ST4234: Bayesian Statistics

Take-home Midterm Exam, AY 19/20

Instructions:

1. This take-home midterm exam consists of **three** problems. The total marks of this take-home exam is **45**=14+13+18.

2. Copying solutions is **strictly prohibited**.

3. Please upload your work in **1 single pdf file** to the LumiNUS folder **"Midterm submission"**. The deadline for submission is **noon 12:00pm, Tuesday, 10 March, 2020**.

4. Please clearly write **your student number and your name** in the your submission.

5. If your student number is XXX, please name your file

   XXX.pdf

   For example, if your matriculation number is A0012345R, your file should be named **A0012345R.pdf**.

6. You must include your programming codes, outputs, and figures in your submitted file. You are allowed to scan your handwritings and combine files, but make sure that your final pdf file is compressed to no more than **5Mb**.

7. **No hard copy** will be accepted. **No late submission** will be accepted (i.e. marks for your midterm = zero).

1. (14 marks) Let $\boldsymbol{y} = \{y_1, \ldots, y_n\}$ be an i.i.d. sample from the Geometric(θ) distribution, with the pdf $p(y|\theta) = \theta(1-\theta)^y$ for $y = 0, 1, 2, \ldots$ and $\theta \in (0, 1)$. Recall that in a sequence of independent Bernoulli trials, a random variable from this Geometric(θ) counts the total number of failures before the first success occurs. The expectation of Geometric(θ) is $\frac{1-\theta}{\theta}$.

(a) (3 marks) Find a class of conjugate prior densities for θ. Find the posterior distribution under this conjugate prior.

(b) (3 marks) Find the prior mean and the posterior mean, using the conjugate prior in Part (a).

(c) (4 marks) Suppose that the prior of θ is Uniform$(0, 1)$. Find the posterior predictive density $p(y_{n+1}|\boldsymbol{y})$ for a new observation y_{n+1} from the same population.

(d) (4 marks) Find the Jeffreys prior for θ. Is it a proper prior?

2. (13 marks) Suppose that $\boldsymbol{x} = \{x_1, \ldots, x_m\}$ is an i.i.d. sample from $\mathrm{N}(\theta_1, \sigma^2)$. $\boldsymbol{y} = \{y_1, \ldots, y_n\}$ is an i.i.d. sample from $\mathrm{N}(\theta_2, \sigma^2)$. $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent. The unknown parameters are $\theta_1 \in (-\infty, \infty)$, $\theta_2 \in (-\infty, \infty)$, and $\sigma^2 \in (0, \infty)$.

(a) (5 marks) Under the improper prior $p(\theta_1, \theta_2, \sigma^2) \propto (\sigma^2)^{-2}$, find the joint posterior distribution of $(\theta_1, \theta_2, \sigma^2)$ by finding the following three distributions:

$$p(\theta_1|\sigma^2, \boldsymbol{x}, \boldsymbol{y}), \quad p(\theta_2|\sigma^2, \boldsymbol{x}, \boldsymbol{y}), \quad p(\sigma^2|\boldsymbol{x}, \boldsymbol{y}),$$

and show that

$$p(\theta_1, \theta_2, \sigma^2|\boldsymbol{x}, \boldsymbol{y}) \propto p(\theta_1|\sigma^2, \boldsymbol{x}, \boldsymbol{y}) \times p(\theta_2|\sigma^2, \boldsymbol{x}, \boldsymbol{y}) \times p(\sigma^2|\boldsymbol{x}, \boldsymbol{y}).$$

(b) (4 marks) Following Part (a), let $\delta = \theta_1 - \theta_2$. Find a suitable change of variable for δ, such that the marginal posterior of the transformed variable from δ has a Student's t distribution.

(Hint: You can first find the conditional posterior $p(\delta|\sigma^2, \boldsymbol{x}, \boldsymbol{y})$, and then calculate $p(\delta|\boldsymbol{x}, \boldsymbol{y}) = \int_0^\infty p(\delta|\sigma^2, \boldsymbol{x}, \boldsymbol{y}) \times p(\sigma^2|\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\sigma^2$.)

(c) (4 marks) Following Part (b), suppose that the following statistics are available:

$$m = 10, \quad \bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i = 1.656, \quad s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2 = 0.6750,$$
$$n = 20, \quad \bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j = 5.766, \quad s_y^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \bar{y})^2 = 0.9281.$$

Find the 90% HPD (highest posterior density) interval for σ^2. Then find the 90% HPD interval for δ.

(Hint: You may use the `qinvgamma` function in the `invgamma` package, and the `hpd` function in the `TeachingDemos` package.)

3. (18 marks) A cancer laboratory is estimating the rate of tumorigenesis in two strains of mice, A and B. They have tumor count data for 10 mice in strain A and 13 mice in strain B. Type A mice are well studied, and information from other laboratories suggests that type A mice have tumor counts that are approximately Poisson-distributed with a mean of 12. Tumor count rates for type B mice are unknown, but type B mice are related to type A mice. The observed tumor counts for the two populations are

$$\boldsymbol{y}_A = (12, 9, 12, 14, 13, 13, 15, 8, 15, 6),$$
$$\boldsymbol{y}_B = (11, 11, 10, 9, 9, 8, 7, 10, 6, 8, 8, 9, 7).$$

Assume independent Poisson sampling distributions for the tumor counts of each type of mice, with mean θ_A for type A and θ_B for type B. Consider the following prior distribution:

$$p(\theta_A, \theta_B) = p(\theta_A)p(\theta_B),$$
$$\theta_A \sim \text{Gamma}(120, 10), \quad \theta_B \sim \text{Gamma}(12 \times n_0, n_0)$$

where $n_0 > 0$.

(a) (3 marks) Find the posterior distribution $p(\theta_A, \theta_B | \boldsymbol{y}_A, \boldsymbol{y}_B)$ in terms of n_0.

(b) (4 marks) Find the posterior mean of θ_A. Find the posterior mean of θ_B in terms of n_0. For $n_0 \in \{1, 2, \ldots, 50\}$, plot the posterior mean of θ_B against n_0. On the same plot, add a horizontal line of the posterior mean of θ_A. Briefly comment on this plot about your findings. (Hint: You can use the function `abline(h=...)` to add a horizontal line at the height of `h`, after using the `plot` function.)

(c) (3 marks) For $n_0 \in \{1, 2, \ldots, 50\}$, obtain $\mathrm{P}(\theta_B < \theta_A | \boldsymbol{y}_A, \boldsymbol{y}_B)$ by Monte Carlo sampling of size $S = 10000$. Plot $\mathrm{P}(\theta_B < \theta_A | \boldsymbol{y}_A, \boldsymbol{y}_B)$ against n_0. How sensitive are the conclusions about the event $\{\theta_B < \theta_A\}$ to the prior distribution on θ_B?

(d) (3 marks) For $n_0 \in \{1, 2, \ldots, 50\}$, obtain $\mathrm{P}(\tilde{Y}_B < \tilde{Y}_A | \boldsymbol{y}_A, \boldsymbol{y}_B)$ by Monte Carlo sampling of size $S = 10000$, where $\tilde{Y}_A$ and $\tilde{Y}_B$ are samples from the posterior predictive distribution. Plot $\mathrm{P}(\tilde{Y}_B < \tilde{Y}_A | \boldsymbol{y}_A, \boldsymbol{y}_B)$ against n_0. How sensitive are the conclusions about the event $\{\tilde{Y}_B < \tilde{Y}_A\}$ to the prior distribution on θ_B? Does this plot display a similar pattern to the plot in Part (c)?

(e) (5 marks) Let's investigate the adequacy of the Poisson model for the data. Generate posterior predictive datasets $\boldsymbol{y}_A^{(1)} \ldots, \boldsymbol{y}_A^{(10000)}$, where each $\boldsymbol{y}_A^{(s)}$ is a sample of size $n_A = 10$ from the Poisson distribution with parameter $\theta_A^{(s)}$, where $\theta_A^{(s)}$ itself is a sample from the posterior distribution $p(\theta | \boldsymbol{y}_A)$. For each s, let $t^{(s)}$ be the sample average of the 10 values of $\boldsymbol{y}_A^{(s)}$ divided by the sample standard deviation of $\boldsymbol{y}_A^{(s)}$. Make a histogram of $t^{(s)}$ and compare to the observed value of this statistic. Based on this statistic, assess whether the fit of Poisson model is adequate for the data $\boldsymbol{y}_A$. With $n_0 = 50$ in the Gamma prior of θ_B, repeat the same procedure to assess whether the fit of Poisson model is adequate for the data $\boldsymbol{y}_B$. Briefly comment on your findings.