UNIVERSITY OF SURREY ©

B. Sc. Honours Courses in Mathematical Studies

Level HE3 Examination

Module MS331 BAYESIAN STATISTICS

Time allowed - 2 hours Autumn Semester 2004

Attempt THREE questions. If any candidate attempts more than THREE questions only the best THREE solutions will be taken into account.
A formula sheet and Cambridge Statistical Tables will be provided.

SEE NEXT PAGE

Question 1

(a) A random sample $t_1, \ldots t_n$ of lifetimes of batteries follows an exponential distribution with unknown mean $\theta^{-1} > 0$. Obtain the likelihood function of the sample and show that the gamma (a,b) distributions with densities

$$p(\theta) = \frac{b^a \theta^{a-1} e^{-b\theta}}{\Gamma(a)}, \quad a, b > 0,$$

form a conjugate family. Obtain the posterior distribution of θ when the prior distribution is a member of this family. Identify this distribution. [6]

(b) Obtain the predictive distribution of the lifetime T of another battery and show that

$$Pr(T \geq t | t_1, \ldots, t_n) = \left(\frac{b+s}{b+s+t} \right)^{a+n}$$

where $s = \sum_i t_i$. The following lifetimes in months of batteries were recorded:

8.33	0.78	13.07	11.83	9.34
2.34	3.23	3.89	2.28	5.59

Find the probability that the lifetime of another battery exceeds 8 months, given the prior distribution of θ is gamma (2,10). [13]

(c) It is required to obtain the predictive density of the total lifetime Y of m future batteries. What is the distribution of Y given θ? Obtain the predictive density of Y up to a constant of proportionality. [6]

Question 2

(a) The results of k experiments on female rats yield the numbers y_i of rats who developed an endrometrial stromal polyp out of the total number n_i of rats in experiment i for $i = 1, \ldots, k$. The data are modelled by assuming that the y_i are independent binomial random variables with parameters n_i, θ_i, respectively, where θ_i is the probability of a rat developing a polyp in the ith experiment. Write $\underline{y} = (y_1, y_2, \ldots, y_k), \underline{\theta} = (\theta_1, \ldots, \theta_k)$. Show that the likelihood function of the data is given by

$$p(\underline{y}|\underline{\theta}) \propto \prod_{i=1}^{k} \theta_i^{y_i}(1-\theta_i)^{n_i - y_i}.$$

[3]

(b) Now assume that the θ_i are an independent sample from a beta (α, β) distribution. That is,

$$p(\underline{\theta}|\alpha,\beta) = \prod_{i=1}^{k} \left\{ \frac{\theta_i^{\alpha-1}(1-\theta_i)^{\beta-1}}{B(\alpha,\beta)} \right\}$$

where $B(.,.)$ is the beta function. Further assume a vague prior for (α, β) of the form $p(\alpha, \beta) \propto (\alpha\beta)^{-1}$. Show that the full posterior density $p(\underline{\theta}, \alpha, \beta|\underline{y})$ is proportional to $(\alpha\beta)^{-1} p(\underline{\theta}|\alpha, \beta) p(\underline{y}|\theta)$. Hence write down an expression for $p(\underline{\theta}, \alpha, \beta|\underline{y})$. (Do not attempt to calculate the constant of proportionality.) [6]

(c) Show that the joint marginal posterior density of (α, β) is given by

$$p(\alpha, \beta|\underline{y}) \propto \frac{\prod_{i=1}^{k} B(\alpha + y_i, \beta + n_i - y_i)}{\alpha\beta\{B(\alpha,\beta)\}^k}.$$

Suppose that the parameter $\mu = \frac{\alpha}{\alpha+\beta}$ is of primary interest. Deduce the form of the joint posterior density of (μ, ϕ), where $\phi = \alpha + \beta$. Suggest a method for evaluating the marginal posterior distribution of μ. [16]

Question 3

(a) Define a $100(1-\alpha)\%$ credible interval and a $100(1-\alpha)\%$ Highest Posterior Density (HPD) interval for an unknown parameter θ. Prove that a $100(1-\alpha)\%$ HPD interval is the shortest possible credible interval, assuming that the posterior distribution has a unimodal density. [8]

SEE NEXT PAGE

(b) Observations are assumed to come from a Cauchy distribution with pdf

$$p(x|\theta) = \pi^{-1}[1 + (x - \theta)^2]^{-1} \qquad -\infty < x < \infty$$

The prior density for θ is assumed to be proportional to a constant. Thus the posterior density for θ is given by

$$p(\theta|\underline{x}) \propto \prod_{i=1}^{n}[1 + (x_i - \theta)^2]^{-1}$$

Suppose there are 5 observations 2.1, 4.9, 3.5, 2.8 and 4.1. Find the missing lines in the following table corresponding to $\theta = 4.0,\ 4.5,\ 5.0$ where

$$H(\theta) = 10^5[1 + (2.1 - \theta)^2]^{-1}[1 + (4.9 - \theta)^2]^{-1} \ldots [1 + (4.1 - \theta)^2]^{-1}$$

and the integrals are calculated using Simpson's rule. [10]

θ	$H(\theta)$	$\int_{-\infty}^{\theta} H(t)dt$	$p(\theta\|\underline{x})$	$\int_{-\infty}^{\theta} p(t\|\underline{x})dt$
0.0	0	0	0	0
0.5	2		.000	
1.0	9	2.8	.000	.000
1.5	56		.006	
2.0	365	102.5	.042	.012
2.5	1643		.187	
3.0	4171	1953.8	.475	.223
3.5	5632		.642	
4.0				
4.5				
5.0				
5.5	48		.005	
6.0	7	8777.5	.000	1.000
6.5	1		.000	
7.0	0	8779.3	.000	1.000

Explain how you would amend the table if the prior for θ was normal with mean μ and variance σ^2. [3]

Why is numerical integration important for the Bayesian approach to statistics? [4]

Question 4

(a) A three-stage linear model is given by

$$\begin{aligned} \underline{y}|\underline{\theta}_1, C_1 &\sim N(A_1\underline{\theta}_1, C_1) \\ \underline{\theta}_1|\underline{\theta}_2, C_2 &\sim N(A_2\underline{\theta}_2, C_2) \\ \theta_2|\underline{\mu}, C_3 &\sim N(\underline{\mu}, C_3) \end{aligned}$$

where $\underline{y}$ is an $n \times 1$ vector, $\underline{\theta}_1$ is a $p_1 \times 1$ vector, $\underline{\theta}_2$ is a $p_2 \times 1$ vector, A_1, A_2,C_1, C_2, C_3 and $\underline{\mu}$ are known.

Using the results for the two stage model given below show that the posterior distribution of $\underline{\theta}_1$ is $N(D\underline{d}, D)$ where

$$\begin{aligned} D^{-1} &= A_1^T C_1^{-1} A_1 + \{C_2 + A_2 C_3 A_2^T\}^{-1}, \\ \underline{d} &= A_1^T C_1^{-1} \underline{y} + \{C_2 + A_2 C_3 A_2^T\}^{-1} A_2 \underline{\mu}. \end{aligned}$$

Use the matrix lemma given below to find the corresponding result when $C_3^{-1} \to 0$. [9]

[Note: The two-stage linear model is given by

$$\begin{aligned} \underline{y}|\underline{\theta}_1, C_1 &\sim N(A_1\underline{\theta}_1, C_1) \\ \underline{\theta}_1|\underline{\mu}, C_2 &\sim N(\underline{\mu}, C_2), \end{aligned}$$

where A_1, C_1, C_2 and $\underline{\mu}$ are known. The marginal distribution of $\underline{y}$ is $N(A_1\underline{\mu}, C_1 + A_1 C_2 A_1^T)$ and the posterior distribution of θ_1 is $N(Bb, B)$ where

$$\begin{aligned} B^{-1} &= A_1^T C_1^{-1} A_1 + C_2^{-1} \\ \underline{b} &= A_1^T C_1^{-1} \underline{y} + C_2^{-1} \underline{\mu} \end{aligned}$$

Matrix lemma: for any matrices A_1, C_1, C_2 of appropriate dimensions for which the inverses stated in the result exist we have

$$\{C_1 + A_1 C_2 A_1^T\}^{-1} = C_1^{-1} - C_1^{-1} A_1 (A_1^T C_1^{-1} A_1 + C_2^{-1})^{-1} A_1^T C_1^{-1}\}.]$$

(b) Consider a simple linear regression model $y_i|\alpha, \beta \sim N(\alpha + \beta(x_i - \bar{x}), \sigma^2)$ for $i = 1, 2, \ldots, n$ where $\alpha \sim N(0, \sigma^2), \beta \sim N(2, \sigma^2)$ and $Cov(\alpha, \beta) = 0$. Write this as a two stage normal linear model. If the values of y_i and x_i are as given below, find the posterior distributions of α and β in terms of σ. [16]

y_i	2.6	4.5	7.1	8.7	11.0
x_i	1	2	3	4	5

SEE NEXT PAGE

Question 5

(a) It is desired to estimate the posterior means of three parameters, θ, δ and ϕ. The full conditional distributions,

$$p(\theta|\delta, \phi, \text{data}), \quad p(\delta|\phi, \theta, \text{data}), \quad p(\phi|\theta, \delta, \text{data})$$

are known. Explain how the posterior means may be estimated using the Gibbs sampling method. [10]

(b) The brightness (J) of a comet may be modelled by a Pareto distribution with density

$$p(J \mid \alpha) = \frac{\alpha J_x^{\alpha}}{J^{\alpha+1}} \qquad J > J_x$$

where J_x is the lower limit of brightness for the sample considered and α is a brightness index.

(i) If a sample of brightness values $J_1, J_2, \ldots, J_n$ for n comets is available show that the maximum likelihood estimate of α is given by

$$\hat{\alpha} = \frac{n}{\sum_{i=1}^{n} \log_e(J_i/J_x)}.$$

[4]

(ii) If the prior distribution for α is taken as Gamma(a, b) find the posterior distribution in terms of a, b, n and $\hat{\alpha}$. [4]

(iii) Three groups of comets of sizes n_1, n_2 and n_3 have brightness parameters α_1, α_2 and α_3. The prior distribution for each α_i is taken as Gamma(a, b). For physical reasons the brightness parameters must be ordered $\alpha_1 < \alpha_2 < \alpha_3$. Explain how the posterior means of the parameters can be estimated by adapting the Gibbs sampling algorithm you have described in part (a). [7]

INTERNAL EXAMINER: K.D.S. Young
EXTERNAL EXAMINER: W. Krzanowski