{n}-1}^{(i)}, \lambda_{\mathrm{n}}^{(i-1)}, \dots, \lambda_{\mathrm{N}+1}^{(i-1)}, T_1^{(i)}, \dots, T_{n-1}^{(i)}, T_{n+1}^{(i-1)}, \dots, T_N^{(i-1)}, y\right)$
    $\lambda_{\mathrm{n}}^{(i)} \sim p\left(\lambda_{\mathrm{n}} | \lambda_1^{(i)}, \dots, \lambda_{\mathrm{n}-1}^{(i)}, \lambda_{\mathrm{n}+1}^{(i-1)}, \dots, \lambda_{\mathrm{N}+1}^{(i-1)}, T_1^{(i)}, \dots, T_n^{(i)}, T_{n+1}^{(i-1)}, \dots, T_N^{(i-1)}, y\right)$
  end
  $\lambda_{\mathrm{N}+1}^{(i)} \sim p\left(\lambda_{\mathrm{N}+1} | \lambda_1^{(i)}, \dots, \lambda_{\mathrm{N}}^{(i)}, T_1^{(i)}, \dots, T_N^{(i)}, y\right)$
end

Calculating the (log) marginal likelihood

One can intuitively assume that the more changepoints in our model, the closer the model's predictions will be to the observed data. However, as per Occam's razor principle, one should instead prefer simpler models even if a bit less precise as more complex models can be more difficult to build, process, and understand.

To compare models with one another, the marginal likelihood will be used as it represents the probability of generating the observed sample for all possible values of the parameters and will serve as a metric to evaluate model performance with a built-in penalty the more complex the model is.

To obtain the marginal likelihood per model:

$$p(M_N|y) = \sum_{T_1=1952}^{1963-N} \sum_{T_2=T_1+1}^{1964-N} \dots \sum_{T_N=T_{N-1}+1}^{1962} \iint_0^\infty \dots$$
$$\dots \int_0^\infty p(y|T_1, \dots, T_N, \lambda_1, \dots, \lambda_{\mathrm{N}+1})\, p(T_1, \dots, T_N, \lambda_1, \dots, \lambda_{\mathrm{N}+1}), d\lambda_{\mathrm{N}+1}, \dots, d\lambda_1$$

which can be simplified to:

$$p(M_N|y) = \sum_{T_1=1952}^{1963-N} \sum_{T_2=T_1+1}^{1964-N} \dots \sum_{T_N=T_{N-1}+1}^{1962} \binom{111}{N}^{-1} \frac{\prod_{i=1}^{N+1}(L_i+1)^{-S_i-2}\Gamma(2+S_i)}{\prod_{i=1851}^{1962} \mathrm{y_i}!} \qquad (10)$$

where L_i is the length (as in, the number of years) in segment i, and S_i is the corresponding sum of the disaster numbers within this segment.

During testing, this calculation's computational cost would skyrocket after 5 changepoints, at which point it was no longer viable to use it to obtain the marginal likelihoods on time.

Due to time constraints, an approximative approach had to be considered. This ended up being the Chib method [6] as it could work in tandem with the Gibbs sampler to produce well approximated results.

The Chib method defines the marginal likelihood of a model as follows:

$$p(y|M_N) = \frac{f(y|\theta_N^*, M_N)\pi(\theta_N^*|M_N)}{\pi(\theta_N^*|y, M_N)}$$

where θ_N^* is the parameter vector for model M_N obtained by calculating the **maximum likelihood estimate** of all the samples generated by the Gibbs sampler using model M_N, $f(y|\theta_N^*, M_N)$ is the density function of the

data under model M_N, and $\pi(\theta_N^*|M_N)$, and $\pi(\theta_N^*|y, M_N)$ is the posterior density for model M_N obtained via a Gibbs sampler.

From the output obtained from the Gibbs sampler, one can easily get $f(y|\theta_N^*, M_N)$ and $\pi(\theta_N^*|M_N)$:

$$f(y|\theta_N^*, M_N) = \left[\prod_{i=1}^{T_1^*} Poi(\mathrm{y_i}|\lambda_1^*)\right] * \dots * \left[\prod_{j=T_N^*+1}^{112} Poi(\mathrm{y_j}|\lambda_{\mathrm{N}+1}^*)\right]$$

$$\pi(\theta_N^*|M_N) = [Gamma(\lambda_1^*|2,1)] * \dots * [Gamma(\lambda_{\mathrm{N}+1}^*|2,1)] * \binom{111}{N}^{-1}$$

The process to obtain the posterior density $\pi(\theta_N^*|y, M_N)$ is more involved however, as it requires one to re-run the Gibbs sampler with **reduced conditional distributions** (i.e.: setting one or more of the model parameters in the Gibbs sampler as constants that are never updated).

The order in which the reduced conditional distributions are processed is of great importance and, through extensive experimentation, has been found to be best when it matches the same order as the full conditional distributions in the Gibbs sampler.

Lastly, the posterior of each parameter must be averaged over all the samples obtained in the Gibbs sampler re-runs (hereinafter in this report, these are called **reduced Gibbs samplers**). This technique is called Rao-Blackwellization [7].

Thus, the reduced conditional distributions are processed as follows:

$$\pi(\theta_N^*|y, M_N) =$$

$$G^{-1}\sum_{g=1}^{G} \pi\left(T_1^*|\ \lambda_1^{\mathrm{g}}, T_2^g, \lambda_2^{\mathrm{g}}, \dots, \mathrm{T_N^g}, \lambda_{\mathrm{N}}^{\mathrm{g}}, \lambda_{\mathrm{N}+1}^{\mathrm{g}}, y\right) * G^{-1}\sum_{g=1}^{G} \pi\left(\lambda_1^*|\ \mathrm{T_1^*}, T_2^g, \lambda_2^{\mathrm{g}}, \dots, \mathrm{T_N^g}, \lambda_{\mathrm{N}}^{\mathrm{g}}, \lambda_{\mathrm{N}+1}^{\mathrm{g}}, y\right)$$
$$* G^{-1}\sum_{g=1}^{G} \pi\left(T_2^*|\ \mathrm{T_1^*}, \lambda_1^*, \lambda_2^g, \dots, \mathrm{T_N^g}, \lambda_{\mathrm{N}}^{\mathrm{g}}, \lambda_{\mathrm{N}+1}^{\mathrm{g}}, y\right) * G^{-1}\sum_{g=1}^{G} \pi\left(\lambda_2^*|\ \mathrm{T_1^*}, \lambda_1^*, T_2^*, \dots, \mathrm{T_N^g}, \lambda_{\mathrm{N}}^{\mathrm{g}}, \lambda_{\mathrm{N}+1}^{\mathrm{g}}, y\right)$$
$$* \dots$$
$$* G^{-1}\sum_{g=1}^{G} \pi\left(T_N^*|\ \mathrm{T_1^*}, \lambda_1^*, T_2^*, \lambda_2^*, \dots, \lambda_{\mathrm{N}}^{\mathrm{g}}, \lambda_{\mathrm{N}+1}^{\mathrm{g}}, y\right) * G^{-1}\sum_{g=1}^{G} \pi\left(\lambda_N^*|\ \mathrm{T_1^*}, \lambda_1^*, T_2^*, \lambda_2^*, \dots, \mathrm{T_N^*}, \lambda_{\mathrm{N}+1}^{\mathrm{g}}, y\right)$$
$$* G^{-1}\sum_{g=1}^{G} \pi(\lambda_{N+1}^*|\ \mathrm{T_1^*}, \lambda_1^*, T_2^*, \lambda_2^*, \dots, \mathrm{T_N^*}, \lambda_{\mathrm{N}}^*, y)$$

where G is equal to the number of samples (after burn-in) obtained from the reduced Gibbs sampler runs.

N.B.: For $\pi\left(T_1^*|\ \lambda_1^{\mathrm{g}}, T_2^g, \lambda_2^{\mathrm{g}}, \dots, \mathrm{T_N^g}, \lambda_{\mathrm{N}}^{\mathrm{g}}, \lambda_{\mathrm{N}+1}^{\mathrm{g}}, y\right)$, the output of the full Gibbs sampler run is used as this corresponds to the density of T_1^* with the full conditional distributions for all parameters.

The formulas used for computing $\pi(T_N^*|\ \dots, y)$ and $\pi(\lambda_N^*|\ \dots, y)$, $\pi(\lambda_{N+1}^*|\ \dots, y)$ will be the same as the generic formulas (8), (9) used for the Gibbs sampler and the densities of T_N^* will also be normalized.

Finally, to simplify the computation and make model comparison easier, the log marginal likelihood will be used:

$$\ln\left(p(y|M_N)\right) = \ln\big(f(y|\theta_N^*, M_N)\big) + \ln\big(\pi(\theta_N^*|M_N)\big) - \ln\left(\pi(\theta_N^*|y, M_N)\right)$$

Results

This section will be organized as follows:

- Present and analyze the results obtained with Gibbs sampler for simple models.
- Present the marginal likelihoods obtained using both methods described for simple models.
- Explore the marginal likelihoods obtained using the Chib method for more complex models.

Gibbs sampler output

The focus being on evaluating the performance of the Gibbs sampler, each model will be presented with a 95% confidence interval plotted over the original data, and the trace and density plots showing the values that the parameters took over the course of the entire sampler run (after burn-in iterations). The number of burn-in iterations is 1000 and the number of usable sample iterations is 9000 for all Gibb sampler runs presented hereinafter.

The **M**aximum **A P**osteriori parameters presented are obtained by calculating the mode of the different T_n posterior distributions.

Considering the size of the plots showing the accuracy and convergence of the different models, only the 1, 4, and 10 changepoint models will be shown on this report.

1 changepoint

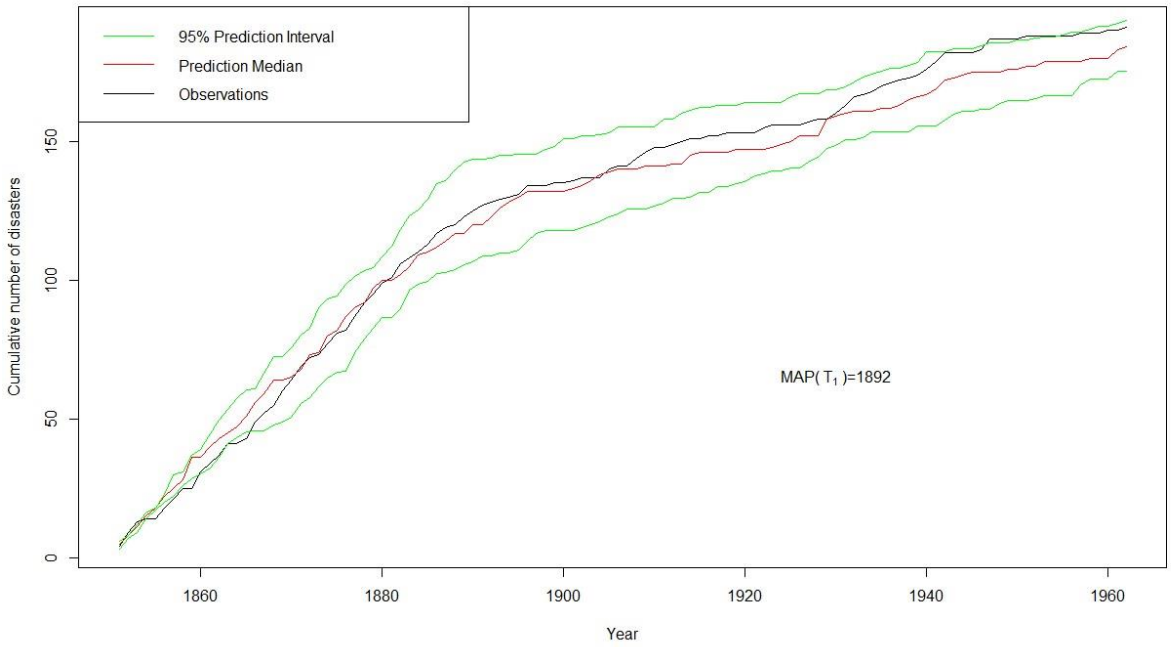


Figure 2: Posterior predictive distributions for a model with 1 changepoint

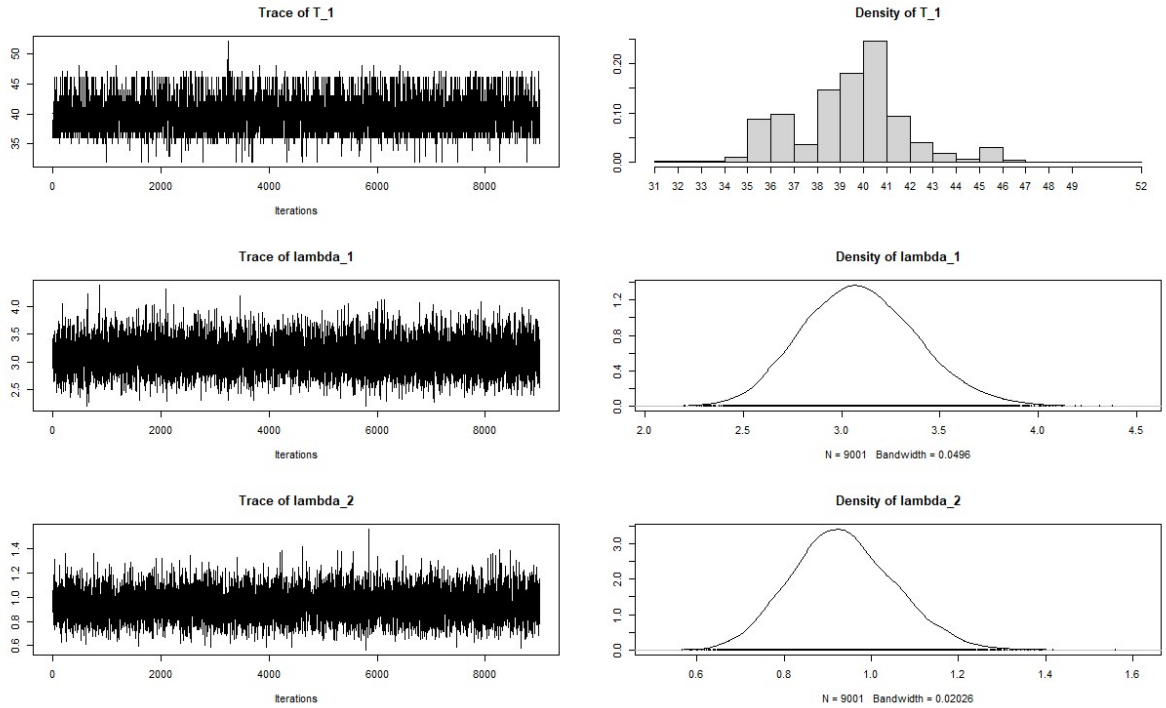


Figure 3: Trace and density plots for a model with 1 changepoint

The model for 1 changepoint is relatively simple. With a MAP at roughly 1892 (or T_1=41 using the scale defined in the previous sections), the posterior density of T_1, shown in Figure 3, looks very similar to a normal distribution of mean 41 and stand deviation of roughly 3.

The chain shown in Figure 2: Posterior predictive distributions for a model with 1 changepoint performed rather well. The median being quite close to the observed data and the 95% confidence interval staying quite close as well. The model prediction tends to underperform at roughly 1930 which may be indicative of another potential changepoint.

The lambdas' posterior densities also follow a normal distribution, and the trace plots show the same convergence tendency as for T_1.

4 changepoints

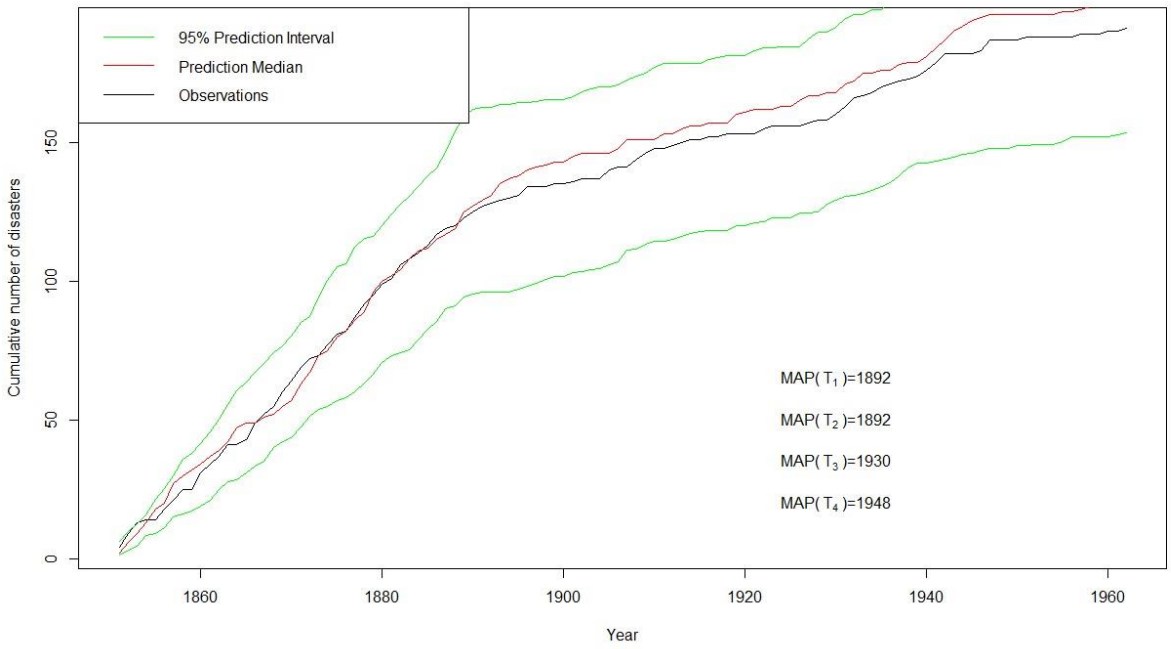


Figure 4: Posterior predictive distributions for a model with 4 changepoints

The model for 4 changepoints starts to already show some interesting points of convergence with the values of parameters T_2 and T_2 having the same MAP yet quite different trace plots. This can likely be explained by the strong correlation between the different values of T_n.

Say for example that T_1 was sampled at a value between 0-20, this would very likely increase the probabilities that T_2 will be sampled around the value 40, which was the MAP of the T parameter in the 1 changepoint model. In Figure 4, what is also interesting to see is the presence of another potential changepoint convergence point at a value of around 80 as reflected by the posterior densities of both T_2 and T_3. It is highly likely that if the value T_1 was low enough, T_2 would favor a value around 40 and T_3 a value around 80, whereas if T_1 was at around 40, then T_2 would favor a value around 80 and T_3 would tend toward the higher end at values of around 90. T_4, as a result, was likely always confined to the range between values 90 and 111, favoring the middle of the range as its MAP was estimated at 1948 or a T value of 97.

Unsurprisingly, the lambdas tend to follow a similar multi-modal posterior distribution to that of the T parameters (this is much more visible in the case of λ_2). This is to be expected as the full conditional distribution of λ_n is also correlated to the T parameters.

The plot in Figure 4 shows a median that is much closer to the observed data than in the case of the one changepoint model, however, the 95% confidence interval is much wider, likely indicating more variance in the different samples obtained. This could also be due to a bad initial starting point for the chain which required many more iterations than in the previous case. Perhaps more burn-in iterations would've helped here.

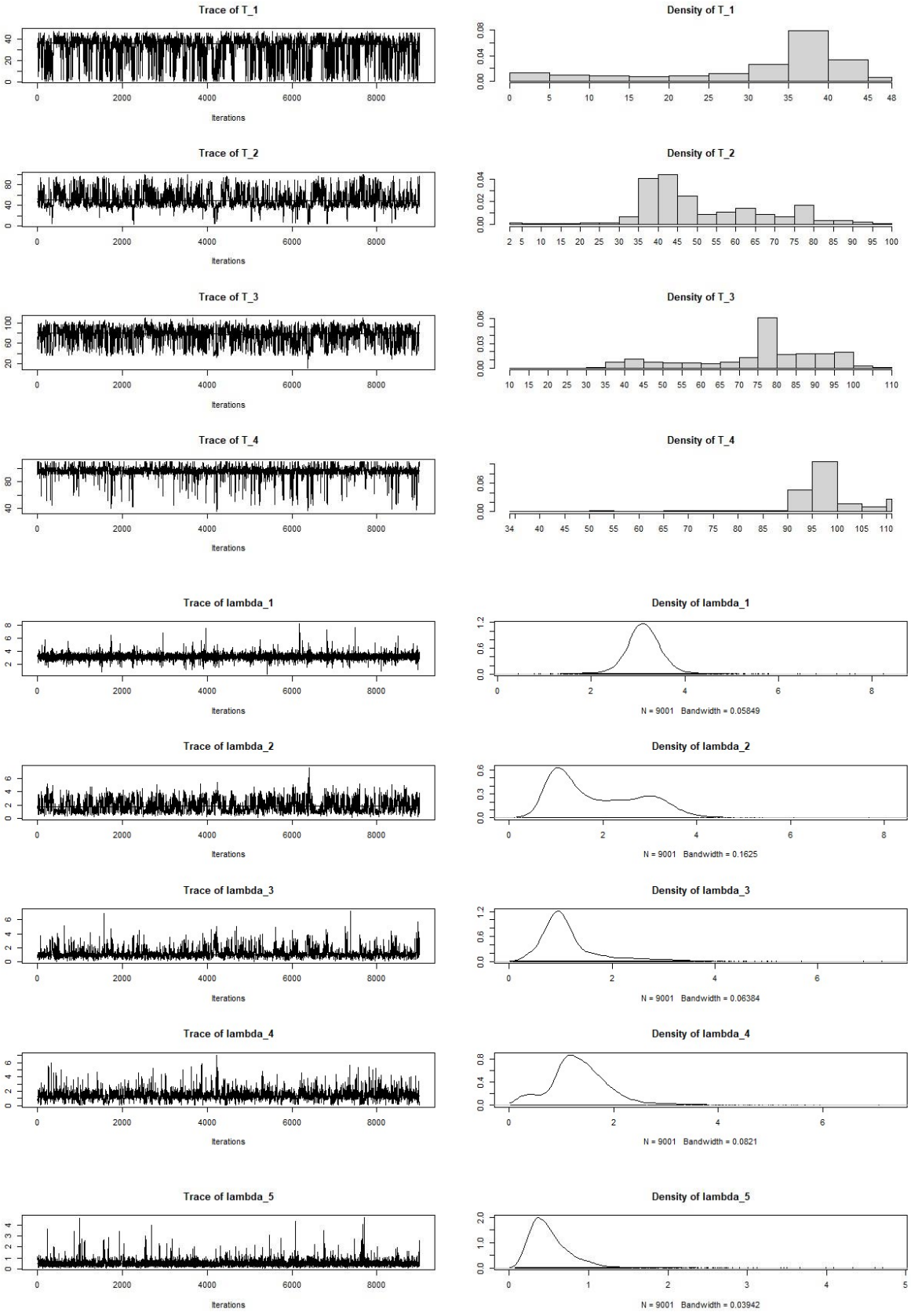


Figure 5: Trace and density plots for a model with 4 changepoints

10 changepoints

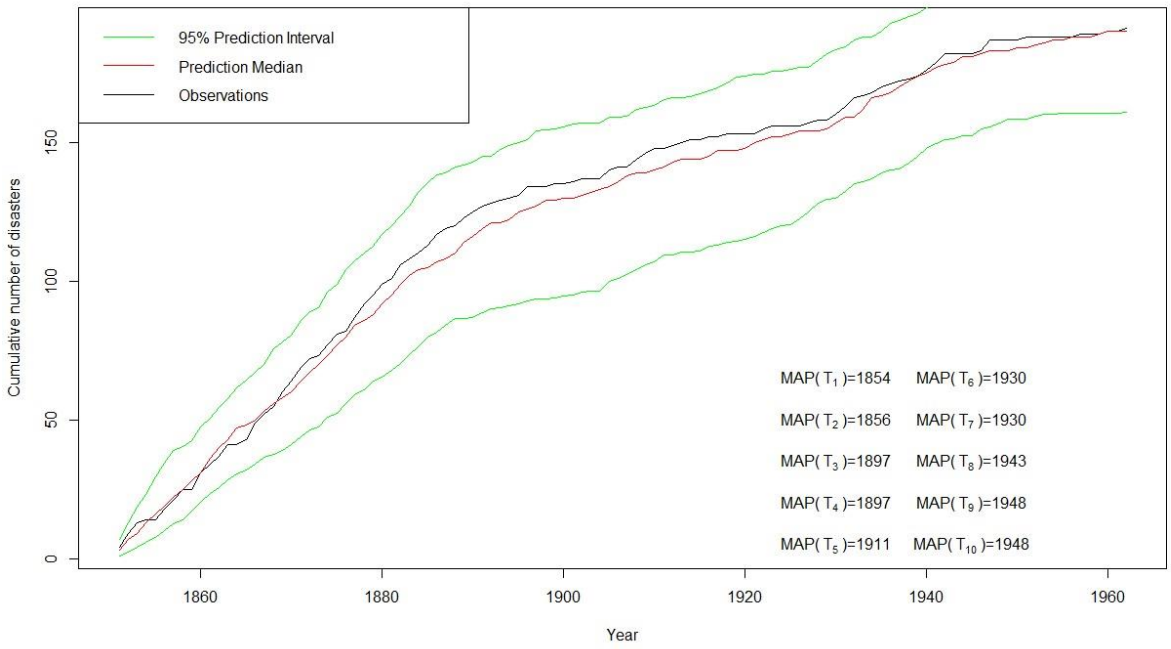


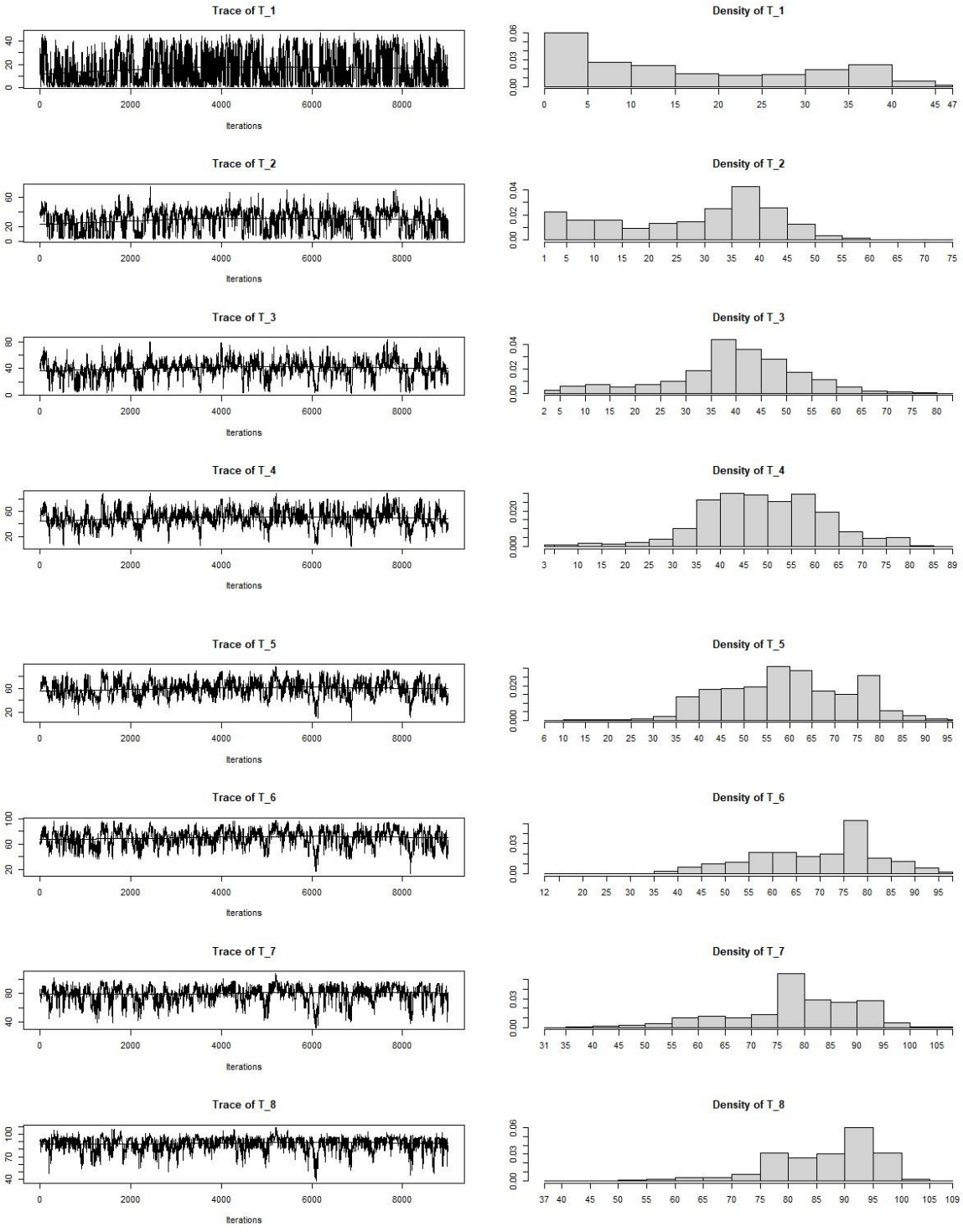
Figure 6: Posterior predictive distributions for a model with 10 changepoints

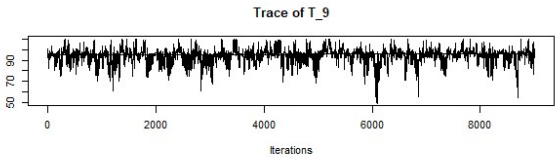
At 10 changepoints, the MAP points for the T parameters are showing strong points of convergence with 3 of the T parameters having the same MAP values and 2 T parameters being very close to one another.

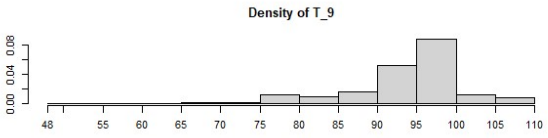
Looking at the plot in Figure 6, the median is very closely following the observed data indicating that the model is well matched for this dataset. As before, the 95% confidence interval is rather wide, which may be more likely due to a bad initial starting point for the chain as now more parameters have to be sampled and the likelihood of the T parameters choosing sub-optimal values is greater.

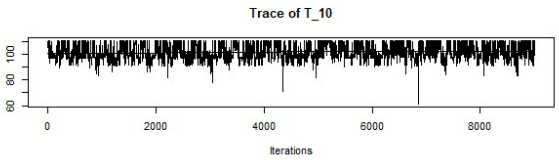
The overlap in the posterior distributions of the T parameters is now even more apparent as T_3 and T_4 show very similar distributions over nearly the same range of values, indicating that these two changepoints were likely sometimes sampled close to one another. T_5 shows a more multi-modal posterior distribution which is why it does not share the same MAP as T_6. However, the range of values of the posterior distribution of T_5 is near identical to that of T_6, which indicates a lot of overlap between these changepoints. T_8 and T_9 also show this same overlap in the range of values of their respective posterior distributions. The modes of these two T parameters are relatively close to one another as well. The fact that T_9 and T_{10} share the same MAP is not surprising as T_4 in the previous 4 changepoint model was showing a convergence point around a value of 97 which is precisely the same MAP as T_9, T_{10}, and T_9 and T_4 in the previous model.

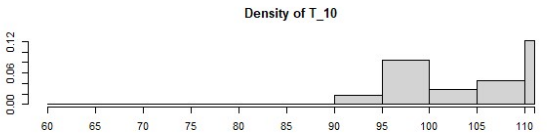
There is strong reason to believe that, although the plot in Figure 6 is showing a strong match between the data and the model, the additional complexity and model parameters (especially the additional T parameters) are incurring apparent diminishing returns as several of these parameters' distributions are seemingly overlapping with one another. This overlap is quite visible when analyzing the different density plots in Figure 7.

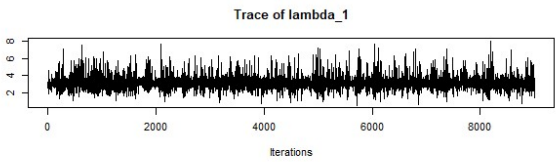

Trace of T_1
Density of T_1
Trace of T_2
Density of T_2
Trace of T_3
Density of T_3
Trace of T_4
Density of T_4
Trace of T_5
Density of T_5
Trace of T_6
Density of T_6
Trace of T_7
Density of T_7
Trace of T_8
Density of T_8
Iterations

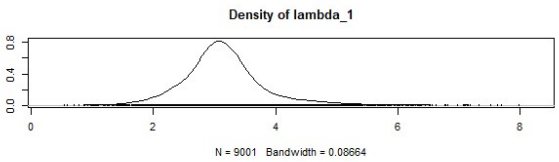

Trace of T_9
Iterations

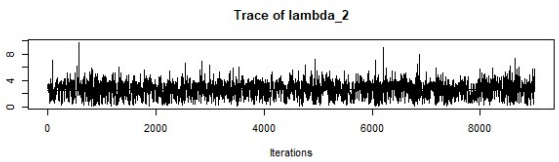

Density of T_9

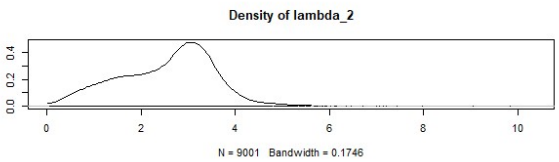

Trace of T_10
Iterations

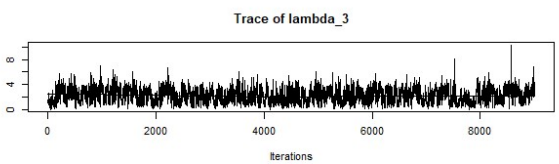

Density of T_10

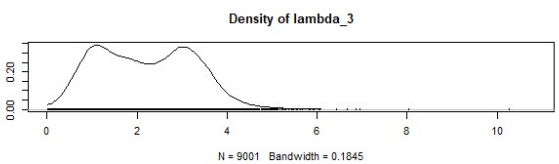

Trace of lambda_1
Iterations

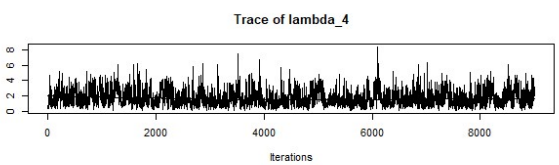

Density of lambda_1
N = 9001 Bandwidth = 0.08664

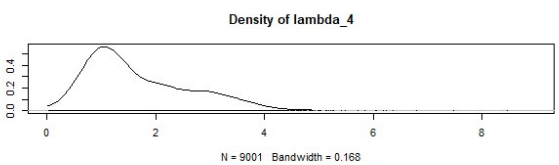

Trace of lambda_2
Iterations

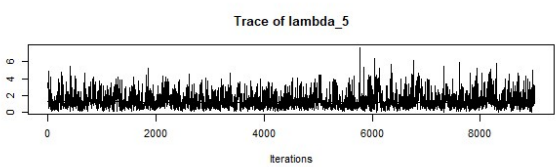

Density of lambda_2
N = 9001 Bandwidth = 0.1746

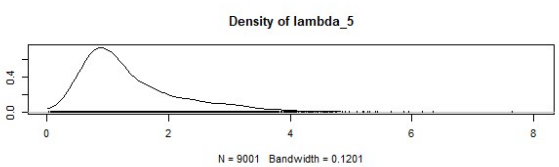

Trace of lambda_3
Iterations

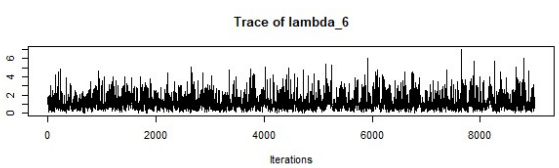

Density of lambda_3
N = 9001 Bandwidth = 0.1845

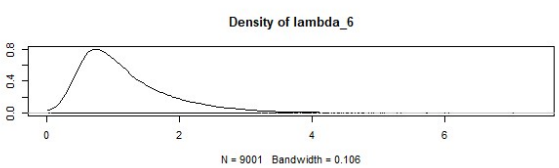

Trace of lambda_4
Iterations


Density of lambda_4
N = 9001 Bandwidth = 0.168


Trace of lambda_5
Iterations


Density of lambda_5
N = 9001 Bandwidth = 0.1201


Trace of lambda_6
Iterations


Density of lambda_6
N = 9001 Bandwidth = 0.106

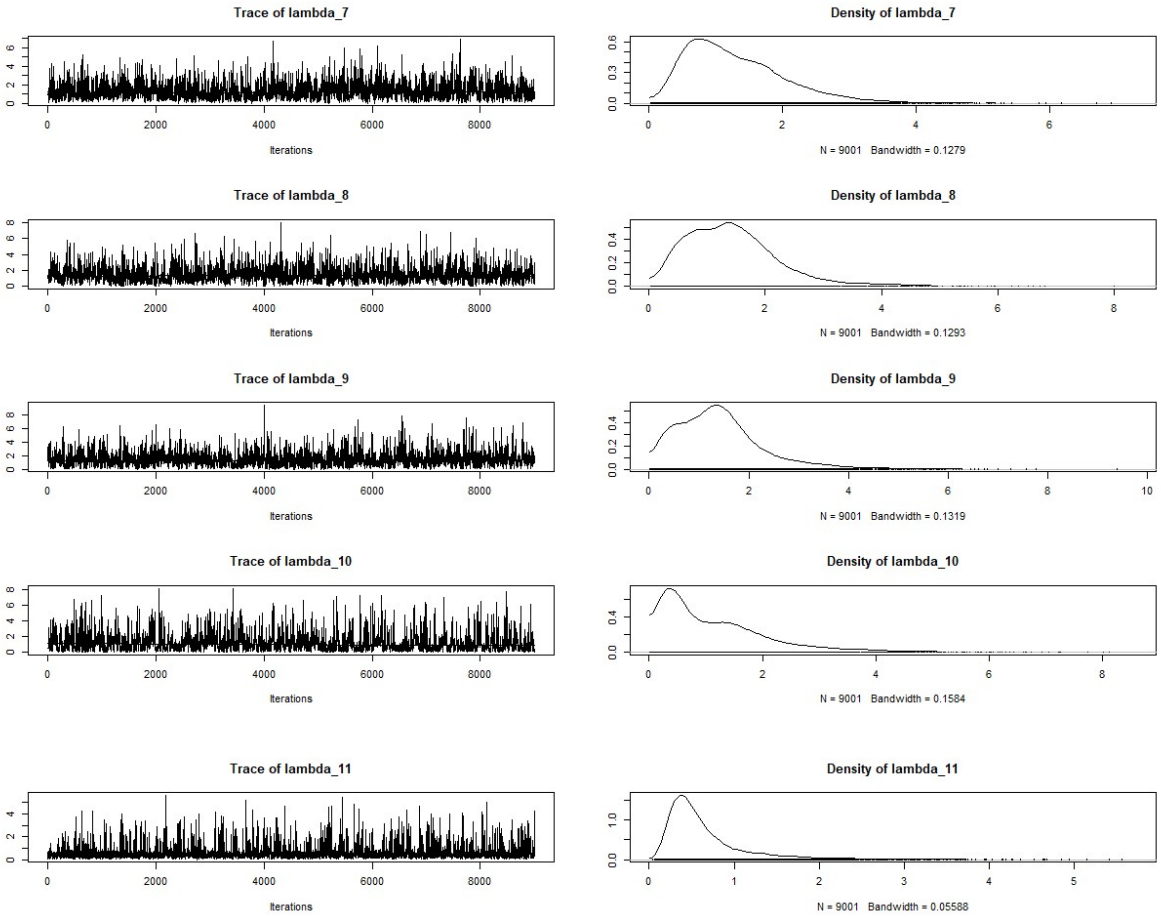


Figure 7: Trace and density plots for a model with 10 changepoints

Log Marginal likelihoods for simpler models

The Chib method values presented in this section were obtained by running 100 parallel chains and computing the mean and standard deviation of the log marginal likelihood of these chains.

Number of changepoints considered	**Log Marginal Likelihood (non-Chib)**	**Log Marginal Likelihood Mean (Chib)**	**Log Marginal Likelihood Standard Deviation (Chib)**
1	-176.4679	-176.4677	0.002386442
2	-175.6190	-175.6184	0.024462167
3	-175.3718	-175.3726	0.046724954
4	-175.2496	-175.2486	0.054651678
5	-175.2511	-175.2418	0.129821324

Figure 8: Table of log marginal likelihood estimations for simple models

Using the non-Chib method (10), one can determine a very accurate estimation of the log marginal likelihood of a given model. These estimations will not only be used to compare the different models but also to evaluate the performance of the Chib method implementation.

Based purely on the plots on Figure 9 and Figure 10, the Chib method implementation is showing good performance with a slight increase in standard deviation as the complexity of the model increases. The mean value of the approximations is very close to the accurate values obtained from the non-Chib method.

Regarding the performance of each model in respect one another, the marginal likelihood values suggest that the models with 4 or 5 changepoints are the best to model this dataset.

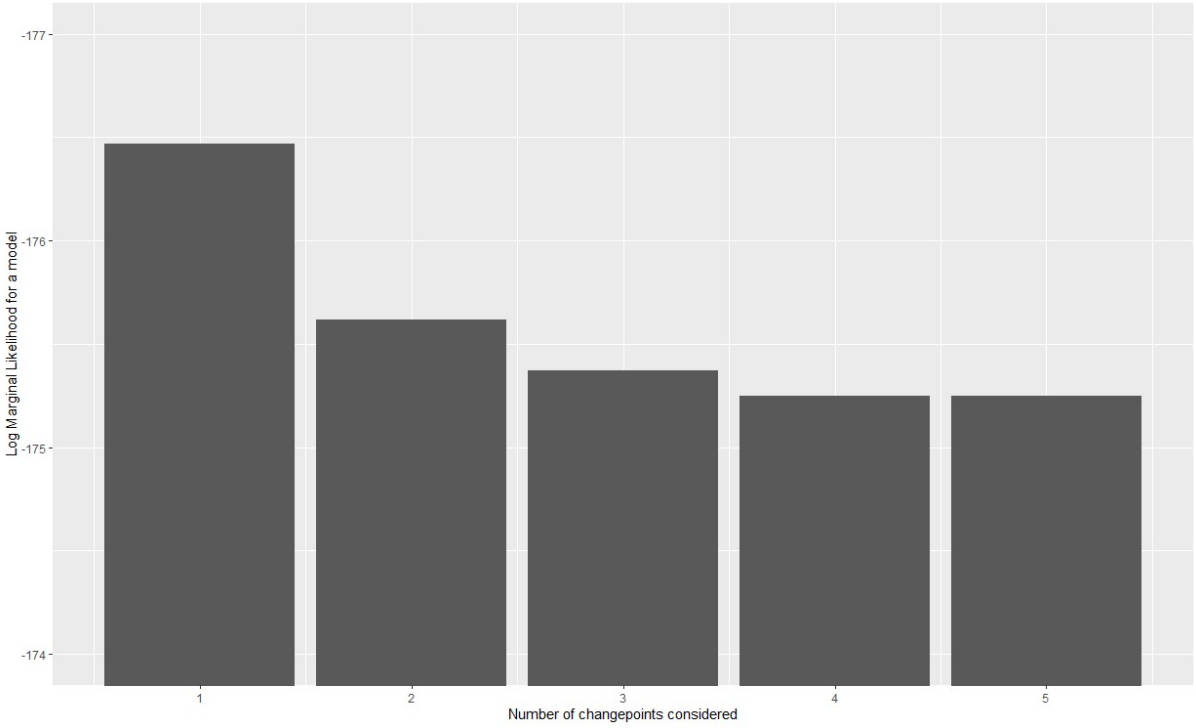


Figure 9: Plot of log marginal likelihood estimations (not using Chib method). The plot's Y-axis is reversed for better readability. Thus, the lower the bar is (closer to 0), the better.

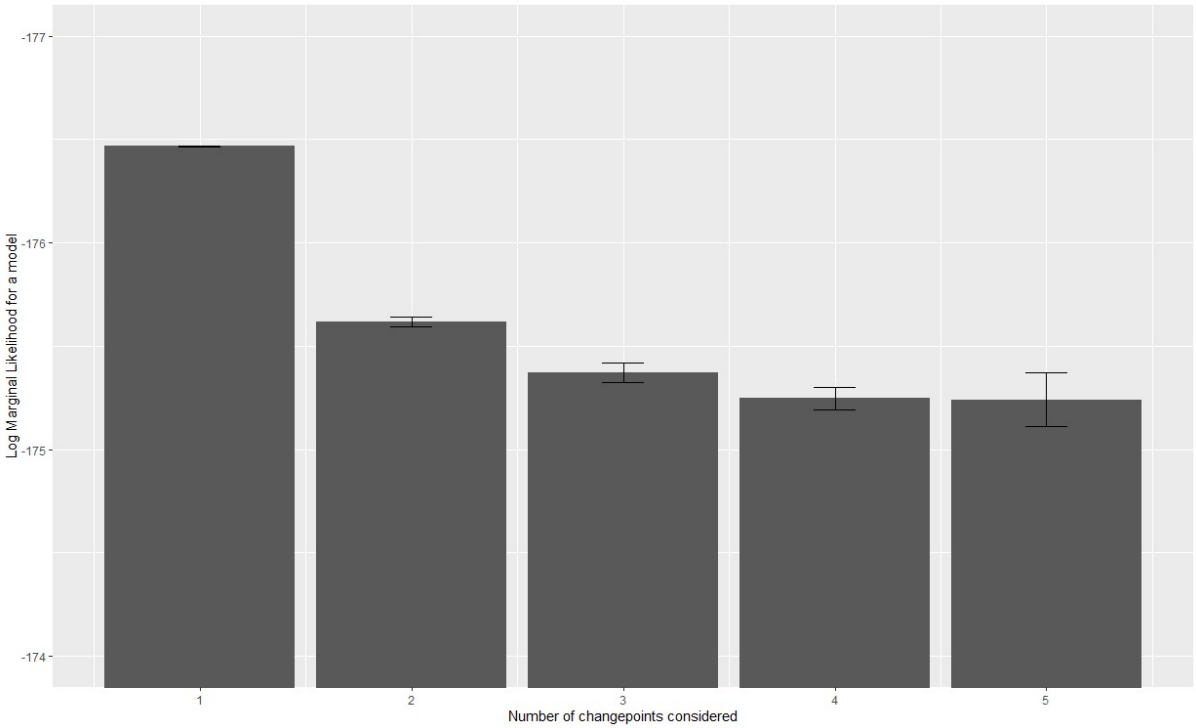


Figure 10: Plot of log marginal likelihood estimations (using Chib method). The plot's Y-axis is reversed for better readability. Thus, the lower the bar is (closer to 0), the better. The error bars represent the standard deviation obtained by running Chib method on 100 parallel chains.

Log Marginal likelihoods for more complex models

As with the previous section, the Chib method values presented in this section were obtained by running 100 parallel chains and computing the mean and standard deviation of the log marginal likelihood of these chains.

Number of changepoints considered	**Log Marginal Likelihood (non-Chib)**	**Log Marginal Likelihood Mean (Chib)**	**Log Marginal Likelihood Standard Deviation (Chib)**
5	-175.2511	-175.2418	0.129821324
10	N/A	-176.0643	0.1545436
21	N/A	-178.6359	0.2315488
32	N/A	-181.2590	0.3200125
43	N/A	-183.7383	0.4231511

Figure 11: Table of log marginal likelihood estimations for more complex models

Unfortunately, the non-Chib method cannot be used to estimate the log marginal likelihood of models with more than 5 changepoints due to the computational cost increasing very dramatically and leading to an eventual combinatorial explosion.

For this reason, the marginal likelihood estimation of any models beyond 5 changepoints was done only using the Chib method. However, despite the lessened computational cost from running the Chib method, it remains an approximation and, as such, should still be run with multiple parallel chains even if the previous section showed good performance numbers.

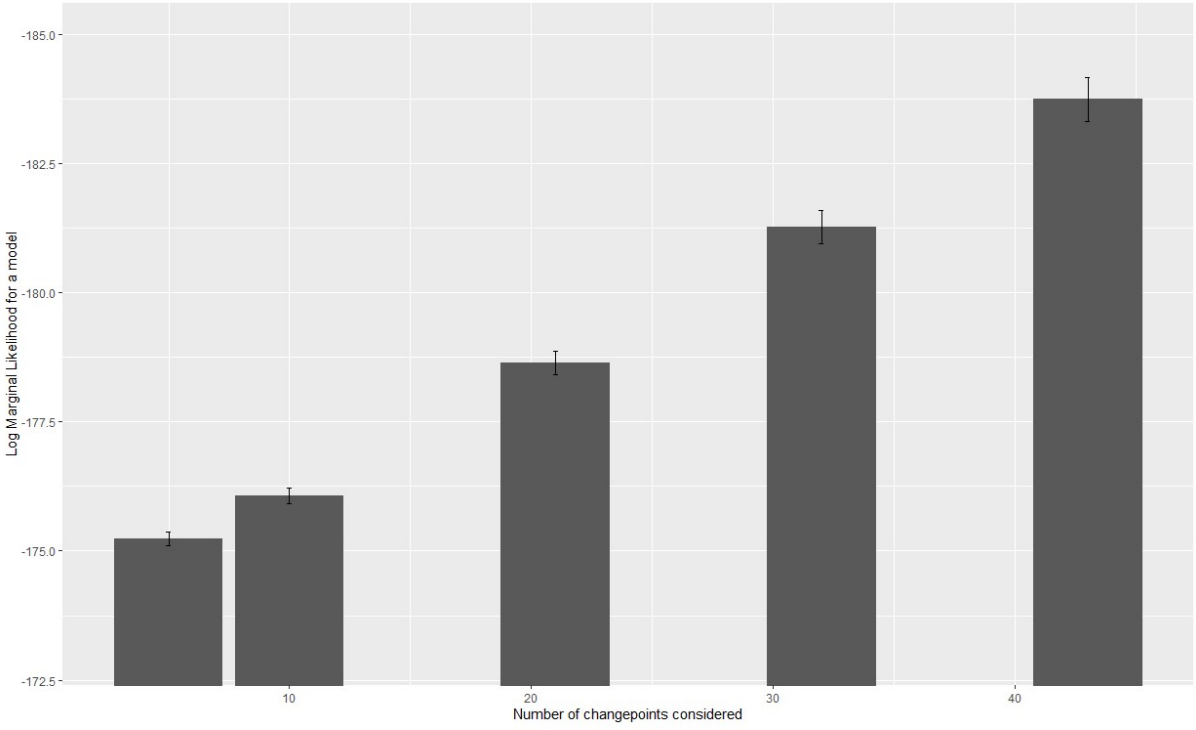


Figure 12: Plot of log marginal likelihood estimations (using Chib method) for 5, 10, 21, 32, and 43 changepoint models. The plot's Y-axis is reversed for better readability. Thus, the lower the bar is (closer to 0), the better. The error bars represent the standard deviation obtained by running Chib method on 100 parallel chains.

As was the case in the previous section, the 5 changepoint model is clearly showing the best performance out the five models presented in Figure 11Figure 12. However, the standard deviation of the Chib approximations has increased quite noticeably.

The increase in standard deviation whilst using Chib is a well-documented issue [8] that was brought up not long after publication of Chib's original article. The issue comes from the Rao-Blackwellized estimators creating biased posterior densities that do not account for $K!$ equally likely posterior modes (where K is the number of mixture components) [9]. The initial fix proposed in [8] increases the computational cost of Chib to the point of making the method very inefficient. Further improvements were in suggested in [10] and, more recently, [9] with varying levels of success.

Due to time constraints, these fixes could not be added to the Chib method implementation used in this project. The increasing standard deviation was only noticed at a very late stage in the project whilst analyzing the full output from the Chib approximations. The papers cited were only briefly skimmed in an effort to determine if the issue was coming from an error in the implementation of the Chib method or if it was a known issue with the method. Thus, even more time would be needed to perform a more thorough investigation of the problem and evaluate which of the proposed fixes would best fit the purposes of the project.

Conclusion and Closing Thoughts

This report addresses how to implement a Gibbs sampler and, by extension, a Chib marginal likelihood estimator for the provided coal mining disasters dataset. Although the Chib estimator showed signs of increase in variance as the models considered get more and more complex, this project has laid the foundations to be able to continue to work on this estimator and refine it until it provides more satisfying results with more complex models.

If this project were to be continued in the future, then the first objective would be to fix the Chib method used. The estimation of the more complex models using the current implementation can take up to a quarter of day at best. Considering this time cost, the best course of action is to at least make sure that the estimator is generating "good" approximations with as low variance as possible. Alternatively, other sampling or marginal likelihood approximation methods can also be considered (whether those methods make use of the already implemented Gibbs sampler or not).

On a more personal note, the investigations done over the course of this project have helped me get a better grip on the subject of Bayesian statistics and I will not hide that the implementation process of the Gibbs sampler and especially the Chib method proved to be quite the challenge given my time constraints and initial understanding of the subject prior to starting this project. Working out the distributions by hand, then implementing them in code and seeing the seemingly good results I was getting really pushed me further into tweaking my implementation to get the most of it and get results for the more complex models. Though I had initially hoped that I would have enough time to compute several spread out models up to the 111 changepoint model and present them in the corresponding log marginal likelihood section.

Bibliography

[1] A. J. Taylor, ""“The Coal Industry.”," in *The Development of British Industry and Foreign Competition 1875–1914*, vol. 66, D. H. ALDCROFT, Ed., University of Toronto Press, 1968, pp. 37-70.

[2] "Coal Mines Regulation Act, 1887," [Online]. Available: http://www.scottishmining.co.uk/256.html.

[3] T. Dwyer and A. Raftery, "Industrial accidents are produced by social relations of work: A sociological theory of industrial accidents," *Applied Ergonomics,* vol. 22, no. 3, p. 167–178, 1991.

[4] R. G. Jarrett, "A Note on the Intervals Between Coal-Mining Disasters," *Biometrika,* vol. 66, no. 1, pp. 191-193, 1979.

[5] "7.2 Change point models," [Online]. Available: https://mc-stan.org/docs/stan-users-guide/change-point.html.

[6] S. Chib, "Marginal likelihood from the Gibbs output.," *Journal of the American Statistical Association,* no. 90, pp. 1313-1321, 1995.

[7] A. E. Gelfand and A. F. M. Smith, "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association,* no. 85, pp. 501-514, 1990.

[8] R. M. Neal, *Erroneous results in “Marginal likelihood from the Gibbs output”,* 1999.

[9] A. Hairault, C. P. Robert and J. Rousseau, "Evidence estimation in finite and infinite mixture models and applications," 2022.

[10] J. Berkhof, I. van Mechelen and A. Gelman, "A Bayesian approach to the selection and testing of mixture models," *Statistica Sinica,* p. 423–442.