University of Iceland

School of Engineering and Sciences

Faculty of Physical Sciences

Department of Mathematics

STÆ529M Bayesian Data Analysis

Fall 2023

Take-home Exam

This is an open-book exam.

Assigned: Friday November 17th 2023 at 16:00.

Due: Monday November 20th 2023 at 10:00.

There is a total of five problems, each problem weighs 20%. Within a problem each scenario weighs the same. Note that the number of scenarios is not the same between problems.

It is assumed that students have access to Matlab, R, S+, or Python, but of course other computer programs can be used. Programs written for Matlab, R, S+, Python or for other software should be should shown as a part of the solutions to the problems.

Students need to turn in their solutions in a pdf format (preferably one pdf file) to Canvas, both printed and handwritten material, before 10:00 on Monday, the 20th of November 2023, along with a statement signed by the student which states that the student worked alone on the solution of this take-home exam. If part of the solution is handwritten then you need to copy this part to a pdf file. Note that a mobile phone can be used for this purpose.

1. (20%)

Let Y_i denote the i-th observed count of pulses from a Geiger counter that was pointed toward a specimen of mineral over a period of t_i minutes, $i = 1, \ldots, n$, where n is the number of measurement periods. The purpose is to measure radiation. The intensity of the radiation is measured in pulses per minute.

Assume that the counts, Y_i, are independent and follow a Poisson distribution. The probability mass function of Y_i is

$$f(Y_i|\theta) = \frac{\exp(-t_i\theta)(t_i\theta)^{Y_i}}{Y_i!}, \quad Y_i \in \{0, 1, 2, \ldots\}, \quad i \in \{1, \ldots, n\},$$

where θ is an unknown parameter such that $\theta > 0$, representing the intensity of the radiation per minute. Assume that the prior density of θ is a gamma density with parameters α and β.

(a) Find the posterior distribution of θ (with the normalizing constant). Hint: conjugate distributions.

(b) The following data were observed.

i	1	2	3	4	5	6	7	8	9	10
t_i	2.9	2.0	2.2	3.4	2.1	2.4	1.9	1.8	1.7	2.7
Y_i	16	10	6	8	3	9	8	11	5	6

Based on these data, calculate the posterior mean, standard deviation and 95% interval of θ. Set the prior density such that $\alpha = 1$, and $\beta = 0.01$.

(c) Draw the posterior density of θ.

2. (20%)

The data set `problem2.txt` contains measurements on the annual maximum daily precipitation (mm per 24 hours, from 9:00 AM to 9:00 AM the next day) recorded at Reykjavík over the years 1926 to 2022. Reykjavík is located in the southwest part of Iceland (64°07.648' N, 21°54.166' W).

(a) Assume the data follow a normal distribution, that is,

$$Y_i \sim \text{Normal}(\mu, \sigma^2), \quad i = 1, \ldots, n,$$

where $\mu \in \mathbb{R}$, $\sigma^2 > 0$. Assign the following prior densities to μ and σ^2,

$$\mu \sim \text{Normal}(0, 100^2), \quad \sigma^2 \sim \text{InvGamma}(0.1, 0.1).$$

Sample from the posterior density of (μ, σ^2) by using a Gibbs sampler. Present the formulas for the Gibbs sampler. Write code for a Gibbs sampler that samples from the posterior density of (μ, σ^2) (do not use JAGS or other modeling libraries). Provide posterior summary for μ and σ^2, that is, present the posterior mean, the posterior standard deviation and the 95% equal-tailed posterior interval (based on 4 chains, each of length 13000 with 3000 for burn-in).

(b) Plot the theoretical cumulative density function (cdf) of the normal distribution using the posterior means of μ and σ^2. Plot the empirical cumulative density of the data on the same graph. Does the proposed normal model capture the shape of the empirical cumulative density?

(c) Draw the trace plots for μ, and σ^2, using the `plot` command in `rjags`. Do these plots indicate that the chains of the two parameters are mixing well?

(d) Use the `autocorr.plot` command in `rjags` to plot the autocorrelation function of each of the sampled parameters. Is the autocorrelation strong or weak? Do the autocorrelation functions indicate that the chains of the parameters will converge quickly or slowly?

(e) Compute the following diagnostic statistics: the lag 1 autocorrelation, the effective sample size, and the Gelman–Rubin statistic for each of the two parameters. Interpret these diagnostic statistics.

(f) Assume here that the data follow a lognormal distribution. This equivalent to assuming that $U_i = \log(Y_i)$ follows a normal distribution, that is,

$$U_i \sim \text{Normal}(\eta, \kappa^2), \quad i = 1, \ldots, n,$$

where $\eta \in \mathbb{R}$, $\kappa^2 > 0$. Assign the following prior densities to η and κ^2,

$$\eta \sim \text{Normal}(0, 100^2), \quad \kappa^2 \sim \text{InvGamma}(0.1, 0.1).$$

Sample from the posterior density of (η, κ^2) by using a Gibbs sampler (same formulas as in (a) but a different response variable). Provide posterior summary for η and κ^2, that is, present their posterior means, the posterior standard deviations and their 95% equal-tailed posterior intervals (based on 4 chains, each of length 13000 with 3000 for burn-in).

(g) Plot the theoretical cumulative density function (cdf) of the normal distribution using the posterior means of η and κ^2. Plot the empirical cumulative density of the log-transformed data on the same graph. Does the proposed normal model capture the shape of the empirical cumulative density of the log-transformed data?

(h) Based on the plots in (b) and (g), which of the two models appears to be better suited for the annual maximum daily precipitation data?

3. (20%)

Here we explore two methods to compare proportions of individuals with a particular trait in two populations, namely DIC and Bayes factor. Assume that we have independent samples from each of the populations, and the sample sizes are $n_1 = 1053$ and $n_2 = 976$, respectively. Let Y_j denote the number of individuals with the trait in the sample from population j, $j = 1, 2$. Assume that that samples are such that $Y_1 = 26$ and $Y_2 = 45$.

To compare the two proportions, we compare two models. Model 1, $\mathcal{M}_1$, is such that

$$Y_1 \sim \text{Bin}(n_1, \theta), \quad Y_2 \sim \text{Bin}(n_2, \theta), \quad \theta \sim \text{Beta}(1, 1).$$

Model 2, $\mathcal{M}_2$, is such that

$$Y_1 \sim \text{Bin}(n_1, \theta_1), \quad Y_2 \sim \text{Bin}(n_2, \theta_2), \quad \theta_1 \sim \text{Beta}(1, 1), \quad \theta_2 \sim \text{Beta}(1, 1).$$

(a) Use DIC to compare the two models. Use simulation to estimate DIC for each of the two models. In the case of the above data, does the difference in DIC indicate that Model 2 is a more suitable model?

(b) Use Bayes factor to compare the models. Assume a priori that the two models are equally likely. Here we use the result for Bayes factor directly. The Bayes factor of Model 2 relative to Model 1 is given by

$$BF = \frac{Y_1!(n_1 - Y_1)!Y_2!(n_2 - Y_2)!(n_1 + n_2 + 1)!}{(n_1 + 1)!(n_2 + 1)!(Y_1 + Y_2)!(n_1 + n_2 - Y_1 - Y_2)!}.$$

In the case of the above data, does the value of BF indicate that Model 2 is a more suitable model?

(c) Derive the formula for the Bayes factor given in (b).

4. (20%)

For one year, the consumption of petrol, Y (in millions of gallons) was measured in 48 states. The variables that may affect the consumption of petrol are; the petrol tax, X_2 (in cents per gallon); the per capita income, X_3 (in 1000 dollars per month); the number of miles of paved highway, X_4 (in 1000 miles); and the percentage of the population with driver's licenses, X_5. The file `problem4.txt` contains data of these variables. The columns contain X_2, X_3, X_4, X_5 and Y, respectively.

The following linear model is proposed

$$Y_i = \sum_{j=1}^{5} X_{ij}\beta_j + \epsilon_i, \quad i = 1, \ldots, 48,$$

with $X_{i1} = 1$ for all i. Assume that the Y's are independent, normally distributed and have equal variance, that is

$$Y_i|\beta_1, \ldots, \beta_5, \sigma^2 \sim \text{Normal}\left(\sum_{j=1}^{5} X_{ij}\beta_j, \sigma^2\right), \quad i = 1, \ldots, 48.$$

A priori, the β_js are independent of each other, and the prior distribution for each β_j is such that $\beta_j \sim \text{Normal}(0, 100^2)$. The prior distribution of σ^2 is such that $\sigma^2 \sim \text{InvGamma}(0.1, 0.1)$.

(a) Plot Y versus X_2, X_3, ..., x_5, a total of 4 figures. Which explanatory variables show a clear relationship with Y, and which don't?

(b) The per capita income, X_3, and the percentage of the population with driver's licenses, X_5, are known to have an effect on the consumption of petrol, Y. The effect of the petrol tax, X_2, and the number of miles of paved highway, X_4, is not as clear. Therefore we test 4 models for the mean, namely,

$$\text{Model 1}: \ \mu_i = \beta_1 + X_{ij}\beta_3 + X_{ij}\beta_5$$

$$\text{Model 2}: \ \mu_i = \beta_1 + X_{i2}\beta_2 + X_{i3}\beta_3 + X_{i5}\beta_5$$

$$\text{Model 3}: \ \mu_i = \beta_1 + X_{i3}\beta_3 + X_{i4}\beta_4 + X_{i5}\beta_5$$

$$\text{Model 4}: \ \mu_i = \beta_1 + X_{i2}\beta_2 + X_{i3}\beta_3 + X_{i4}\beta_4 + X_{i5}\beta_5$$

Create a table with these 4 models where the columns of the table show the models (Model 1 to Model 4), DIC of the models, and the effective number of parameters of the models. Based on the table, select one model for these data. This model will be used below. Sample from the posterior densities of these models using JAGS.

(c) Draw a normal probability plot of the residuals. Do the residuals appear to follow a normal distribution?
Hint: Create a vector of residuals, call it `resid` and use the commands `qqnorm(resid)` and `abline(0,sd(resid))`.

(d) Plot the residuals versus the predictions of the Y_is according to the model selected in (b). Also, plot the residuals versus each of the explanatory variables. Does the variance appear to be fixed when the residuals are plotted against these variables? Is is possible that the expected value of the residuals as a function of these variables is not equal to zero?

(e) Compute the posterior mean and 95% marginal posterior intervals for the parameters in the final model selected in (b), that is, the βs and σ^2.

(f) Interpret the parameters in the model, that is, explain the effect of each explanatory variable on the expected value of mean consumption of petrol by looking at the posterior mean of β_j when increasing the j-th explanatory variable by one unit while holding the other explanatory variables fixed.

(g) Based on the model found in (b), sample from the posterior predictive distribution (PPD) of the petrol consumption for a state that has the following values; $x_2 = 7$ cents per gallon, $x_3 = 3.3 \times 10^3$ dollars per month, $x_4 = 6.7 \times 10^3$ miles, $x_5 = 51$ %. Based on these samples, compute the mean of the posterior predictive distribution and its 95% equal-tailed prediction interval.

5. (20%)

The data in the file `problem5.txt` contain the height of 26 boys living in Oxford in England over a two year period. Each boy is measured on nine different occasions over a period of two years. The first column contains the identity number of each boy, the second column contains the time variable (in years, starts at -1.0, ends at a number close to 1.0) and the third column contains the height of the boys at the specified time point. The table below illustrates the form of the data.

id:	time	height
Boy 1:	-1.0000	140.52
Boy 1:	-0.7479	143.42
⋮	⋮	⋮
Boy 1:	0.9945	155.89
Boy 2:	-1.0000	136.91
⋮	⋮	⋮
Boy 2:	0.9945	148.39
⋮	⋮	⋮

Let Y_{ij} be the height of the i-th boy at time X_{ij} (the j-th time point for the i-th boy). It is assumed that the height increases linearly with time for each boy, however, the growth may vary between boys. Thus, the following model is proposed

$$Y_{ij} = \alpha_{i1} + \alpha_{i2} X_{ij} + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathrm{N}(0, \sigma^2),$$

$j = 1, \ldots, T$, $i = 1, \ldots, n$, where T is the number of measurements taken of each boy i ($T = 9$), and n is the number of boys ($n = 26$). Here ϵ_{ij} is the mean-zero deviation of the j-th measurement from the linear model for the i-th boy and σ^2 is its variance. It is assumed that the ϵ_{ij}s are independent of each other.

Each $\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2})^\mathsf{T}$ is assigned a bivariate Gaussian distribution with mean $\boldsymbol{\beta}$ and covariance matrix $\boldsymbol{\Omega}$, that is,

$$\boldsymbol{\alpha}_i \sim \mathrm{Normal}(\boldsymbol{\beta}, \boldsymbol{\Omega}), \quad i \in \{1, \ldots, n\}.$$

The following prior distributions are assigned to σ^2, $\boldsymbol{\beta}$, and $\boldsymbol{\Omega}$,

$$\sigma^2 \sim \text{InvGamma}(0.1, 0.1),$$
$$\boldsymbol{\beta} \sim \text{Normal}(\mathbf{0}, 100^2 \boldsymbol{I}_2),$$
$$\boldsymbol{\Omega} \sim \text{InvWishart}(2.1, \boldsymbol{I}_2/2.1).$$

(a) Find the posterior distribution of the unknown parameters in the hierarchical model. Present it in terms of the densities of the Y_{ij}s, the $\boldsymbol{\alpha}_i$s, σ^2, $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$.

(b) Specify the data layer, the process layer and the prior layer of the hierarchical model, that is, specify which variables/parameters are modeled at each layer, and what probability models are assumed at each level.

(c) Write code in JAGS within R to sample from the posterior density of the unknown parameters, and present the code. If another programming language than R (with JAGS) is used, please present the code that you use to sample from the posterior density of the unknown parameters.

(d) Compute the posterior means and 95% posterior intervals for α_{11}, ..., α_{n1}, α_{12}, ..., α_{n2}, σ^2, $\boldsymbol{\beta}$, and $\boldsymbol{\Omega}$. Use four chains where each chain consists of 13000 iterations and the first 3000 are used for burn-in. Present the α_{i1}s and the α_{i2}s in two separate figures, that is, plot the id-number of the boys on the x-axis and the 95% posterior intervals along with a point for the posterior mean on the y-axis. Present σ^2, $\boldsymbol{\beta}$, and $\boldsymbol{\Omega}$ in a table.

(e) For each of the 26 boys plot as dots the observed heights on the y-axis and corresponding times on the x-axis. This gives 26 figures. These figures can, for example, be arranged in a seven by four matrix (with two spots empty). In the case of the plot for the i-th boy, draw a line based on the formula $\alpha_{i1} + \alpha_{i1}x$ where x goes from -1.1 to 1.1. Use the posterior means of α_{i1} and α_{i1}. Does the model appear to fit the data adequately well?