Statistical Methods for Data Science

DATA7202

Semester 1, 2023

Assignment 3 (Weight: 25%)

Assignment 3 is due on 18 May 23 16:00).

Please answer the questions below. For theoretical questions, you should present rigorous proofs and appropriate explanations. Your report should be visually appealing and all questions should be answered in the order of their appearance. For programming questions, you should present your analysis of data using Python, Matlab, or R, as a short report, clearly answering the objectives and justifying the modeling (and hence statistical analysis) choices you make, as well as discussing your conclusions. Do not include excessive amounts of output in your reports. All the code should be copied into the appendix and the sources should be packaged separately and submitted on the blackboard in a zipped folder with the name:

```
"student_last_name.student_first_name.student_id.zip".
```

For example, suppose that the student name is John Smith and the student ID is 123456789. Then, the zipped file name will be `John.Smith.123456789.zip`.

1. **[20 Marks]** Conjugate Uniform random variable analysis: Consider n iid Uniform random variables Y_i, $(i = 1, \ldots, n)$, each with p.d.f.

$$\mathbb{P}(y \mid \theta) = \frac{1}{\theta}.$$

Suppose that the prior for θ is the Pareto distribution $\mathsf{Pareto}(\alpha, x_m)$. Namely

$$p(x \mid \alpha, x_m) = \begin{cases} \frac{\alpha x_m^\alpha}{x^{\alpha+1}} & x \geqslant x_m \\ 0 & x < x_m. \end{cases}.$$

Derive the posterior distribution of θ.

2. **[20 Marks]** Conjugate Categorical random variable analysis: Consider n iid categorical random variables Y_i, $(i = 1, \ldots, n)$, each with p.d.f.

$$\mathbb{P}(y \mid p_1, \ldots, p_k) = \prod_{j=1}^{k} p_j^{1_{\{y=j\}}}$$

Suppose that the prior for $\boldsymbol{\theta} = \{p_1, \ldots, p_k\}$ is the Dirichlet distribution $\mathsf{Dirichlet}(\alpha_0^{(1)}, \ldots, \alpha_0^{(k)})$. Namely

$$p(p_1, \ldots, p_k \mid \alpha_0^{(1)}, \ldots, \alpha_0^{(k)}) \propto \prod_{j=1}^{k} p_j^{\alpha_0^{(j)}-1}$$

Derive the posterior distribution of $\boldsymbol{\theta}$.

3. **[25 Marks (see details below)]** Consider a sampling from the 2-dimensional pdf

$$f(x, y) = c\,\mathrm{e}^{-(xy+x+y)}, \quad x \geqslant 0, \quad y \geqslant 0,$$

for some normalization constant c, using a Gibbs sampler. Let $(X, Y) \sim f$.

(a) **[10 Marks]** Find the conditional pdf of X given $Y = y$, and the conditional pdf of Y given $X = x$.

(b) **[15 Marks]** Write working code that implements the Gibbs sampler and outputs 1000 points that are approximately distributed according to f.

4. **[45 Marks (see details below)]** Consider a survey data from J groups (such as locations for example), where respondents were asked to reply YES or NO for a specific question (such as if you support a certain candidate or a specific policy). Let the number of responders in group $1 \leqslant j \leqslant J$ be n_j. Then,

$$y_{ji} = \begin{cases} 1 & \text{respondent } i \text{ in group } j \text{ answered YES} \\ 0 & \text{oherwise.} \end{cases}$$

We now define $r_j = \sum_{i=1}^{n_j} y_{ji}$ to be the number of YES answers in group j. The data can now be modeled via:

$$r_j \sim \mathsf{Bin}(\theta_j, n_j), \text{ for } 1 \leqslant j \leqslant J,$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ are the parameter(s) of interest. We can now consider two models.

Model 1

$r_j \sim \mathsf{Bin}(\theta_j, n_j)$

$\theta_j \sim \mathsf{Beta}(\alpha, \beta)$.

Model 2

$r_j \sim \mathsf{Bin}(\theta_j, n_j)$

$\ln\left(\frac{\theta_j}{1-\theta_j}\right) \sim \mathsf{N}(\beta_0 + \beta_1 d_j, \sigma^2)$.

In Model 2, d_j stands for some scalar that might indicate a sentiment towards a question of interest in a group. For example, if we survey say a question of a negative attitude towards a placement of hydrogen facilities, d_j can stand for a "distance" from group (location) j to these facilities.

For Model 1 (M_1), α and β are hyper parameters; for Model 2 (M_2), the hyper parameters are β_0, β_1, and σ^2. So, prior densities of hyper parameters will have to be specified for the complete model description.

For Model 1, let $\alpha, \beta \sim \mathsf{U}(0, 100)$. For Model 2, define $\beta_0 \sim \mathsf{U}(-10, 10)$, $\beta_1 \sim \mathsf{U}(-10, 10)$, $\sigma^2 \sim \mathsf{U}(0, 100)$.

The obtained data from 7 districts is as follows.

district id	# of answers	distance	YES answers
1	100	7	10
2	200	6	40
3	300	5	90
4	400	4	160
5	300	3	150
6	200	2	120
7	100	1	70

(a) **[15 Marks]** For Model 1, implement an appropriate MCMC algorithm (you can use JAGS or similar software). You will need to produce 3 independent MCMC chains.

(b) **[15 Marks]** For Model 2, implement an appropriate MCMC algorithm (you can use JAGS or similar software). You will need to produce 3 independent MCMC chains.

(c) **[5 Marks]** Using the following `R` code, report the inference results.

```
library(coda)
library(MCMCvis)

m1 <- mcmc(read.csv("mcmc1.csv"))
m2 <- mcmc(read.csv("mcmc2.csv"))
m3 <- mcmc(read.csv("mcmc3.csv"))

mlist <- mcmc.list(m1,m2,m3)


MCMCdiag(mlist)
MCMCsummary(mlist)
MCMCtrace(mlist)
MCMCplot(mlist)
```

(d) **[10 Marks]** In Model 2, explain the effect of β_1.