Avd. Matematisk statistik

KTH Matematik

HOME ASSIGNMENT 2, SF2955 COMPUTER INTENSIVE METHODS IN MATHEMATICAL STATISTICS

Examiner: Jimmy Olsson
All MATLAB-files needed are available through the course home page.
The following is to be submitted in Canvas by **Thursday 20 May, 12:00**:

- A report, named `group number-HA2-report.pdf`, of **maximum 7 pages** in `pdf` format. The report should provide detailed solutions to all problems. The presentation should be self-contained and understandable without access to the code.
- *All* your `m`-files (or similar depending on your language of choice) along with a file named `group number-HA2-matlab.m` that runs your analysis.

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

Statistical inference from coal mine disaster and mixture model data using Markov chain Monte Carlo and the EM-algorithm

5 maj 2021

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.csv`). 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?

EM-based inference in mixture models

A *mixture model* comprises an unobservable, $\{0, 1\}$-valued random variable X (referred to as *index*) such that $\mathbb{P}(X = 1) = 1 - \mathbb{P}(X = 0) = \theta$ and an observable random variable Y, taking on values in some possibly continuous space, such that

$$Y \mid X = 0 \sim g_0(y)\, dy,$$
$$Y \mid X = 1 \sim g_1(y)\, dy,$$

where g_i, $i \in \{0, 1\}$, are known probability densities. We will consider the special case where g_0 and g_1 are Gaussian with known means $(\mu_0, \mu_1) = (0, 1)$ and standard deviations $(\sigma_0, \sigma_1) = (1, 2)$, respectively. The probability θ is however unknown. We are given a set of independent observations $\mathbf{y} = (y_1, \ldots, y_n)$ from the model and denote by $\mathbf{x} = (x_1, \ldots, x_n)$ the corresponding unobserved index variables. Our aim is to infer this parameter by means of the frequentist approach. More specifically, since the model comprises missing data, we compute the maximum likelihood estimator of θ using the EM algorithm.

Problem 2

(a) Write, up to additive constants not depending on θ, the complete data log-likelihood function $\theta \mapsto \log f_\theta(\mathbf{x}, \mathbf{y})$.

(b) Determine the conditional distribution $f_\theta(\mathbf{x} \mid \mathbf{y})$.

(c) The file `mixture-observations.csv`, available on the home page, contains $n = 1000$ observations $\mathbf{y}$ from the model. Inspect the data by plotting a histogram. Using (a) and (b), derive the EM-updating formula for θ based on these observations, and run the EM algorithm for some suitable initial guess of θ. Report the EM learning curve as well as the numerical value of your final estimate of θ.

Good luck!