Project Bayesian Data Analysis

Oswaldo Gressani & Christel Faes

2023-2024

INSTRUCTIONS

- Make a report in question-answer style
- Describe the flow of your procedures and the reasoning behind them.
- Limit your report to 30 pages (excluding title page, but including tables and figures)
- **Deadline of report is 10/01/2024**
- Send also your programs so that we can check that the programs really work!
- Submit report and R program via Blackboard

1 PART 1

The Gibbs algorithm

An experiment is conducted in a laboratory to assess the strength of infection of $n = 8$ different artificial viruses on microorganisms. The number of infected organisms (y_i) during an exposure period (t_i) to a virus (v_i) is given below.

Virus (v_i)	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8
Infections (y_i)	4	1	5	14	3	19	7	6
Exposure period (t_i)	95	16	63	126	6	32	16	19

Consider that the infections generated by the ith virus follow a Poisson process with parameter λ_i, $i \in \{1, \ldots, 8\}$, so that for an exposure time t_i, the number of infections y_i is Poisson distributed $y_i \sim \mathcal{P}(\lambda_i t_i)$ with mean $\mathbb{E}(y_i) = \lambda_i t_i$. In addition, assume the following Gamma prior distributions $\lambda_i \sim_{i.i.d.} \mathcal{G}(\alpha, \beta)$ for $i = 1, \ldots, n$ with $\alpha = 1.8$ and $\beta \sim \mathcal{G}(\gamma, \delta)$ with $\gamma = 0.01$ and $\delta = 1$, where $\mathcal{G}$ denotes a gamma distribution with the shape-rate parameterization.

1. Using Bayes' theorem compute the joint posterior distribution $p(\boldsymbol{\lambda}, \beta|\mathcal{D})$, where $\mathcal{D}$ denotes the observed data and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_8)^\top$.

2. Compute the following conditional posterior distributions $p(\lambda_i|\boldsymbol{\lambda}_{-i}, \beta, \mathcal{D})$ for $i = 1, \ldots, 8$ and $p(\beta|\boldsymbol{\lambda}, \mathcal{D})$. To what well-known parametric family do these conditional posterior distributions belong?

3. Based on the full conditionals obtained in the previous step, write a pseudo-code to obtain a random sample of total size 35 000 (including a burn-in of size 5 000) from $p(\boldsymbol{\lambda}, \beta|\mathcal{D})$ using the Gibbs sampler.

4. Using R and without relying on external software/packages (except the *coda* package), write the Gibbs sampler based on your pseudo-code in step 3. Please specify *set.seed(2025)* at the beginning of your code for reproducibility of your results. What is your acceptance rate?

5. Use the Geweke convergence diagnostic test and the Heidelberger-Welch stationarity test on the generated chains and interpret your results.

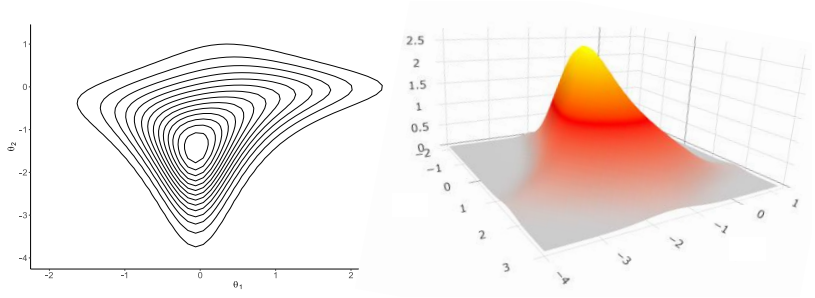
6. Based on the posterior mean of your generated chain, provide a point estimate and 95% quantile-based credible interval for $\mathbb{E}(y_6)$, i.e. the average number of infections generated by virus v_6 for $t_6 = 32$ minutes. (Round your results to the nearest integer).

7. Using your generated chain, estimate the probability $\mathbb{P}(\lambda_6 > 0.53)$.

The Metropolis algorithm

Let $\boldsymbol{\theta} = (\theta_1, \theta_2)^\top \in \mathbb{R}^2$ and assume the following posterior distribution (Gelman and Meng, 1991) with $A = 1.90$, $B = 0.54$, $C_1 = 0.50$ and $C_2 = -1.40$:

$$p(\boldsymbol{\theta}|\mathcal{D}) \propto \exp\left(-\frac{1}{2}(A\theta_1^2\theta_2^2 + \theta_1^2 + \theta_2^2 - 2B\theta_1\theta_2 - 2C_1\theta_1 - 2C_2\theta_2)\right).$$



1. Write the log posterior distribution $\log p(\boldsymbol{\theta}|\mathcal{D})$ and obtain analytically the gradient $\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\mathcal{D}) = (\partial \log p(\boldsymbol{\theta}|\mathcal{D})/\partial\theta_1, \partial \log p(\boldsymbol{\theta}|\mathcal{D})/\partial\theta_2)^\top$ and Hessian matrix:

$$\nabla^2_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\mathcal{D}) = \begin{pmatrix} \frac{\partial^2 \log p(\boldsymbol{\theta}|\mathcal{D})}{\partial\theta_1^2} & \frac{\partial^2 \log p(\boldsymbol{\theta}|\mathcal{D})}{\partial\theta_1\partial\theta_2} \\ \frac{\partial^2 \log p(\boldsymbol{\theta}|\mathcal{D})}{\partial\theta_2\partial\theta_1} & \frac{\partial^2 \log p(\boldsymbol{\theta}|\mathcal{D})}{\partial\theta_2^2} \end{pmatrix}.$$

2. Using the gradient and Hessian matrix obtained in the previous step, implement a Newton-Raphson algorithm in R to find the posterior mode of $p(\boldsymbol{\theta}|\mathcal{D})$ and denote by $\boldsymbol{\theta}^*_{NR} = (\theta_1^*, \theta_2^*)^\top$ the mode after convergence of the algorithm (Note: round $\boldsymbol{\theta}^*_{NR}$ to five digits after the decimal point).

3. Compute a Laplace approximation to $p(\boldsymbol{\theta}|\mathcal{D})$ around $\boldsymbol{\theta}^*_{NR}$. Report the covariance matrix Σ^* of the Laplace approximation (Note: round your results to five digits after the decimal point).

4. In R write a random-walk Metropolis algorithm to explore the joint posterior $p(\boldsymbol{\theta}|\mathcal{D})$ using a Gaussian proposal with covariance matrix $\widetilde{\Sigma} = c\Sigma^*$. Use a chain of length $M = 50\ 000$ and tune c to (approximately) reach the optimal acceptance rate of 23%. Please specify *set.seed(1993)* for the sake of reproducibility of your results.

5. Using your generated chains, estimate of the probability $\mathbb{P}\left((\theta_1/\theta_2) > 0.45\right)$.

2 PART 2

In this part, you have to use Bayesian software such as OpenBUGS, JAGS, R2OpenBUGS, R2jags or Nimble and any additional R packages such as CODA.

Data description

The German socioeconomic panel study data was taken from the first twelve annual waves (1984 through 1995) of the German Socioeconomic Panel (GSOEP) which surveys a representative sample of East and West German households. The data provide detailed information on the utilization of health care facilities, characteristics of current employment, and the insurance schemes under which individuals are covered. We consider a random sample of 100 individuals aged 25 through 65 from the West German subsample and of German nationality, which participated throughout all years.

The dataset contains the following variables:

- id: person - identification number
- female: female = 1; male = 0
- year: calendar year of the observation
- age: age in years
- hsat: health satisfaction, coded 0 (low) - 10 (high)
- handdum: handicapped = 1; otherwise = 0
- handper: degree of handicap in percent (0 - 100)
- hhninc: household nominal monthly net income in German marks / 1000
- hhkids: children under age 16 in the household = 1; otherwise = 0
- educ: years of schooling
- married: married = 1; otherwise = 0
- working: employed = 1; otherwise = 0
- docvis: number of doctor visits in last three months
- hospvis: number of hospital visits in last calendar year
- public: insured in public health insurance = 1; otherwise = 0

Research Question

We want to investigate whether there is a link between employment status and individual characteristics (age, gender, health status, etc), and the evolution of the employment status over time.

Specific Questions

1. Fit a logistic regression model and investigate the link between working status and the individual variables. Take vague priors for all model parameters. Select the most important variables explaining the working status of an individual using DIC and/or WAIC. Note: standardize variables to improve convergence of the MCMC procedure.

2. Fit a logistic mixed effects model with a random intercept and/or slope to the longitudinal profiles. Assume normality for the random effects. Take vague priors for all model parameters.

3. Check the posterior predictive performance to evaluate the model.

4. Check whether the logistic mixed effects model can be improved by changing the distribution of the random effects.

5. Perform a sensitivity analysis on the selected priors. Check what happens to the mixing of the chains and the posterior results.