HA2 Report

Damian Sobecki

May 15, 2025

1 Bayesian analysis of coal mine disasters—constructing a complex MCMC algorithm

In Problem 1, we model coal mine disaster events $\boldsymbol{\tau}$ using an inhomogeneous Poisson process with intensities $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)$ over intervals defined by breakpoints $\boldsymbol{t} = (t_1, \ldots, t_{d+1})$.

Priors are $\theta \sim \Gamma(2, \vartheta)$, $\lambda_i \sim \Gamma(2, \theta)$, and $f(\boldsymbol{t}) \propto \prod_{i=1}^{d}(t_{i+1} - t_i)$.

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

1.1 Problem 1 (a): Computing marginal posteriors $f(\theta \mid \boldsymbol{\lambda}, \boldsymbol{t}, \boldsymbol{\tau})$, $f(\boldsymbol{\lambda} \mid \theta, \boldsymbol{t}, \boldsymbol{\tau})$, and $f(\boldsymbol{t} \mid \theta, \boldsymbol{\lambda}, \boldsymbol{\tau})$.

The joint posterior can be expressed as:

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

After substituting the given likelihood and priors:

$$f(\theta, \lambda, t \mid \tau) \propto \left[\exp\left(-\sum_{i=1}^{d} \lambda_i(t_{i+1} - t_i)\right) \prod_{i=1}^{d} \lambda_i^{n_i(\tau)}\right] \cdot \left[\theta^{2d} \prod_{i=1}^{d} \lambda_i e^{-\theta\lambda_i}\right] \cdot \left[\theta e^{-\vartheta\theta}\right] \cdot \left[\prod_{i=1}^{d}(t_{i+1} - t_i)\right].$$

From here we can derive the marginal posteriors by simplifying the expression, and then collecting only the terms that involve the respective variables.

In the end we obtain:

- $f(\theta \mid \boldsymbol{\lambda}, \boldsymbol{t}, \boldsymbol{\tau}) \propto \theta^{(2d+2)-1} e^{-\theta\left(\vartheta + \sum_{i=1}^{d} \lambda_i\right)} \sim \Gamma\left(2d + 2, \vartheta + \sum_{i=1}^{d} \lambda_i\right)$,
- $f(\boldsymbol{\lambda} \mid \theta, \boldsymbol{t}, \boldsymbol{\tau}) \propto \prod_{i=1}^{d} \lambda_i^{(n_i(\boldsymbol{\tau})+2)-1} e^{-\lambda_i(t_{i+1} - t_i + \theta)}$, with $\lambda_i \sim \Gamma(n_i(\boldsymbol{\tau}) + 2, t_{i+1} - t_i + \theta)$,
- $f(\boldsymbol{t} \mid \theta, \boldsymbol{\lambda}, \boldsymbol{\tau}) \propto \prod_{i=1}^{d}(t_{i+1} - t_i) e^{-\lambda_i(t_{i+1} - t_i)} \lambda_i^{n_i(\boldsymbol{\tau})}$, not a standard distribution.

1.2 Problem 1 (b): Constructing a hybrid MCMC algorithm.

We will use Gibbs sampling for variables that have conjugate priors (θ and $\boldsymbol{\lambda}$), and Metropolis-Hastings (MH) for variables that don't, in our case the breakpoints $\boldsymbol{t}$.

As the update for MH we pick *random walk proposal*.

Hybrid MCMC algorithm to sample from the posterior $f(\theta, \boldsymbol{\lambda}, \boldsymbol{t} \mid \boldsymbol{\tau})$

Initialize $\theta^{(0)}, \boldsymbol{\lambda}^{(0)}, \boldsymbol{t}^{(0)}$ randomly
Parameters: number of iterations N, tuning parameter $\rho > 0$

Loop for each iteration $k = 1$ **to** N**:**
// Gibbs sampling for θ
Sample $\theta^{(k)} \sim \Gamma\left(2d + 2, \vartheta + \sum_{i=1}^{d} \lambda_i^{(k-1)}\right)$
// Gibbs sampling for $\boldsymbol{\lambda}$
For $i = 1$ to d:
Sample $\lambda_i^{(k)} \sim \Gamma\left(n_i(\boldsymbol{\tau}) + 2, t_{i+1}^{(k-1)} - t_i^{(k-1)} + \theta^{(k)}\right)$
End
// Metropolis-Hastings for $\boldsymbol{t}$ *(random walk proposal)*
Pick a random breakpoint index i from $\{2, \ldots, d\}$
Propose $t_i^* = t_i^{(k-1)} + \epsilon$, where $\epsilon \sim \text{Unif}(-R, R)$, $R = \rho(t_{i+1}^{(k-1)} - t_{i-1}^{(k-1)})$
Compute acceptance ratio $\alpha = \min\left(1, \frac{f(\boldsymbol{\tau}|\boldsymbol{\lambda}^{(k)}, t^*) f(t^*)}{f(\boldsymbol{\tau}|\boldsymbol{\lambda}^{(k)}, t^{(k-1)}) f(t^{(k-1)})}\right)$
Set $t_i^{(k)} = t_i^*$ with probability α; **otherwise** $t_i^{(k)} = t_i^{(k-1)}$
Update other $t_j^{(k)} = t_j^{(k-1)}$ for $j \neq i$
End
Return samples $\{\theta^{(k)}, \boldsymbol{\lambda}^{(k)}, \boldsymbol{t}^{(k)}\}_{k=1}^{N}$

1.3 Problem 1 (c): Investigating the behaviour of the MCMC chain for 1, 2, 3, and 4 breakpoints.

We will investigate the behaviour of MCMC chain by analysing histograms for the suggested breakpoints setting.

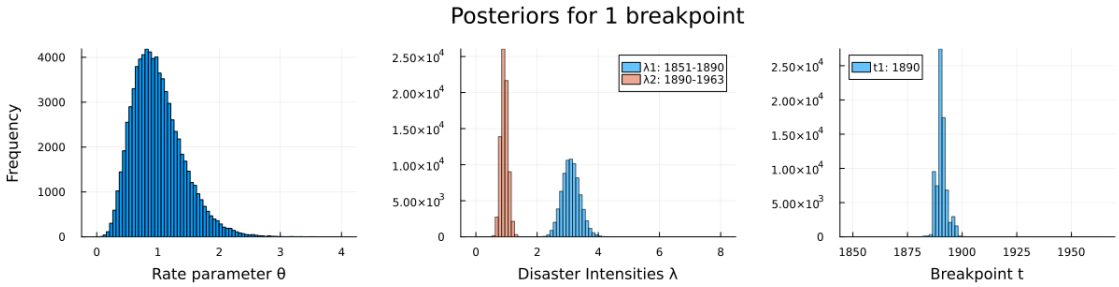

Figure 1: Histograms showing the posterior behaviour of hyperparameter θ, disaster intensities λ and breakpoints t ($\rho = 0.2$, $\vartheta = 2.0$, number of iterations = 77000).

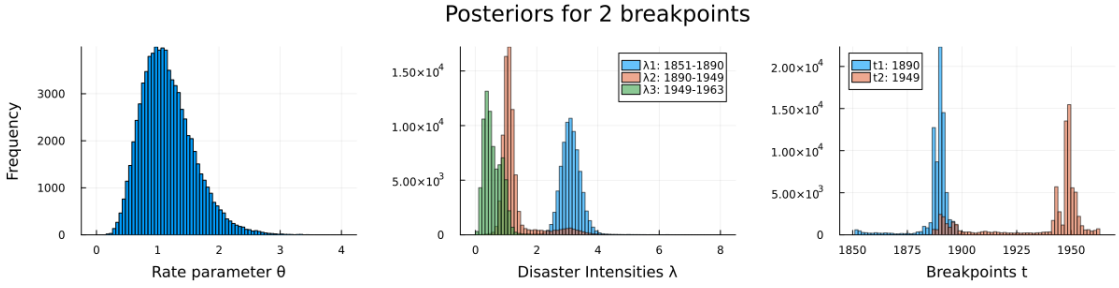

Figure 2: Histograms showing the posterior behaviour of hyperparameter θ, disaster intensities λ and breakpoints t ($\rho = 0.2$, $\vartheta = 2.0$, number of iterations = 77000).

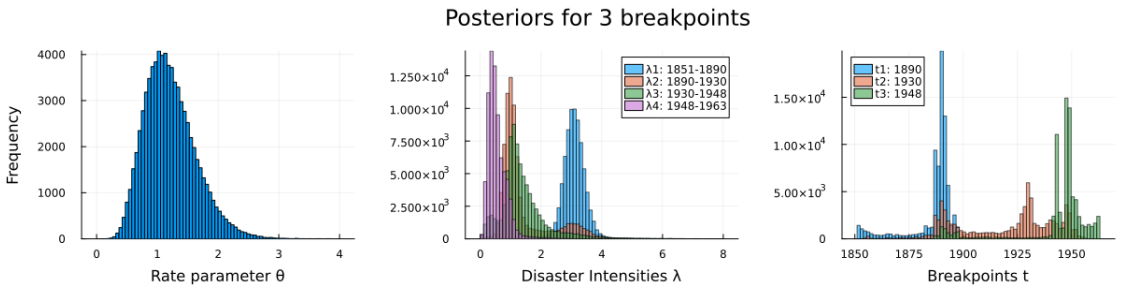

Figure 3: Histograms showing the posterior behaviour of hyperparameter θ, disaster intensities λ and breakpoints t ($\rho = 0.2$, $\vartheta = 2.0$, number of iterations = 77000).

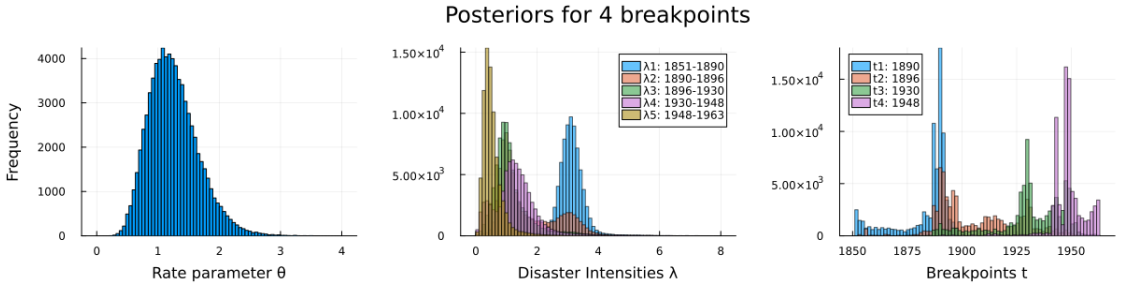

Figure 4: Histograms showing the posterior behaviour of hyperparameter θ, disaster intensities λ and breakpoints t ($\rho = 0.2$, $\vartheta = 2.0$, number of iterations = 77000).

From the figures, we can see that clear breakpoints can be set around years $t_1 \approx 1890$ and $t_2 \approx 1948$. This results in three intervals with distinct disaster rates per year cumulated around $\lambda_1 \approx 3.2$ before 1890, $\lambda_2 \approx 1.2$ between 1890-1948, and $\lambda_3 \approx 0.5$ per year after 1948.

With three breakpoints, the third is found around 1930. The disaster rate distributions for the intervals 1890-1930 and 1930-1948 overlap for the most part. This suggests that there was not much change in mine disasters between these intervals. Regardless, looking closer we might notice that, perhaps surprisingly, they got slightly worse in 1930-1950.

When trying to add a fourth breakpoint, it is hard to pinpoint it (probably around 1920), and the respective disaster rates are scattered across different values overlapping other distributions. This may suggest that it is too many and two or three breakpoints seem more suitable.

The results suggest 1890 was a big breakthrough in mining safety. All additional breakpoints further split the interval 1890-1963. This is an intuitive result, as we may have expected that the conditions would improve with time and innovation.

1.4 Problem 1 (d): How sensitive are the posteriors to the choice of the hyperparameter ϑ?

By testing a range of $\vartheta = \{1, 2, 5, 10, 20\}$ and visually comparing the obtained histograms (all other settings like in part 1c) we could see very little difference, if any, and conclude that the posteriors are not sensitive to the ϑ changes.

1.5 Problem 1 (e): How sensitive is the mixing and the posteriors to the choice of ρ in the proposal distribution?

Similarly for the step $\rho = \{0.1, 0.2, 0.5, 0.8, 0.9\}$, we do not notice much change and conclude the posteriors and mixing are not sensitive to the change of ρ.

2 Sampling from a circle-shaped posterior using Hamiltonian Monte Carlo

2.1 Problem 2 (a)

Compute the logarithm of the posterior $\log f(\boldsymbol{\theta}, \mathbf{y})$, up to a constant:

In general, the posterior is $f(\boldsymbol{\theta} \mid \mathbf{y}) \propto f(\mathbf{y} \mid \boldsymbol{\theta}) \cdot f(\boldsymbol{\theta})$. First we'll derive the likelihood, i.e., $f(\mathbf{y} \mid \boldsymbol{\theta})$, from the PDF of the normal distribution that our observations follow, take the logarithm and drop the terms constant with respect to $\boldsymbol{\theta}$, eventually obtaining:

$$\ln f(\mathbf{y} \mid \boldsymbol{\theta}) \propto -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - (\theta_1^2 + \theta_2^2)\right)^2$$

Second, the prior. It is a bivariate normal distribution $\boldsymbol{\theta} \sim \mathcal{N}_2(\mathbf{0}, \Sigma)$. The PDF of a multivariate normal distribution for a 2D vector $\boldsymbol{\theta}$ is:

$$f(\boldsymbol{\theta}) = \frac{1}{(2\pi)^{2/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\boldsymbol{\theta}^T\Sigma^{-1}\boldsymbol{\theta}\right) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\boldsymbol{\theta}^T\Sigma^{-1}\boldsymbol{\theta}\right),$$

where $|\Sigma|$ is the determinant of Σ, and $\boldsymbol{\theta}^T\Sigma^{-1}\boldsymbol{\theta} = [\theta_1, \theta_2]\Sigma^{-1}[\theta_1, \theta_2]^T$. Taking the logarithm:

$$\ln f(\boldsymbol{\theta}) = -\frac{1}{2}\ln(2\pi)^2 - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}\boldsymbol{\theta}^T\Sigma^{-1}\boldsymbol{\theta}.$$

The terms $-\ln(2\pi) - \frac{1}{2}\ln|\Sigma|$ are constant with respect to $\boldsymbol{\theta}$, so:

$$\ln f(\boldsymbol{\theta}) \propto -\frac{1}{2}\boldsymbol{\theta}^T\Sigma^{-1}\boldsymbol{\theta}.$$

Finally, the log-posterior is:

$$\ln f(\boldsymbol{\theta} \mid \mathbf{y}) \propto \ln f(\mathbf{y} \mid \boldsymbol{\theta}) + \ln f(\boldsymbol{\theta}) \propto -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - (\theta_1^2 + \theta_2^2)\right)^2 - \frac{1}{2}\boldsymbol{\theta}^T\Sigma^{-1}\boldsymbol{\theta}. \tag{1}$$

Compute its gradient $\nabla_\theta \log f(\boldsymbol{\theta}, \mathbf{y})$:

To do so, we need to differentiate the log-posterior with respect to θ_1 and θ_2, this can be expressed as:

$$\nabla_{\boldsymbol{\theta}} \ln f(\boldsymbol{\theta} \mid \mathbf{y}) = \left(\frac{\partial}{\partial\theta_1} \ln f(\boldsymbol{\theta} \mid \mathbf{y}),\ \frac{\partial}{\partial\theta_2} \ln f(\boldsymbol{\theta} \mid \mathbf{y})\right),$$

where $f(\boldsymbol{\theta} \mid \mathbf{y}) \propto -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - (\theta_1^2 + \theta_2^2)\right)^2 - \frac{1}{2}\boldsymbol{\theta}^T\Sigma^{-1}\boldsymbol{\theta}$ – from 1.

First, we'll treat the likelihood term $-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - (\theta_1^2 + \theta_2^2)\right)^2$.

We substitute $u = \theta_1 + \theta_2$, then the derivative of the inner expression is:

$$\frac{\partial}{\partial u}(y_i - u)^2 = -2(y_i - u),$$

and compute the partial derivatives:

- For θ_1, we note that $u = \theta_1^2 + \theta_2^2$, so $\frac{\partial u}{\partial\theta_1} = 2\theta_1$. Using the chain rule:

$$\frac{\partial}{\partial\theta_1}\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - u)^2\right] = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[-2(y_i - u)\cdot 2\theta_1\right] = \frac{2\theta_1}{\sigma^2}\sum_{i=1}^{n}(y_i - (\theta_1^2 + \theta_2^2)).$$

- For θ_2, since $\frac{\partial u}{\partial\theta_2} = 2\theta_2$:

$$\frac{\partial}{\partial\theta_2}\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - u)^2\right] = \frac{2\theta_2}{\sigma^2}\sum_{i=1}^{n}(y_i - (\theta_1^2 + \theta_2^2)).$$

Second, the prior term $\frac{1}{2}\boldsymbol{\theta}^T\Sigma^{-1}\boldsymbol{\theta}$.

The gradient of $\boldsymbol{\theta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}$ with respect to $\boldsymbol{\theta}$ is $\nabla_{\boldsymbol{\theta}}(\boldsymbol{\theta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}) = 2\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}$, since $\boldsymbol{\Sigma}^{-1}$ is symmetric.

Thus, $\nabla_{\boldsymbol{\theta}}\left(-\frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}\right) = -\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta} = -\boldsymbol{\Sigma}^{-1}\left[\theta_1, \theta_2\right]^{\mathrm{T}}$.

Let $\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$, then we have $-\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta} = -\begin{bmatrix} a\theta_1 + b\theta_2 \\ b\theta_1 + c\theta_2 \end{bmatrix}$.

Finally, the total gradient is:

$$\nabla_{\boldsymbol{\theta}} \ln f(\boldsymbol{\theta} \mid \mathbf{y}) = \begin{bmatrix} \frac{\partial}{\partial\theta_1} \ln f(\boldsymbol{\theta} \mid \mathbf{y}) \\ \frac{\partial}{\partial\theta_2} \ln f(\boldsymbol{\theta} \mid \mathbf{y}) \end{bmatrix} = \begin{bmatrix} \frac{2\theta_1}{\sigma^2}\sum_{i=1}^{n}(y_i - (\theta_1^2 + \theta_2^2)) - (a\theta_1 + b\theta_2) \\ \frac{2\theta_2}{\sigma^2}\sum_{i=1}^{n}(y_i - (\theta_1^2 + \theta_2^2)) - (b\theta_1 + c\theta_2) \end{bmatrix}. \tag{2}$$

2.2 Problem 2 (b)

HMC algorithm to sample from the posterior $f(\boldsymbol{\theta} \mid \mathbf{y})$

Input: Data $\mathbf{y}$, parameters σ, Σ, initial $\boldsymbol{\theta}^0$, step size ε, number of steps L, number of iterations N
Output: Samples $\{\boldsymbol{\theta}^0, \boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^N\}$

1. Initialize $\boldsymbol{\theta} = \boldsymbol{\theta}^0$
2. For $t = 1$ to N:
   (a) Sample momentum $\mathbf{p} \sim \mathrm{N}_2(\mathbf{0}, I)$
   (b) Set current state $(\boldsymbol{\theta}_{\text{current}}, \mathbf{p}_{\text{current}}) = (\boldsymbol{\theta}, \mathbf{p})$
   (c) // Leapfrog Integration for L Steps
   For $l = 1$ to L:
   $\mathbf{p} = \mathbf{p} + \frac{\varepsilon}{2}\nabla_{\boldsymbol{\theta}} \ln f(\boldsymbol{\theta} \mid \mathbf{y})$ // half-step update for momentum
   $\boldsymbol{\theta} = \boldsymbol{\theta} + \varepsilon\mathbf{p}$ // full-step update for position
   $\mathbf{p} = \mathbf{p} + \frac{\varepsilon}{2}\nabla_{\boldsymbol{\theta}} \ln f(\boldsymbol{\theta} \mid \mathbf{y})$ // the other half of the step update for momentum
   End For
   (d) Set proposed state $(\boldsymbol{\theta}^*, \mathbf{p}^*) = (\boldsymbol{\theta}, \mathbf{p})$
   (e) // Compute Hamiltonian
   $H_{\text{current}} = -\ln f(\boldsymbol{\theta}_{\text{current}} \mid \mathbf{y}) + \frac{1}{2}\mathbf{p}_{\text{current}}^T\mathbf{p}_{\text{current}}$
   $H_{\text{proposed}} = -\ln f(\boldsymbol{\theta}^* \mid \mathbf{y}) + \frac{1}{2}\mathbf{p}^{*T}\mathbf{p}^*$
   (f) // Metropolis-Hastings Acceptance
   $\alpha = \min\left(1, \exp(-H_{\text{proposed}} + H_{\text{current}})\right)$
   If random $u \sim \text{Uniform}(0, 1) < \alpha$:
   $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ // Accept proposal
   Else:
   $\boldsymbol{\theta} = \boldsymbol{\theta}_{\text{current}}$ // Reject and keep current state
   End If
   (g) Store $\boldsymbol{\theta}$ as sample $\boldsymbol{\theta}^t$
   End For
3. Return samples $\{\boldsymbol{\theta}^0, \boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^N\}$

2.3 Problem 2 (c)

How efficiently is the support of the target distribution explored in each case?

Figure 5 depicts heatmaps for the distributions obtained with both methods. We can see that posterior samples from HMC resemble the ground truth closely. While MH sampling produced a ring-shaped distribution as well, the desired pattern is less apparent.

Figure 6 on the other hand shows the autocorrelation function (ACF) for θ_1 and θ_2 for both HMC and MH, where "Lag" is the number of steps separating samples in the chain. It is apparent that the MH samples remain highly correlated even at a lag of 50, while HMC's correlation quickly vanishes and becomes negligible beyond a lag of ~ 10.

The acceptance rates are 91.3% for HMC, and 49.5% for MH. This suggests that the proposals in HMC are way closer to the target posterior distribution, compared to MH.

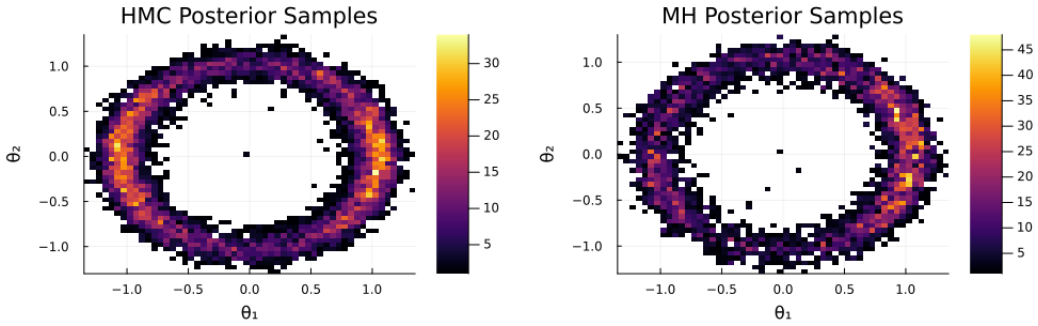

Figure 5: Heatmaps of the posterior density $f(\boldsymbol{\theta} \mid \mathbf{y})$ obtained with two different algorithms: HMC ($\varepsilon = 0.1$, $L = 10$) and MH ($\zeta = 0.2$).

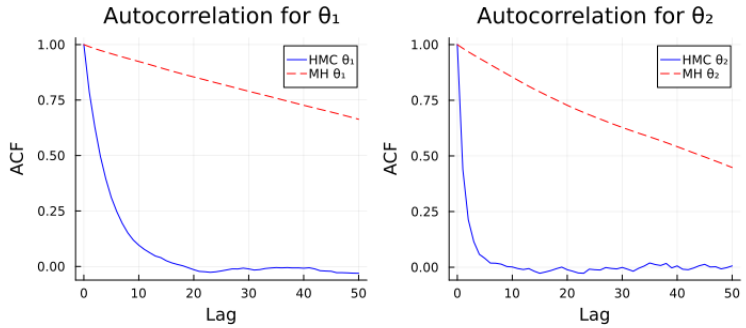

Figure 6: Autocorrelation functions of the samples for posterior density $f(\boldsymbol{\theta} \mid \mathbf{y})$ obtained with two different algorithms: HMC ($\varepsilon = 0.1$, $L = 10$) and MH ($\zeta = 0.2$).

All evidence indicates that HMC outperforms MH in exploring the target distribution's support. This outcome is expected, as the posterior's complex, ring-shaped geometry poses a challenge for MH's relatively simple random-walk proposals. Its small, uniform steps lead to slow exploration, poor mixing, and high autocorrelation. In contrast, HMC's gradient-guided proposals efficiently navigate the ring's narrow, curved support, achieving faster mixing and better coverage.

Investigate the performance of both algorithms varying the ε, L, ζ.

Let's first analyze MH. The standard deviation ζ controls the spread of the proposal distribution. A smaller ζ reduces the variance ζ^2 making $\boldsymbol{\theta}^*$ more likely to be close to $\boldsymbol{\theta}$. This means the proposed

exploration steps ($\boldsymbol{\theta}^* - \boldsymbol{\theta}$) are smaller in magnitude.

The tested values were $\zeta = \{0.1, 0.2, 0.5, 0.8\}$.

The performance of the algorithm is sensitive to ζ changes. Smaller ζ results in significantly improved heatmap coverage and acceptance rate, as well as higher autocorrelation. Considering the steps in MH are uniform, intuitively we can see that as overshooting the narrow target with too large steps, while smaller steps make it easier to not do that, but are slower at exploration. Ideally we would be able to make larger, informed steps - like in HMC.

In HMC, we test ε, which sets the size of each leapfrog step, and L, the number of leapfrog steps, which determines the trajectory length by controlling how many steps of size ε the particle takes before proposing a new sample $\boldsymbol{\theta}^*$. This gives us more flexibility in controlling the distance between the current and proposed samples ($\boldsymbol{\theta}^* - \boldsymbol{\theta}$).

The tested values were $\varepsilon = \{0.01, 0.1, 0.2, 0.5\}$, and $L = \{2, 5, 10, 20, 30, 50\}$.

Empirically, relatively small ε and large L combination results in steps that adapt to do distribution the best – we obtain good heatmap coverage and acceptance rate, and a low autocorrelation.

When the step ε is large, and L is small it's very easy to overshoot. Holding both values small or large results in behaviour similar to when modifying step size in MH.

To sum up, varying the parameters adjusts the step sizes in both HMC and MH: with large steps we risk overshooting, while with small steps we explore too slowly. HMC gives us possibility to combine many small steps into a single proposal. This results in adaptability that allows efficient navigation of a complex posterior like our ring-shaped distribution.