ISDS First Year Exam

May 6, 1995

You have four hours to complete the exam. The exam is closed book; you may not use any reference materials (books, notes, friends, etc). Please answer 6 of the following 12 questions. Make sure that each answer is on a separate sheet. Also, please make up a three letter code and write it on the top of each page.

Problem 1

The beta density for a random variable X in $(0,1)$ is

$$f_X(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$$

for positive parameters a and b. Find the variance of X. Show your work.

Problem 2

Assume that $X_1, ..., X_n$ are independent and identically distributed with distribution:

$$p(x|\theta) = e^{-(x-\theta)} \quad x > \theta > 0$$

a) Find a sufficient statistic. Is it minimal sufficient? Please explain.

b) Find the form of the natural conjugate prior distribution.

c) Find the conjugate prior distribution that incorporates the expert opinion that $P(\theta > 1) = .1$ and $P(\theta < 2) = 1$

d) Take $n = 2$, and $x_1 = 1$ and $x_2 = 1.23$. Find the Bayes factor for testing the null hypothesis $H_0 : \theta < .5$ against $H_1 : \theta \geq .5$ Use the prior as in c) if you succeeded in solving (c); otherwise make one up for the purpose of d).

Problem 3

Consider simple linear regression: $y_i = a + bx_i + \epsilon_i$ where $\epsilon_i \overset{ind}{\sim} N(0, \sigma^2)$. You are to select the x_i; all must be in the interval [0,100]. Your goal is to estimate b as precisely as possible. You can make 10 observations. Which 10 $x's$ would you choose? Show that your answer is best (define what you mean by best). In a practical setting you might deviate from your answer–why?

Problem 4

Urinary fluoride concentration was measured on 3 randomly chosen livestock at the beginning and in the middle of their grazing period in a region previously exposed to fluoride pollution.

The data are as follows:

Subject:	1	2	3
Beginning	24.7	46.1	26.3
Middle	12.4	14.1	19.7

a) Specify a statistical model for the analysis of these data.

b) Is there any evidence that there has been a decrease in the fluoride concentration during the period under consideration? Please explain.

Problem 5.

Let $\pi(x_1, x_2)$ be a p.m.f. (probability mass function) defined on the finite state space $S \times S$. We can define a Gibbs sampler with the transition probability matrices

$$\begin{aligned} P_1\left((x_1, x_2) \to (x_1', x_2')\right) &= \pi(x_1'|x_2) \cdot I[x_2' = x_2] \\ P_2\left((x_1, x_2) \to (x_1', x_2')\right) &= \pi(x_2'|x_1) \cdot I[x_1' = x_1] \end{aligned}$$

where $\pi(\,\cdot\,|x_2)$ is the conditional p.m.f. of x_1 given x_2 and $\pi(\,\cdot\,|x_1)$ is the conditional p.m.f. of x_2 given x_1. $I[x' = x] = 1$ if $x' = x$ and is 0 otherwise.

Show that both P_1 and P_2 have $\pi(x_1, x_2)$ as a stationary distribution.

Problem 6

Use the fact that the ratio of two independent standard normal variables has a Cauchy distribution to prove that for a random sample (X_1, X_2) of size $n = 2$ from $N(0, \sigma^2)$,

$$t = \bar{X}/S = \frac{(X_1 + X_2)/2}{|X_1 - X_2|/2}$$

has a Cauchy distribution.

Problem 7

There are 100 light sockets in a certain prison. Whenever a bulb goes out, an alarm rings. 100 brand new bulbs are installed and these are assumed to have independent exponential distributions with parameter λ, so the density is $\lambda e^{-\lambda t}$, $t > 0$. The alarm rings at time T.

(a) Find the maximum likelihood estimate of λ.

(b) Each time the alarm sounds, the spent bulb is replaced with a new bulb (also $\exp(\lambda)$). Having observed $T_1, \ldots, T_{150}$, find an approximate 95% confidence interval for λ.

Problem 8

An experimenter thinks that the time it takes a mouse to run a maze varies in a cyclic pattern within a period of 24 hours. He wishes to estimate the phase and amplitude of the cycle. Beginning at time 0, he takes measurements at times $t_1, \ldots, t_n$ (in hours) and tries to fit the non-linear model

$$Y_i = .25 + A \cos(\pi t_i/12 + \Phi) + \varepsilon_i$$

where A is the amplitude and Φ is the phase. Here Y_i is the time it takes a mouse to run the maze if the measurement is taken at t_i hours after the start of the experiment (every 24 hours a new day starts, but t_i keeps increasing).

a) Convert this to the usual linear model setup, $Y = X\beta + \epsilon$, and write out what the appropriate X matrix and β vector are. (Remember: $\cos(\alpha + \beta) = \cos\alpha \cos\beta - \sin\alpha \sin\beta$).

b) What are the MLE's of A and Φ?

Problem 9

Let y_f be a future observation taken independent of $(y_1, \ldots y_n)$, $y_i \, iid \, U(0, \theta)$. Let the prior for θ be $p(\theta) \propto 1/\theta$, $\theta > 0$.

(a) Find the predictive distribution of y_f.

(b) Find the $E(\, y_f | y_1, \ldots, y_n)$.

(c) Find the Var $(y_f | y_1, \ldots, y_n)$.

(d) Further let $y_{f_1}, \ldots y_{f_m}$ be m future observations. Find the predictive distribution of max $(y_{f_1}, \ldots y_{f_m})$ and the predictive distribution of min $(y_{f_1}, \ldots y_{f_m})$.

Problem 10

We want to investigate the relation between the occurrence of a certain rare disease and some covariate. Let y_i be an indicator for the disease, i.e. $y_i = 1$ if subject i has the condition, and $y_i = 0$ otherwise. Let x_i denote the covariate for the ith subject. We use a logistic regression model

$$Pr(y_i = 1|x_i, \alpha, \beta) = \left(1 + e^{-\alpha - x_i\beta}\right)^{-1}, \quad (1)$$

where $-\infty < \alpha < \infty$ and $-\infty < \beta < \infty$. Since the disease is rare, a simple random sample from the population might not contain any cases $y_i = 1$. Instead of taking a random sample, we therefore select m patients known to have the disease, and record their covariates $x_1, \ldots, x_m$, and k subjects who are known not to have the disease, and record their covariates $x_{m+1}, \ldots, x_{m+k}$. I.e. we fix y_i and record x_i. Denote with $n = m + k$ the total number of subjects.

Assume

$$p(\alpha, \beta) \propto c \quad (2)$$

$$x_i \overset{iid}{\sim} N(\mu, \Sigma). \quad (3)$$

In other words, we assume an – improper – constant prior on (α, β), and a normal distribution for the covariates x_i. Assume the hyperparameters (μ, Σ) are known. Use y and x to denote $y = (y_1, \ldots, y_n)$ and $x = (x_1, \ldots, x_n)$. Assume x is a priori independent of (α, β) and the y_i's are conditionally independent given α, β and x.

In the probability model (1) – (2) find the following distributions.

a) The marginal distribution of y_i given (α, β) but **not** conditional on x_i, i.e. $p(y_i|\alpha, \beta)$.

b) The "retrospective likelihood" $p(x_i|y_i, \alpha, \beta)$.

c) The joint posterior: $p(\alpha, \beta|x, y)$ where $y = (y_1, \ldots, y_n)$ and $x = (x_1, \ldots, x_n)$

Hints: The answers may not involve a closed form solution – you may have to write the answer as an integral/expectation. You may use $p(y_i|x_i, \alpha, \beta)$ for (1), $p(x_i)$ for (3), and $p(\alpha, \beta)$ for (2) in your solutions.

Problem 11

Implantable heart defibrillators (IHD) are small devices which are implanted into heart patients. When the device detects heart arrhythmia it discharges a small electric shock of a certain strength x to stop the fibrillation. In a trial with dogs as subjects we induce heart fibrillation for n dogs. Dog i has a defibrillator set at a certain strength x_i, for $i = 1, \ldots, n$. We record indicators y_i for successful ($y_i = 1$) or failed ($y_i = 0$) defibrillation (don't worry for the dogs – if defibrillation fails they get a "back-up" shock to bring them back to normal).

It is reasonable to model the probability of successful defibrillation by a logistic regression:

$$Pr(y_i = 1 | x_i, \alpha_i, \beta_i) \quad = \quad \left(1 + e^{-\alpha_i - x_i \beta_i}\right)^{-1}, \tag{4}$$

with (α_i, β_i) as subject-specific logistic regression parameters. The pairs (α_i, β_i) are related by

$$(\alpha_i, \beta_i) \quad \stackrel{iid}{\sim} \quad N(\mu, \Sigma), \; i = 1, \ldots, n, \tag{5}$$

$$\mu \quad \sim \quad N(a, A). \tag{6}$$

Assume all other hyperparameters (a, A, Σ) are known. Denote with $y = (y_1, \ldots, y_n)$ the data vector. Assume the y_i's are conditionally independent given α, β and the x_i's.

Find the following distributions.
Hint: Only (d) takes the form of a well known distribution. For (a)–(c) just write out the appropriate expression for the desired p.d.f. You may write $p(y_i | x_i, \alpha_i, \beta_i)$ for (4), $p(\alpha_i, \beta_i | \mu)$ for (5), and $p(\mu)$ for (6).

a) The joint likelihood $p(y | \alpha_1, \beta_1, \ldots, \alpha_n, \beta_n, \mu)$.

b) The posterior distribution $p(\alpha_1, \beta_1, \ldots, \alpha_n, \beta_n, \mu | y)$.

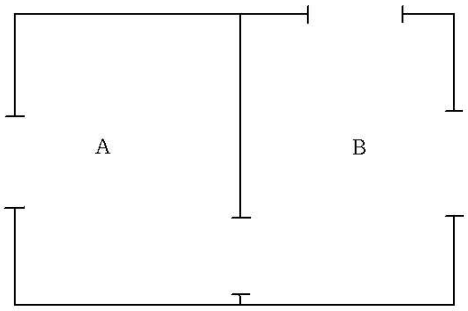
c) The conditional posterior $p(\alpha_i, \beta_i | \mu, \alpha_j, \beta_j, j \neq i, y)$.

d) The conditional posterior distribution $p(\mu | \alpha_1, \beta_1, \ldots, \alpha_n, \beta_n, y)$.
Hint: Just state the distribution, you don't need to give the explicit formulas for the particular parameters.

e) Propose a Markov Chain Monte Carlo scheme to implement posterior inference in the model defined by (4) – (6).

Problem 12

A bee buzzes about a 2 room house shown below. At each successive stage independently, it is equally likely to leave its current compartment by any of the available exits. Once the bee leaves the house, it never returns.



If the bee starts in room A, what is the expected number of visits to room B before it leaves the house.