Excercise 2

TMA4300 Computer intensive statistical methods Spring 2021

Johan Fredrik Agerup Arne Rustad

09 mars, 2021

Problem A

In this problem we analyze a data set of time intervals between successive coal-mining disaster in the UK involving ten or more men killed. The data is for the period March 15th 1851 to March 22nd 1962.

You should note that the first and last records in this variable are the start and end dates, respectively. IS THIS TRUE??

1)

First we try to get an impression of the data set by making a plot with year along the x-axis and the cumulative number of disasters along the y-axis.

```
df.coal = coal
df.coal$cum.n.explosions = cumsum(rep(1, nrow(coal)))
head(df.coal)
```

```
##       date cum.n.explosions
## 1 1851.203                1
## 2 1851.632                2
## 3 1851.969                3
## 4 1851.975                4
## 5 1852.314                5
## 6 1852.347                6
```

```
ggplot(df.coal, aes(x = date, y = cum.n.explosions)) + geom_line() + ggtitle("Cumulative number of explo
```

From Figure 1 the rate of accidents appear approximately constant from year 1850 until around 1890. Then the rate of large explosions appear to dampen a bit, perhaps due to better safety rutines, equipment, change in societal norms and laws or another reasson.

2)

To analyze the data set we adopt a hierarchical Bayesian model. Assume the coal-mining disasters to follow an inhomogeneous Poisson process with intensity function $\lambda(t)$ (number of events per year). Assume $\lambda(t)$ to be piecewise constant with n breakpoints. Let t_0 and t_{n+1} denote the start and end times for the dataset and let $t_k;\ k = 1, 2, \ldots, n$ denote the break points of the intensity function. Thus,

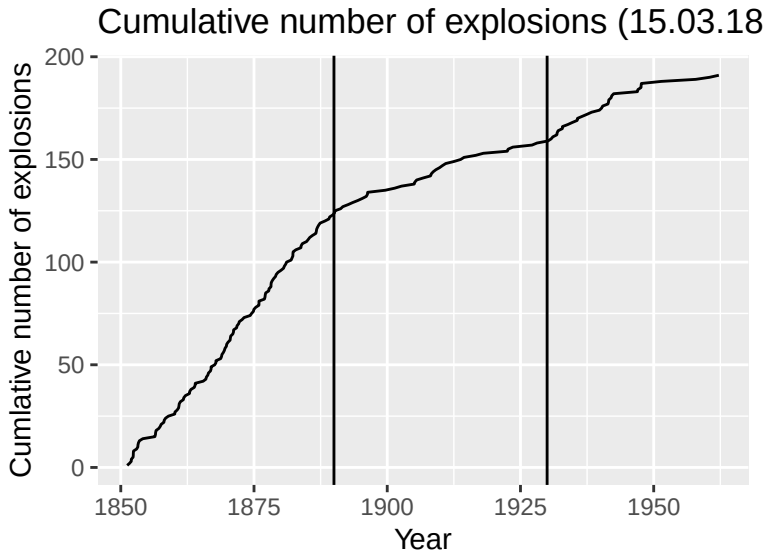

Figure 1: Cumulative umber of explosions which resulted in 10 or more fatalities. The time span is from March 15 1851 until March 22 1962.

$$\lambda(t) = \begin{cases} \lambda_{k-1} & \text{for } t \in [t_{k-1}, t_k] \\ \lambda_n & \text{for } t \in [t_n, t_{n+1}] \end{cases}$$

Thereby the parameters of the model is $t_1, \ldots, t_n$ and $\lambda_0, \ldots, \lambda_n$ where $t_0 < t_1 < \cdots < t_n < t_{n+1}$. By subdividing the observations into short intervals and taking the limit when the length of these intervals go to zero, the likelihood function for the observed data can be derived as

$$\begin{aligned} f(x|t_1, \ldots, t_n, \lambda_0, \ldots, \lambda_n) &= e^{-\int_{t_0}^{t_{n+1}} \lambda(t)dt} \prod_{k=0}^{n} \lambda_k^{y_k} \\ &= e^{-\sum_{k=0}^{n} \lambda_k (t_{k+1} - t_k)} \prod_{k=0}^{n} \lambda_k^{y_k}. \end{aligned}$$

?????????????????????????????????? should it here be a normalizing constant as well in the equation

Here x is the observed data and y_k is the number of observed disasters in the period t_k to t_{k+1}. Assume $t_1, \ldots, t_n$ to be apriori uniformly distributed on the allowed values and $\lambda_0, \ldots, \lambda_n$ to be apriori independent of $t_1, \ldots, t_n$ and apriori independent of each other. Apriori we assume all $\lambda_0, \ldots, \lambda_n$ to be distributed from the same gamma distribution with shape parameter $\alpha = 2$ and scale parameter β, i.e

$$f(\lambda_i|\beta) = \frac{1}{\beta^2} \lambda_i e^{\frac{-\lambda_i}{\beta}} \text{ for } \lambda_i \geq 0$$

Finally, for β we use the improper prior

$$f(\beta) \propto \frac{e^{\frac{-1}{\beta}}}{\beta} \text{ for } \beta > 0$$

In the following it is assumed $n = 1$, resulting in $\theta = (t_1, \lambda_0, \lambda_1, \beta)$.

The posterior distribution is then

$$\begin{aligned} f(\theta|x) &= f(x|\theta)f(\theta) \\ &= f(x|t_1, \lambda_0, \lambda_1)f(t_1)f(\lambda_0|\beta)f(\lambda_1|\theta)f(\beta) \end{aligned}$$

Here it is used that t_1, λ_0 and λ_1 all are independent of each other. Inserting the expressions for the likelihood and the priors we get an expression for the posterior distribution up to a proportionality constant

$$\begin{aligned} f(\theta = (t_1, \lambda_0, \lambda_1, \beta)|x) &\propto e^{-\lambda_0(t_1-t_0)-\lambda_1(t_2-t_1)}\lambda_0^{y_0}\lambda_1^{y_1}\frac{1}{t_2-t_0}\frac{1}{\beta^2}\lambda_0 e^{\frac{-\lambda_0}{\beta}}\frac{1}{\beta^2}\lambda_1 e^{\frac{-\lambda_1}{\beta}}\frac{e^{\frac{-1}{\beta}}}{\beta} \\ &\propto e^{-\lambda_0(t_1-t_0)-\lambda_1(t_2-t_1)}\lambda_0^{y_0}\lambda_1^{y_1}\frac{1}{\beta^2}\lambda_0 e^{\frac{-\lambda_0}{\beta}}\frac{1}{\beta^2}\lambda_1 e^{\frac{-\lambda_1}{\beta}}\frac{e^{\frac{-1}{\beta}}}{\beta}. \end{aligned}$$

3)

Next we want to find the full conditions of θ for later use in the implementation of MCMC algorithms to find the posterior. The full conditional of t_1 is

$$f(t_1|x, \lambda_0, \lambda_1, \beta) \propto e^{t_1(\lambda_1-\lambda_0)}, \qquad t_1 \in [t_0, t_2].$$

??????????????????????????????????+ Should β and x be included after | here?

The full conditional of t_1 is not recognized as a known distribution???????.

The full conditional of λ_0 is

$$f(\lambda_0|x, t_1, \lambda_1, \beta) \propto \lambda_0^{y_0+1}e^{-\lambda_0(t_1-t_0+\frac{1}{\beta})}, \qquad \lambda_2 \geq 0.$$

We recognize the full conditional of λ_0 as the Gamma$(y_0 + 2, \frac{1}{t_1-t_0+\frac{1}{\beta}})$ distribution. Similarily, the full conditional of λ_1 is

$$f(\lambda_1|x, t_1, \lambda_0, \beta) \propto \lambda_1^{y_1+1}e^{-\lambda_1(t_2-t_1+\frac{1}{\beta})}, \qquad \lambda_1 \geq 0.$$

From this we see that the full conditional of λ_1 is Gamma$(y_1+2, \frac{1}{t_2-t_1+\frac{1}{\beta}})$ distributed. Lastly, the expression for the full conditional of β is

$$f(\beta|x, t_1, \lambda_0, \lambda_1) \propto \frac{1}{\beta^5}e^{\frac{-(1+\lambda_0+\lambda_1)}{\beta}}, \qquad \beta > 0.$$

Here $\frac{1}{\beta}$ appear Gamma$(6, \frac{1}{1+\lambda_0+\lambda_1})$ distributed.

#Exercise B: Our main inferential interest lies in the posterior marginal for the smooth effect $\Pi(\eta(t)|y), t = 1, ..., T$.

1. Explain why this model is a latent Gaussian model and why it is possible to use INLA to estimate the parameters.

A latent variable model relates a set of observable variables to a set of inferred variables. A latent Gaussian model is a model which infers a variable based on observed variables with that have a Gaussian distribution.

2. Define and implement a block Gibbs sampling algorithm for $f(\eta, \Theta|y)$ using the following two (block) proposals: Propose a new value for Θ from the full conditional $\Pi(\Theta|\eta, y)$ Propose a new value for the vector η from the full conditional $\Pi(\eta|\Theta, y)$ Use the samples to get an estimate for the posterior marginal for the hyperparameter $\Pi(\Theta|y)$ Use the samples to get an estimate of the smooth effect using the mean and pointwise a 95% confidence bound around the mean.
3. We want to approximate the posterior marginal for the hyperparameter $\Theta, \Pi(\Theta|y)$ using the INLA scheme. We start from:

$$\Pi(\Theta|y) \propto \frac{\Pi(y|\eta, \Theta)\Pi(\eta|\Theta)\Pi(\Theta)}{\Pi(\eta|\Theta, y)} \tag{1}$$

Note that since the likelihood is Gaussian then also $\Pi(\eta|\Theta, y)$ is Gaussian. Consider a grid of value between 0 and 6 and use 5 to construct an approximation for $\Pi(\Theta|y)$. Compare your result with the MCMC estimate you obtained in point 1) 4. We now want to implement the next step in the INLA scheme, the approximation of the marginal posterior for the smooth effect, $\Pi(\eta_i|y)$. We have that:

$$\pi(\eta_i|y) = \int \pi(\eta_i|y, \Theta)\Pi(\Theta|y)d\Theta \tag{2}$$

Use the grid of Θ value from point 2) to approximate the integral above for $i = 10$. Compare your approximation for $\Pi(\eta_i|y)$ with the estimation obtained in point 1) via Gibbs sampling.