SF2955 Computer Intensive Methods
Home Assignment 2

Group : Klara Zimmerman and Ville Wassberg

May 2024

Problem 1

In this problem, we look at coal mining disasters between the years of 1851 and 1963. We are given the time instances of each disaster, τ, and our task is to find the posterior distribution of the $d-1$ breakpoints t, the d intensities λ of each interval $[t_i t_{i+1})$, and the parameter θ, given τ.

We are given the prior distribution of the d breakpoints $t = t_1, ..., t_d$

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

as well as the prior distribution $\theta \sim \Gamma(2, \nu)$. Further, we know the conditional distribution $\lambda_i | \theta \sim \Gamma(2, \theta)$ and the the conditional distribution $f(\tau|\lambda, t) \propto e^{-\sum_{i=1}^{d} \lambda_i (t_{i+1} - t_i)} \prod_{i=1}^{d} \lambda_i^{n_i(\tau)}$, where $n_i(\tau)$ gives the number of disasters in the sub-interval $[t_i, t_{i+1})$.

We will now find a posterior distribution $f(\theta, \lambda, t|\tau)$, from which a hybrid MCMC algorithm samples.

We begin by finding the conditional distributions $f(\theta|\lambda, t, \tau), f(\lambda|\theta, t, \tau)$ and $f(t|\theta, \lambda, \tau)$, up to normalizing constants. If we can identify these distributions, we can update the variables using Gibbs sampling. Otherwise, we will use a Metropolis-Hastings step.

In each of the calculations below, we have used that

$$f_{X_1|X_2,X_3,X_4} = \frac{f_{X_1,X_2,X_3,X_4}}{f_{X_2,X_3,X_4}} \propto f_{X_1,X_2,X_3,X_4} \propto f_{X_1} \cdot f_{X_2|X_1} \cdot f_{X_3|X_1,X_2} \cdot f_{X_4|X_1,X_2,X_3},$$

i.e. the conditional distribution equals the joint distribution, up to a normalizing constant, which can be factored using the chain rule for distributions.

$$\begin{aligned} f(\theta|\lambda, t, \tau) \propto f(\theta) f(\lambda|\theta) f(t) f(\tau|\lambda, t) \propto f(\theta) \prod_{i=1}^{d} f(\lambda_i|\theta) &= \nu^2 \theta e^{-\nu\theta} \prod_{i=1}^{d} \theta^2 \lambda_i e^{-\theta\lambda_i} \\ &= \nu^2 (\prod_{i=1}^{d} \lambda_i) \theta^{1+2d} e^{-\theta\nu + \theta \sum_{i=1}^{d} \lambda_i)} \propto \theta^{1+2d} e^{-\theta(\nu + \sum_{i=1}^{d} \lambda_i)} \end{aligned}$$

By letting $\alpha = 2 + 2d$ and $\beta = \nu + \sum_{i=1}^{d} \lambda_i$, we thus get

$$f(\theta|\lambda, t, \tau) \propto \theta^{\alpha-1} e^{-\beta\theta} \propto \beta^{\alpha} \theta^{\alpha-1} e^{-\beta\theta},$$

which we recognize as a gamma distribution. That is,

$$\theta|\lambda, t, \tau \sim \Gamma(2 + 2d, \nu + \sum_{i=1}^{d} \lambda_i)). \tag{1}$$

We also have that

$$f(\lambda_i|\theta,t,\tau) \propto f(\theta)f(\lambda_i|\theta)f(t)f(\tau|\lambda_i,t) \propto f(\lambda_i|\theta)f(\tau|\lambda_i,t) \propto$$
$$\theta^2\lambda_i e^{-\theta\lambda_i}e^{-\lambda_i(t_{i+1}-t_i)}\lambda_i^{n_i(\theta)} \propto \lambda_i^{1+n_i(\theta)}e^{-\lambda_i(\theta+\lambda_i(t_{i+1}-t_i)},$$

thus we set $\alpha = 2 + n_i(\tau)$ and $\beta = \theta + t_{i+1} - t_i$, and get $f(\lambda_i|\theta,t,\tau) \propto \beta^\alpha\theta^{\alpha-1}e^{-\beta\theta}$, which means that up to a normalizing constant,

$$\lambda_i|\theta,t,\tau \sim \Gamma(2+n_i(\tau),\theta+t_{i+1}-t_i). \tag{2}$$

We assume that λ_i are independent, so finding the conditional density of $\lambda|\theta,t,\tau$ corresponds to multiplying the marginal distributions for each i.

$$f(t|\theta,\lambda,\tau) \propto f(\theta)f(\lambda|\theta)f(t)f(\tau|\lambda,t) \propto f(t)\prod_{i=1}^{d}f(\tau|\lambda_i,t) =$$
$$= \begin{cases} \prod_{i=1}^{d}(t_{i+1}-t_i)e^{\sum_{i=1}^{d}\lambda_i(t_{i+1}-t_i)}\prod_{i=1}^{d}\lambda_i^{n_i(\tau)} & \text{for } t_1 < ... < t_{d+1} \\ 0 & \text{else} \end{cases} \tag{3}$$
$$\propto \begin{cases} \prod_{i=1}^{d}(t_{i+1}-t_i)e^{\sum_{i=1}^{d}\lambda_i(t_{i+1}-t_i)} & \text{for } t_1 < ... < t_{d+1} \\ 0 & \text{else} \end{cases}$$

We do not recognize the distribution (3), hence we will use a Metropolis-Hastings step to update the breakpoints t.

To sample from the posterior $f(\theta,t,\lambda|\tau)$, we will now construct the hybrid MCMC algorithm. For $1 \le i \le N$ steps, we will create a Markov chain by updating λ^i and θ^i with Gibbs sampling using the conditional densities (1) and (2), and using a random walk proposal to update t^i with a Metropolis-Hastings step.

Selecting θ^{i+1} is done by drawing a new candidate from the posterior distribution $f(\theta|\lambda^i,t^i,\tau)$. λ_j^{i+1} is found by sampling from $f(\lambda_j|\theta^{i+1},t_j^i,\tau)$, where we can use the updated θ^{i+1}, and where j loops from 0 to d, updating λ_j individually for each interval $[t_j,t_{j+1})$.

In accordance with the random walk proposal, we updated each t_j^{i+1} in t^{i+1}, one at a time, by setting $t_j^* = t_j^i + \epsilon_j$ where $\epsilon_j \sim Unif(-R,R)$ and $R = \rho(t_{j+1}^i - t_{j-1}^i)$. By the Metropolis-Hastings algorithm, we then calculate $\alpha = 1 \wedge \frac{f(t^*)r(t^*;t_j^i)}{f(t_j^i)r(t_j^i;t^*)}$, where r is the transition kernel.

The random walk update of t_j^* is symmetric, that is the probability of moving from t_j* to t_j^i is equal to the probability of moving from t_j^i to t_j*. This means that $r(t_j^*|t_j^i) = r(t_j^i|t_j^*)$, so these will cancel out with the division.

As for the density of t, we will use the conditional density (3). Thus, we get

$$\alpha = 1 \wedge \frac{f(t_j^*|\lambda_j^{i+1}, \theta^{i+1}, \tau)}{f(t_j^i|\lambda_j^{i+1}, \theta^{i+1}, \tau)}.$$

We then draw $u \sim Unif(0,1)$. If $u \leq \alpha$, we set $t_j^{i+1} = t_j^*$, and otherwise we set $t_j^{i+1} = t_j^i$. This completes the Metropolis-Hastings step.

Investigating the Results

We now want to investigate the behaviour of the MCMC chain for 1, 2, 3 and 4 breakpoints. We thus plot histograms of the parameters λ and t, with $N = 10000$.

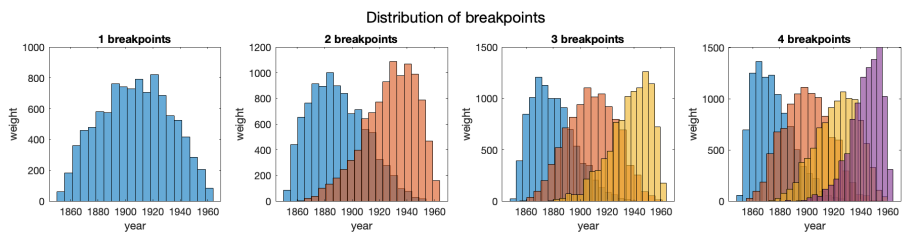

Figure 1: Distributions of breakpoints t, for 1,2,3 and 4 breakpoints.

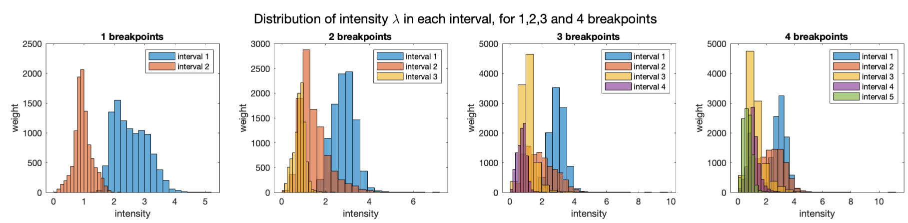

Figure 2: Distributions of intensity λ, in each interval for 1,2,3 and 4 breakpoints.

In figure 1 we see that with one breakpoint, t_1 is expected to be around 1910. With 2 breakpoints we expect $t_1 \approx 1880, t_2 \approx 1935$, with 3 breakpoints we expect $t_1 \approx 1870, t_2 \approx 1910, t_3 \approx 1950$ and with 4 breakpoints we expect $t_1 \approx 1865, t_2 \approx 1895, t_3 \approx 1930, t_4 \approx 1950$.

As for the intensities in figure 2, we expect slightly decreasing intensities for each new interval added, and for 4 breakpoints this is $\lambda_1 \approx 3, \lambda_2 \approx 2.5, \lambda_3 \approx 1.5, \lambda_4 \approx 1$ and $\lambda_4 \approx 0.5$.

Further, we see that the number of breakpoints does not seem to affect the distribution of θ. This is shown in figure 3, where we see that the distributions are approximately equal, regardless of how many breakpoints are used.

We want to investigate how sensitive the posteriors are to the choice of the hyperparameter ν. This parameter is the rate parameter of the prior distribution of θ. In figure 3 we have plotted the posterior distribution of θ, with different values of ν. As we see, increasing ν squeezes the height of the distribution of θ slightly, and it is shifted towards 0 as ν grows very large, while centered around 1.5 for small ν. However, we see in figure 4 that the distribution of breakpoints maintains the same, so the posterior of t does not seem sensitive to ν.

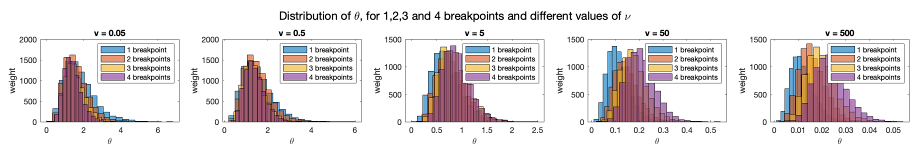

Figure 3: Distribution of θ, for 1,2,3 and 4 breakpoints, with $\nu = 0.05, 0.5, 5, 50$ and 500.

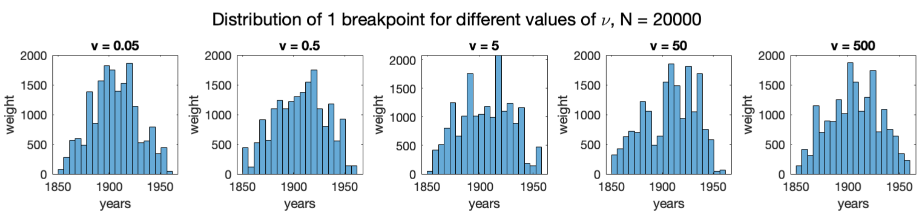

Figure 4: Distribution of t, for 1 breakpoint, with $\nu = 0.05, 0.5, 5, 50$ and 500.

Next, we want to look at how sensitive the mixing and the posteriors is to the choice of parameter ρ, used in the random walk proposal to find t^* in the MH-step of our MCMC algorithm.

As we can see in figure 5, where the expectation of 1 breakpoint is plotted, the expected value of the breakpoint seems to converge best when ρ is approximately 1. To investigate this further, we look at the state space exploration (figure 6) and the ACF plots (figure 7) for the corresponding values of ρ, which further confirms that proposal variance is too small for $\rho = 0.0125$, and too large for $\rho = 20$.

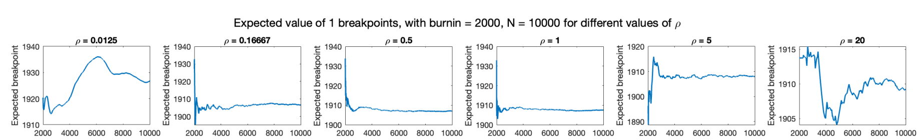

Figure 5: Expected value of one breakpoint, for different values of ρ.

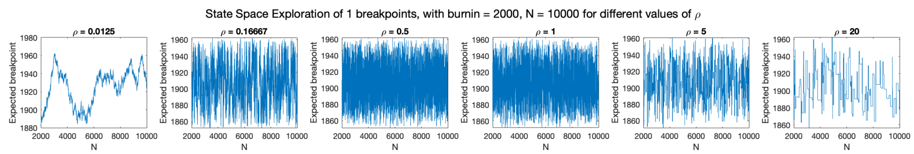

Figure 6: State space exploration of one breakpoint, for different values of ρ.

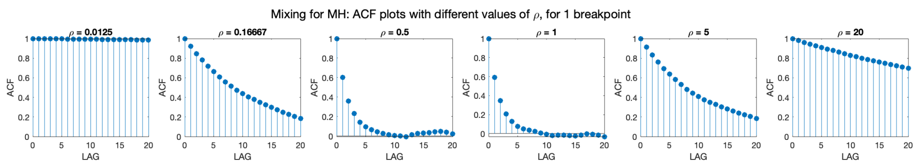

Figure 7: ACF for one breakpoint, for different values of ρ.

Problem 2

In this problem we use Hamiltonian Monte Carlo (HMC) to sample from a circle-shaped posterior. In order to write the HMC algorithm, we first need to find the logarithm of the posterior density $f(\theta|y)$, up to an additive constant. We know that $f(\theta|y) \propto f(y|\theta)f(\theta)$, which means that

$$lnf(\theta|y) \propto ln(f(y|\theta)f(\theta)) = lnf(y|\theta) + lnf(\theta).$$

The distributions

$$\begin{cases} f(y|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{y_i - (\theta_1^2+\theta_2^2)}{\sigma})^2} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(\frac{y_i-(\theta_1^2+\theta_2^2)}{\sigma})^2} \\ f(\theta) = \frac{1}{2\pi det(\Sigma)} e^{-\frac{1}{2}\theta^T\Sigma^{-1}\theta} \end{cases}$$

are given, so we get

$$lnf(\theta|y) \propto \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(\frac{y_i-(\theta_1^2+\theta_2^2)}{\sigma})^2}\right) + \ln\left(\frac{1}{2\pi det(\Sigma)} e^{-\frac{1}{2}\theta^T\Sigma^{-1}\theta}\right)$$

$$\propto -\frac{1}{2}\left(\sum_{i=1}^{n}\left(\frac{y_i-(\theta_1^2+\theta_2^2)}{\sigma}\right)^2 + \theta^T\Sigma^{-1}\theta\right) = -\frac{1}{2}\left(\sum_{i=1}^{n}\left(\frac{y_i-\theta^T\theta}{\sigma}\right)^2 + \theta^T\Sigma^{-1}\theta\right).$$

Thus, since Σ is symmetric, the gradient becomes

$$\nabla_\theta \ln f(\theta|y) \propto -\left(\sum_{i=1}^{n}\left(\frac{y_i-\theta^T\theta}{\sigma^2}\right)(-2\theta) + \Sigma^{-1}\theta\right).$$

Hence, in a Hamiltonian Monte Carlo setting, we have

$$\begin{cases} U(\theta|y) = -\ln f(\theta|y) = \frac{1}{2}\left(\sum_{i=1}^{n} \left(\frac{y_i - (\theta^T \theta)}{\sigma} \right)^2 + \theta^T \Sigma^{-1} \theta \right) \\ \nabla U(\theta|y) = -\nabla_\theta \ln f(\theta|y) = \left(\sum_{i=1}^{n} \left(\frac{y_i - \theta^T \theta}{\sigma^2} \right)(-2\theta) + \Sigma^{-1}\theta \right). \end{cases}$$

Algorithm 1 Hamiltonian Monte Carlo

Require: $\sigma, \Sigma, N, L, \varepsilon,\ H(\theta, v) = U(\theta) + K(v)$

```
function leapFrog(θ, v, σ, Σ, ε, L)
    v ← v − 0.5ε∇_θU
    for m = 1 to L − 1 do:
        θ ← θ + εv
        v ← v − ε∇_θU
    end for
    θ ← θ + εv
    v ← v − 0.5ε∇_θU
    v ← −v
    return θ, v
end function
function HMC(θ, y, v):
    for n = 2 to N do:
        draw v_0 ∼ N_2(0, I)
        θ, v ← leapFrog(θ, v_0, σ, Σ, ε, L)
        α ← min 1, exp (H(θ_{n−1}, v_0) − H(θ_n, v))
        draw rand ∼ U(0, 1)
        if rand ≤ α then
            accept θ
        else
            θ_n ← θ_{n−1}
        end if
    end for
end function
```

The HMC algorithm is now compared to the Metropolis Hastings algorithm, with random walk proposal ($\theta^* \sim \mathcal{N}_2(\theta, \zeta^2\mathcal{I})$), for finding the posterior density $f(\theta|y)$. The parameters L and ε were adjusted for the HMC while the variance ζ was adjusted for the MH algorithm, in order to reach appropriate acceptance rates ([0.23, 0.44] for MH, slightly higher for HMC) as well as a rapidly decreasing ACF plots and extensive state space explorations. A grid search over a set of different possible parameter values was made in order to do this. In figure 8 the results of both methods are plotted.

As we see in figure 10, the autocorrelation function from the Hamiltonian Monte Carlo method with our chosen parameters looks good, as it decreases rapidly.

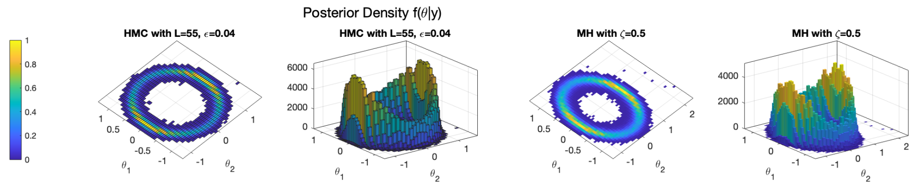

Figure 8: Posterior density simulated with HMC algorithm with $L = 55$ and $\epsilon = 0.04$ and MH algorithm where $\zeta = 0.5$, both with $N = 1000000$.

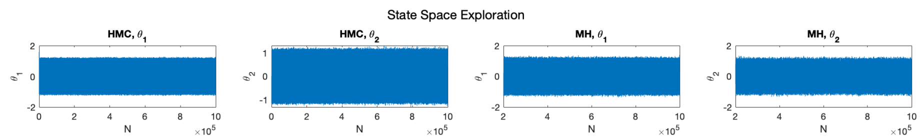

Figure 9: State Space Exploration of θ_1 and θ_2 with $N = 1000000$ and parameter values $L = 55$, $\epsilon = 0.04$ (HMC) and $\zeta = 0.5$ (MH).

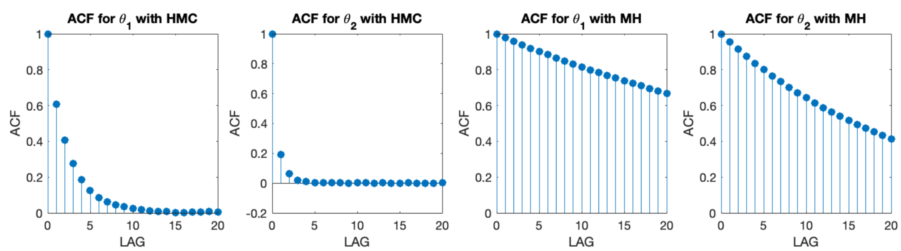

Figure 10: ACF functions for the two methods, with $N = 1000000$ and parameter values $L = 55$, $\epsilon = 0.04$ (HMC) and $\zeta = 0.5$ (MH).

We used the parameters $L = 55$ and $\epsilon = 0.04$, and got an an acceptance rate of 53%. If ϵ is too large we get the wrong distribution, and if it's too small, we get a slower decrease of the autocorrelation function. As we see in figure 9, the state space is well explored using these parameter values.

For the Metropolis-Hastings method however, the autocorrelation function decreases very slowly, which indicates that the proposal variance is too small. The decrease rate can be improved by increasing ζ, as we would get a proposal distribution that sets $\theta*$ further from the current state. However, this would lead to fewer proposals being accepted, so the acceptance rate would decrease drastically. In this trade-off between a good acceptance rate and a rapid decrease of ACF, we choose to set $\zeta = 0.5$, giving us an acceptance rate of 23.4%.

In conclusion, the HMC algorithm is better suited for this problem, as obtaining the same results using MH would require a much larger N for ACF convergence, which is computationally very expensive. However, both methods are successfull in finding the target distribution.