Assignment 3: Markov Chain Monte Carlo

Bayesian Inference and Computation

1 Introduction

On the module's Canvas page you will find a comparative judgement dataset (csv file). Your assignment is to fit a model to data using a Bayesian approach. This will involve writing an MCMC algorithm.

2 Purpose

As well as developing your understanding of Chapters 2 and 5 of the lecture notes, the coursework is designed to give you an opportunity to

- Analyse data using Bayesian methods.
- Express your mathematical understanding in code.
- Develop your sense of what good, clear code looks like.

3 Problem

Female genital mutilation (FGM) is a procedure where the female genitals are deliberately cut, injured or changed, but there's no medical reason for this to be done. It is illegal in the UK and a form of child abuse. South Yorkshire Police wanted to find out where girls are most at risk of FGM in Sheffield. They ran a comparative judgement study, where safeguarding professionals (social services, health workers, etc.) were shown pairs of wards in Sheffield and asked which of the ward had a higher risk of FGM. Fit a model to estimate the risk of FGM in each ward.

4 Data

The data (file `win_matrix.csv`) is in the form of a win/loss matrix. The matrix has 28 rows and columns and element (i, j) contains the number of times ward i was judged to have a lower risk than ward j. The file `wards.csv` contains the numeric encoding of the wards.

5 Model

The wards are labelled $1, \dots, N = 28$, and subward i is assigned a value $\lambda_i \in \mathbb{R}$. Larger values are associated with lower risk wards, smaller (potentially negative) values with higher risk wards.

When comparing wards i and j, the probability $\pi_{i,j}$ that ward i is judged to be lower risk than ward j depends on the difference in their parameters

$$\text{logit}(\pi_{ij}) = \lambda_i - \lambda_j \iff \pi_{ij} = \frac{\exp(\lambda_i)}{\exp(\lambda_i) + \exp(\lambda_j)} \qquad (i \neq j, 1 \leq i, j \leq N).$$

Under the assumption that each comparison is independent, if subwards i and j are compared n_{ij} times, then the number of times i is chosen to be less deprived than j is

$$Y_{ij} \sim \text{Bin}(n_{ij},\, \pi_{ij}).$$

Due to identifiability constraints, λ_1 (city ward) is set to 0, so that the magnitude of all other λ_i are relative to this fixed value. The likelihood function is then given as the product of the density functions for each pair of subwards (for $i < j$)

$$\pi(\boldsymbol{y} \mid \boldsymbol{\lambda}) \propto \prod_{i=1}^{N} \prod_{j=1}^{N} \pi_{ij}^{y_{ij}}$$

where $\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_N\} = \{0, \lambda_2, \ldots, \lambda_N\}$ is the set of ward risk parameters. Note that do derive the likelihood we used that $n_{i,j} = y_{i,j} + y_{j,i}$, $y_{i,i}$ is set to 0, and proportionality is understood with respect to $\boldsymbol{\lambda}$.

6 Task

Using independent $N(0, \sigma^2)$ prior distributions for $\lambda_2, \ldots, \lambda_N$ derive the posterior distribution $\pi(\boldsymbol{\lambda} \mid \boldsymbol{y})$. Letting $\sigma = 10$, construct an MCMC algorithm with appropriate hyperparameter choices to generate samples from the posterior distribution for $(\lambda_2, \ldots, \lambda_{28})$. Write an `R` script to run the MCMC algorithm and use it to generate samples from the posterior distribution. Interpret your results statistically.

7 Submission

You should submit two files:

1. A report of no more than four pages deriving the posterior distribution and acceptance probabilities, displaying your MCMC algorithm (also discussing your choice of proposal distribution and burn-in periods) and summarising your results.
2. An `R` script that implements your MCMC algorithm.

The coursework will be designed to be completed with roughly 4–5 hours' work. You should also make sure you are familiar with the school/university regulations and guidelines on **plagiarism**.

8 Assessment

14 marks are available for the theoretical statistical and data analysis element of the task. Three marks are available for coding style and presentation. Three marks are available for the presentation and writing in the report.

Marks	Requirements
14	Completes the task in full with no errors
9 - 13	Completes the tasks in full but contains minor errors
5 - 8	Makes some progress towards completing the task but is incomplete or contains serious errors.
0 - 4	Makes little or no progress towards completing the task.

Table 1: Mark scheme for the theoretical statistical and data analysis element of the task

Marks	Requirements
3	Code is fully commented, with suitably and consistently named variables. Any plots are fully labelled.
1-2	Code is mostly commented and coding style is mostly consistent, but some parts are either not commented or consistently named
0	Code contains little or no comments and coding style is incoherent.

Table 2: Mark scheme for the coding style and presentation element

Marks	Requirements
3	The report is well presented and in a logical format. The report is written to a high standard of scientific english. All relevant parts of the data analysis are presented
1-2	The report is mostly well presented and in a logical format. But the report may contain spelling or grammatical errors, or parts of the data analysis are missing
0	The report is poorly presented, has incoherent reasoning, or does not describe the data analysis carried out.

Table 3: Mark scheme for the report style and presentation element