Mathematical Statistics **Home assignment 3: 2023–02–21**
Centre for Mathematical Sciences **MASM11/FMSN50 Monte Carlo and Empirical Methods**
Lund University **for Stochastic Inference 7.5 hp**

Detailed solutions, making sure that you describe your algorithms in text, and not only as code, must be submitted **before Tuesday Mar 7, 13:00:00**. You are strongly encouraged to work in groups of two.

Report submissions are accepted in PDF format only.

Also submit an email with your MATLAB-files (or implementation in other language), with a file named `proj3.m` that can be used to run your analysis (Remember to submit **all** of the files you use to the create your solution). Submit the projects in CANVAS.

Discussion between groups is permitted, as long as your report reflects your own work.

Coal mine disasters—constructing a complex MCMC algorithm

1. In the first problem we are going to take a closer look at coal mine disasters in Great Britain between 1658 and 1980. The data is taken from `http://www.cmhrc.co.uk/site/disasters/index.html` Since the dataset span over more than 300 years it is natural to assume that the conditions have changed over the years. Opening of new mines and closing of old mines, development of new technology and varying demand for coal are some of the factors that can cause a change in disaster intensity.

   First we need some notation. Let $t_1 = 1658$ and $t_{d+1} = 1980$ be the fixed end points of the dataset and denote by t_i, $i = 2, \ldots, d$, the breakpoints ($d = 1$ implies no breakpoints only one fixed intensity λ_1 over the whole data set). We collect end points and break points in a vector $\mathbf{t} = (t_1, \ldots, t_{d+1})$. The disaster intensity in each interval $[t_i, t_{i+1})$ is λ_i and we let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)$.

   The data consist of time continuous data where $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n)$ denotes the time points of the $n = 751$ disasters (available in the file `coal_mine_disasters.mat`). We model the data on the interval $t_1 \leq t \leq t_{d+1}$ using an inhomogeneous Poisson process with intensity

   $$\lambda(t) = \sum_{i=1}^{d} \lambda_i \mathbb{1}_{[t_i, t_{i+1})}(t).$$

   This implies that the times between accidents have an exponential distribution with intensity λ_i for accidents in interval i i.e. $[t_i, t_{i+1})$. The time between the last accident in interval i and t_{i+1} is an observation of a truncated exponential random variable. All we know here is that it is larger than the time span left in the interval.

   From the time points of the disasters we compute

   $$n_i(\boldsymbol{\tau}) = \text{number of disasters in the sub-interval } [t_i, t_{i+1}) = \sum_{j=1}^{n} \mathbb{1}_{[t_i, t_{i+1})}(\tau_j).$$

   We put a Gamma$(2, \theta)$-prior[1] on the intensities with a Gamma$(2, \Psi)$-hyperprior on θ, where Ψ is a fixed hyperparameter that needs to be specified. In addition, we put a prior

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

[1] Here we use a parametrization of the gamma distribution where Gamma(k, θ), $k > 0$ and $\theta > 0$, has density function

$$f(x) = \frac{\theta^k}{\Gamma(k)} x^{k-1} \exp(-\theta x), \quad x \geq 0.$$

MATLAB uses a different parametrisation for the second argument which then should be set to $1/\theta$.

on the breakpoints, preventing the same from being located too closely. This implies that

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

To sample from the posterior $f(\theta, \boldsymbol{\lambda}, \mathbf{t}|\boldsymbol{\tau})$ we will construct a hybrid MCMC algorithm as follows. All components except the breakpoints $\mathbf{t}$ can be updated using Gibbs sampling. To update the breakpoints we use a Metropolis-Hastings (MH) sampler. **Note that all proposals which change the order of the breakpoints should have zero acceptance probability due to the assumption of ordered points in the model setup.** There are are several possible proposal distributions for the MH step. We will look at at two slightly different approaches:

- *Random walk proposal one at a time:* Update one breakpoint at a time. For each breakpoint t_i we generate a candidate t_i^* according to

  $$t_i^* = t_i + \varepsilon, \quad \text{with} \quad \varepsilon \sim \mathcal{U}(-R, R),$$

  and $R = \rho(t_{i+1} - t_{i-1})$.
- *Random walk proposal all at once:* The random walk proposal where we suggest a new position for all the breakpoints at once. Each of the points is now proposed as

  $$t_i^* = t_i + \varepsilon, \quad \text{with} \quad \varepsilon \sim \mathcal{U}(-\rho, \rho),$$

  but we either accept the entire vector and move the breakpoints or reject the move and stay at the same breakpoints.

In both cases ρ is a tuning parameter of the proposal distributions. It also possible to try to use a different ρ for each breakpoint.

(a) Compute, up to normalizing constants, the marginal posteriors $f(\theta|\boldsymbol{\lambda}, \mathbf{t}, \boldsymbol{\tau})$, $f(\boldsymbol{\lambda}|\theta, \mathbf{t}, \boldsymbol{\tau})$, and $f(\mathbf{t}|\theta, \boldsymbol{\lambda}, \boldsymbol{\tau})$. In addition, try to identify the distributions.

(b) Construct a hybrid MCMC algorithm that samples from the posterior $f(\theta, \boldsymbol{\lambda}, \mathbf{t}|\boldsymbol{\tau})$. Pick *one* of the possible updating options for $\mathbf{t}$; *motivate* why the proposal yields a correct MH algorithm.

(c) Investigate the behavior of the chain for 1, 2, 3, 4 or perhaps even more breakpoints.

(d) How sensitive are the posteriors to the choice of the hyperparameter Ψ?

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

Parametric bootstrap for the 100-year Atlantic wave

2. The data file `atlantic.txt` contains the significant wave-height recorded 14 times a month during several winter months in the north Atlantic. A *Gumbel 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 $\beta > 0$, is a good fit to the data. The parameters can be estimated using the matlab function `est_gumbel.m`.

   The expected 100-year return value of the significant wave-height gives the largest expected value during a 100-year period. The Tth return value is given by $F^{-1}(1-1/T; \mu, \beta)$. We note that we have 14 observations during a month and three winter months during a year, thus $T = 3 \cdot 14 \cdot 100$.

   (a) Find the inverse $F^{-1}(u; \mu, \beta)$.

   (b) Provide a parametric bootstrapped 95% confidence intervals for the parameters.

   (c) Provide a one sided upper bounded parametrically bootstrapped 95% confidence interval for the 100-year return value.

Good luck!