UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Thursday, 6 June, 2013 1:30 pm to 3:30 pm

PAPER 34

APPLIED BAYESIAN STATISTICS

*Attempt no more than **THREE** questions.*

*There are **FOUR** questions in total.*

The questions carry equal weight.

STATIONERY REQUIREMENTS

Cover sheet
Treasury Tag
Script paper

SPECIAL REQUIREMENTS

None

You may not start to read the questions printed on the subsequent pages until instructed to do so by the Invigilator.

1

Assume $y_1, .., y_n$ are independent observations of lengths of caterpillar, assumed drawn from a density $Y|\theta \sim \text{Unif}(0, \theta),\ Y < \theta$, where θ is the maximum size this breed of caterpillar can grow.

(a) Show that the likelihood function for θ is $\propto \theta^{-n}, \theta > M$, where $M = \max(y_1, .., y_n)$.

(b) Suppose we assume θ follows a Pareto distribution, $\theta \sim \text{Pareto}(\alpha, \beta), \alpha, \beta > 0$. Show that the Pareto is conjugate to the Uniform distribution, and the posterior distribution for θ is $\text{Pareto}(\alpha + n, \max(\beta, M))$. [A $\text{Pareto}(\alpha, \beta)$ distribution has density $p(\theta) = \alpha\beta^\alpha\theta^{-(\alpha+1)}; \theta > \beta$, 0 otherwise.]

(c) Derive a form for the predictive density $p(y_1, .., y_n|\alpha, \beta)$.

(d) Suppose we now have a series of I different breeds, each with n_i observations denoted $\mathbf{y}_i = (y_{i1}, .., y_{in_i})$, where $Y_{ij}|\theta \sim \text{Unif}(0, \theta_i), j = 1, ..., n_i$. Explain the meaning of an assumption of exchangeability of the θ_i's, and when it might be reasonable.

(e) Each θ_i is assumed to have a common prior Pareto distribution with known β but unknown α. Write down an expression for $p(\mathbf{y}_1, .., \mathbf{y}_I|\alpha, \beta)$.

(f) Define the Type II maximum likelihood estimate $\hat{\alpha}$ for α, and show that it obeys the equation

$$\frac{I}{\hat{\alpha}} - \sum_i \frac{1}{n_i + \hat{\alpha}} = \sum_i \log \max(1, M_i/\beta),$$

where $M_i = \max(y_{i1}, .., y_{in_i})$. Show that if $n_1 = n_2 = .. = n_I$, then $\hat{\alpha}$ is a solution to a quadratic equation.

(g) If β were greater than all the observed data-points, what would $\hat{\alpha}$ be? Would this be sensible?

(h) Suppose someone claimed that the sampling distributions for the Y_{ij} were not uniform, but exponential. Briefly outline the steps to compare these models using DIC.

2

(a) Suppose a random variable Y is assumed to have probability density $p(y|\theta)$ for a scalar parameter θ. Define a Jeffreys prior $p_J(\theta)$ for θ.

(b) In a scale model $p(y|\sigma) = \frac{1}{\sigma} f\left(\frac{y}{\sigma}\right)$, show that $E[g(Y/\sigma)]$ is independent of σ for any function g.

(c) Hence show that the Jeffreys prior in a scale model $p(y|\sigma) = \frac{1}{\sigma} f\left(\frac{y}{\sigma}\right)$ is $p_J(\sigma) \propto 1/\sigma, \quad \sigma > 0$.

(d) Show that this prior is scale invariant, in that $c\sigma$ has the same (improper) distribution as σ for all $c > 0$

(e) Benford's Law states that, in many collections of numbers in the real world, the leading digit i, for $i = 1, 2, 3, 4, 5, 6, 7, 8, 9$, occurs in proportion given by $\frac{\log(1+1/i)}{\log 10}$. Show that if X has a density proportional to $1/x$ on the range (1,10), then X obeys Benford's Law exactly.

(f) Show that if X has a density proportional to $1/x$ on the range $(10^a, 10^b)$, for any $b > a$, then X obeys Benford's Law exactly.

(g) Given a null hypothesis that fully specifies a probability density $p_0(x)$, and and a set of observations $\mathbf{x} = x_1, \dots, x_n$ that may or may not obey that distribution, what is meant by a checking (or discrepancy) function $T(\mathbf{X})$?

(h) Suppose you had a set $(y_1, \dots, y_9)$ comprising the counts of the leading digits in a collection of numbers. Suggest one or more appropriate checking functions for Benford's Law, and describe roughly how you would implement them using R or WinBUGS.

(i) The following table from Eurostat shows the leading digits in 140 values in the National Accounts of Greece in 2009 [real data].

Leading digit	Benford prediction	Greece 2009	prop
1	0.30	41	0.29
2	0.18	37	0.26
3	0.12	28	0.20
4	0.10	14	0.10
5	0.08	3	0.02
6	0.07	6	0.04
7	0.06	7	0.05
8	0.05	4	0.03
9	0.05	0	0.00
		140	

If there is a distribution over the leading digit of sums of money, why might this distribution be scale-invariant?

(j) Without doing any calculations, are you suspicious about these figures?

3

Suppose there are N individuals, grouped into I groups, with n_i in the ith group all having covariate vector $\mathbf{x}_i = (x_{i1}, .., x_{ip})$. Each individual is then classified into one of K disjoint categories, and let $Y_{ik}, i = 1, ..., I; k = 1, .., K$ be the number of individuals in group i that are classified into category k.

(a) We assume $\mathbf{Y}_i = (Y_{i1}, .., Y_{iK})$ is multinomial with parameters $\mathbf{p}_i = (p_{1i}, .., p_{iK})$, $\sum_k p_{ik} = 1$ and $n_i = \sum_k Y_{ik}$. Give the form of the density for $\mathbf{Y}_i$.

(b) We assume a regression model for each category $k > 1$ versus category 1 given by

$$\log \frac{p_{ik}}{p_{i1}} = \boldsymbol{\beta}_k' \mathbf{x}_i,$$

where $\boldsymbol{\beta}_k = (\beta_{k1}, .., \beta_{kp})$. Assuming $\boldsymbol{\beta}_1 = 0$, show the overall likelihood for $\boldsymbol{\beta}$ based on observations $\mathbf{y}_i$, $i = 1, \ldots, I$ is proportional to

$$\prod_{i=1}^{I} \frac{e^{\sum_{k=1}^{K} y_{ik} \boldsymbol{\beta}_k' \mathbf{x}_i}}{\left[\sum_{k=1}^{K} e^{\boldsymbol{\beta}_k' \mathbf{x}_i})\right]^{n_i}}$$

(c) Suppose we assume the data in fact were generated by

$$\begin{aligned} Y_{ik} &\sim \text{Poisson}(\mu_{ik}) \\ \mu_{ik} &= \mu_{i1} e^{\boldsymbol{\beta}_k' \mathbf{x}_i} \end{aligned}$$

Show that if we assume that the μ_{i1}'s have independent Gamma(a, b) prior distributions, the marginal likelihood $\propto p(\mathbf{y}_1, .., \mathbf{y}_I | \boldsymbol{\beta}_1, .., \boldsymbol{\beta}_K)$ has the form

$$\propto \prod_{i=1}^{I} \frac{e^{\sum_{j=1}^{K} y_{ij} \boldsymbol{\beta}_j' \mathbf{x}_i}}{\left[\sum_{j=1}^{K} e^{\boldsymbol{\beta}_j' \mathbf{x}_i} + b\right]^{n_i + a}}$$

(d) The table shows the feeding choices of 219 alligators, where the response measure for each alligator is one of $K = 5$ categories: fish, invertebrate, reptile, bird, other. Possible explanatory factors are the length of alligator ($\leqslant 2.3$ metres and > 2.3 metres), and the lake (Hancock, Oklawaha, Trafford, George).

		Primary Food Choice				
Lake	Size	Fish	Invertebrate	Reptile	Bird	Other
Hancock	$\leqslant 2.3$	23	4	2	2	8
	> 2.3	7	0	1	3	5
Oklawaha	$\leqslant 2.3$	5	11	1	0	3
	> 2.3	13	8	6	1	0
Trafford	$\leqslant 2.3$	5	11	2	1	5
	> 2.3	8	7	6	3	5
George	$\leqslant 2.3$	16	19	1	2	3
	> 2.3	17	1	0	1	3

There are $I = 8$ groups, and the covariate $\mathbf{x_i} = (x_{i1}, .., x_{i5})$ is coded as

- $x_{i1} = 1$ if alligators in group i are from Lake Hancock, 0 otherwise
- $x_{i2} = 1$ if alligators in group i are from Lake Oklawaha, 0 otherwise
- $x_{i3} = 1$ if alligators in group i are from Lake Trafford, 0 otherwise
- $x_{i4} = 1$ if alligators in group i are from Lake George, 0 otherwise
- $x_{i5} = 1$ if alligators in group i are are > 2.3 metres, 0 if < 2.3 metres

β_1 is set to 0, and β_k, $k > 1$, given locally uniform prior distributions.

BUGS code includes the section

```
for (i in 1 : I) {     # loop around groups
     lambda[i] ~ dnorm(0, 0.00001) # vague priors
     for (k in 1 : K) {     # loop around foods
        y[i, k] ~ dpois(mu[i, k])
        log(mu[i, k]) <- lambda[i] + inprod(beta[k,], x[i,])
        }
     }
```

where `inprod(beta[k,], x[i,])` represents $\sum_j \beta_{kj} x_{ij} = \boldsymbol{\beta}'_k \mathbf{x_i}$.

Explain why this model should lead to a posterior distribution for the $\boldsymbol{\beta}$'s proportional to the Multinomial likelihood of part (a). Why might this be a more efficient (from a computational perspective) way of representing the model?

(e) Parts of the output of a BUGS run was as follows.

node	mean	sd
beta[2,1]	-1.852	0.542
beta[2,2]	0.880	0.411
beta[2,3]	1.077	0.430
beta[2,4]	-0.093	0.303
beta[2,5]	-1.524	0.397

	Dbar	Dhat	pD	DIC
y	164.6	137.7	26.8	191.5

Interpret the estimates for β_{21} and β_{25}. How does the estimated effective number of parameters in the DIC output compare with the actual number of free parameters estimated?

[A Poisson(μ) distribution has density $p(y|\mu) = \frac{\mu^y}{y!} e^{\mu}; y = 0, 1, ...$

A Gamma(a, b) distribution has density $p(\mu|a, b) = \frac{b^a}{\Gamma(a)} \mu^{a-1} e^{-b\mu}; \mu > 0$, 0 otherwise.]

4

(a) When expressing a distribution P to predict a random quantity X, define a *scoring rule* $S(P, X)$. If we really believe a distribution Q, find the expected score and give the conditions for the scoring rule to be proper and strictly proper.

(b) Suppose a weather forecaster is providing probabilities p_t of it raining on day t, where the outcome $X_t = 1$ if it rains, 0 otherwise. It is proposed to score them by the probability they give to the correct outcome, so that they score p_t if $X_t = 1$, and score $1 - p_t$ if $X_t = 0$. Show this is not a proper scoring rule, and find the expected score if the forecaster maximally exaggerates their confidence.

(c) When forecasting mean daily temperature Y, a logarithmic scoring rule is used, so that if a forecaster provides a predictive density $p_{\tilde{Y}}(\tilde{y})$, and y is then observed, they are rewarded $\log p_{\tilde{Y}}(y)$. Show that this is a strictly proper scoring rule. [Hint: the Kullback-Leibler inequality states that for two densities f, g (for which $f(x) = 0$ whenever $g(x) = 0$), that $E_f[\log(f/g)] = \int \log \frac{f(x)}{g(x)} f(x)dx > 0$, with equality if and only if $f = g$ almost everywhere.]

(d) Suppose we have two forecasters making sequential forecast distributions $f_{1t}(\tilde{y}_t)$ and $f_{2t}(\tilde{y}_t)$ for days $t = 1, .., T$. These forecast distributions are constructed as follows. The first forecaster has a model $p_1(y|\theta)$ and prior distribution $p_1(\theta)$, assesses a predictive density for the first observation $\tilde{Y}_1$ given by $f_{11}(\tilde{y}_1) = p_1(\tilde{y}_1) = \int p_1(\tilde{y}_1|\theta)p_1(\theta)d\theta$. Having observed y_1, they update to a posterior distribution $p_1(\theta|y_1) \propto p_1(y_1|\theta)p_1(\theta)$, create a predictive distribution $f_{12}(\tilde{y}_2) = p_1(\tilde{y}_2|y_1)$, and so on. When the temperature y_t is observed, the forecaster is scored according to a logarithmic scoring rule, so that on day t forecaster 1 scores $L_{1t} = \log f_{1t}(y_t)$. Their total score $T_1 = \sum_t L_{1t}$ is recorded.

The second forecaster goes through a similar process using their own model. Show that the quantity $e^{T_1 - T_2}$ is equivalent to the Bayes factor for comparing models p_1 and p_2 based on the data $(y_1, ..., y_T)$.

(e) Suppose we decided to take a weighted average of the two forecasters, so that we adopted our own forecast distribution $f_t(\tilde{y}_t) = w_{1,t}f_{1t}(\tilde{y}_t) + w_{2,t}f_{2t}(\tilde{y}_t)$, where $w_{1,t} + w_{2,t} = 1$ and the ratio of the weights starts at $w_{1,1}/w_{2,1} = 1$, and then is updated according to the formula

$$\frac{w_{1,t+1}}{w_{2,t+1}} = \frac{e^{L_{1t}}}{e^{L_{2t}}} \frac{w_{1,t}}{w_{2,t}}.$$

Show that this is equivalent to assuming a full Bayesian procedure in which we consider the hypotheses that each of the forecasters had the 'correct' model, and initially assign equal prior probability to these two hypotheses.

(f) The Bayes factor gives the relative support for the two forecasters. Suppose we only had the predictive distributions $f_{1t}(\tilde{y}_t)$ for forecaster 1, and the observations $(y_1, ..., y_T)$. Briefly, how we might assess the absolute quality of these predictions?

END OF PAPER