Avd. Matematisk statistik

KTH Matematik

HOME ASSIGNMENT 2, SF2955 COMPUTER INTENSIVE METHODS IN MATHEMATICAL STATISTICS

Teacher: Johan Westerborn
All MATLAB-files needed are available through the course home page.
The following is to be submitted:

- An email containing two report files (see below), one with names and one anonymous as well as *all* your `m`-files with a file named `group_number_HA2_matlab.m` that runs your analysis, or similar depending on your language of choice. This email has to be sent to `johawes@kth.se` by **Thursday 18 May, 13:00:00**.
- A report, named `group number-HA2-report.pdf`, of maximum 7 pages in `pdf` format with names of group members, a version of this report named `HA2-report.pdf` without names should also be included. The report should provide detailed solutions to all problems. The presentation should be self-contained and understandable without access to the code. **One** printed and stitched copy of the report (with name of group members) is brought to the lecture on Thursday 18 May.

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

Statistical inference from coal mine disaster and Atlantic wave data using Markov chain Monte Carlo and the bootstrap

4 maj 2017

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

In this problem we will generalise the coal mining example in the book (See Chapter 11.2.1) from one breakpoint to $d-1$ breakpoints. First we need some notation. Let $t_1 = 1851$ and $t_{d+1} = 1963$ be the fixed end points of the dataset and denote by t_i, $i = 2, \ldots, t_d$, the breakpoints. We collect end points and break points in a vector $\boldsymbol{t} = (t_1, \ldots, t_{d+1})$. The disaster intensity in each interval $[t_i, t_{i+1})$ is λ_i and we let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)$.
Another difference from the example in the book is that instead of calculating the number of disasters each year we will use time continuous data where $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n)$ denotes the time points of the $n = 191$ disasters (available in the file `coal_mine.txt`). We model the data on the interval $t_1 \leq t \leq t_{d+1}$ using an inhomogeneous Poisson process with intensity

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

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 on the intensities with a $\Gamma(2, \vartheta)$ hyperprior on θ, where ϑ is a fixed hyperparameter that needs to be specified. In addition, we put a prior

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

on the breakpoints, preventing the same from being located too closely. Using theory of Poisson processes, it can be shown that

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

To sample from the posterior $f(\theta, \boldsymbol{\lambda}, \boldsymbol{t} \mid \boldsymbol{\tau})$ we will construct a hybrid MCMC algorithm as follows. All components except the breakpoints $\boldsymbol{t}$ can be updated using Gibbs sampling. To update the breakpoints we use a Metropolis-Hastings step. There are several possible proposal distributions for the MH step:

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

  $$t_i^* = t_i + \epsilon, \quad \text{with} \quad \epsilon \sim \text{Unif}(-R, R)$$

  and $R = \rho(t_{i+1} - t_{i-1})$.

- *Independent proposal:* update one breakpoint at a time. For each breakpoint t_i we generate a candidate t_i^* according to

  $$t_i^* = t_{i-1} + \varepsilon(t_{i+1} - t_{i-1}), \quad \text{with} \quad \varepsilon \sim \text{Beta}(\rho, \rho).$$

This corresponds to a scaled and shifted beta-distribution for t_i^* with density function

$$f(t_i|t_{i+1}, t_{i-1}) = \frac{\Gamma(2\rho)}{\Gamma(\rho)^2} \frac{(t_i - t_{i-1})^{\rho-1} (t_{i+1} - t_i)^{\rho-1}}{(t_{i+1} - t_{i-1})^{2\rho-1}}.$$

In both cases ρ is a tuning parameter of the proposal distributions.

Problem 1

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

(b) Construct a hybrid MCMC algorithm that samples from the posterior $f(\theta, \boldsymbol{\lambda}, \boldsymbol{t} \mid \boldsymbol{\tau})$. Pick *one* of the possible updating options for $\boldsymbol{t}$.

(c) Investigate the behavior of the MCMC chain for 1, 2, 3, and 4 breakpoints.

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

(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

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 $T = 3 \cdot 14 \cdot 100$.

Problem 2

(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 parametric bootstrapped 95% confidence interval for the 100-year return value.

Good luck!