Durham University

EXAMINATION PAPER

Examination Session:	Year:	Exam Code:
May	2018	MATH4031-WE01

Title: Bayesian Statistics IV

Time Allowed:	3 hours	
Additional Material provided:	None	
Materials Permitted:	None	
Calculators Permitted:	Yes	Models Permitted: Casio fx-83 GTPLUS or Casio fx-85 GTPLUS.
Visiting Students may use dictionaries: No		

Instructions to Candidates:	Credit will be given for: the best **TWO** answers from Section A, the best **THREE** answers from Section B, **AND** the answer to the question in Section C. Questions in Section B and C carry **TWICE** as many marks as those in Section A.

Revision:	

SECTION A

1. (a) Let $x_{1:n} = (x_1, ..., x_n)$ be an exchangeable sequence of observables, $\theta \in \Theta$ be the unknown parameter with prior distribution $\pi(\theta)$, and posterior distribution $\pi(\theta|x_{1:n})$, where Θ is a countable parametric space. Suppose $\theta^* \in \Theta$ is the only true value of θ, with $\pi(\theta^*) > 0$, and $-\mathrm{KL}(f(\cdot|\theta^*), f(\cdot|\theta)) := \int \log \frac{f(x|\theta)}{f(x|\theta^*)} f(\mathrm{d}x|\theta^*) < 0$ for all $\theta \neq \theta^*$. Prove that

$$\lim_{n\to\infty} \pi(\theta|x_{1:n}) = \begin{cases} 1 & , \theta = \theta^* \\ 0 & , \theta \neq \theta^* \end{cases}$$

   (b) Give an interpretation of the above result.

2. Consider the Bayesian model $(f(x_{1:n}|\theta), \pi(\mathrm{d}\theta))$, where $x_{1:n} = (x_1, ..., x_n)$ is an exchangeable sequence of observables, and $\theta \in \Theta$ is the uncertain parameter.

   (a) Give a definition of the Bayesian point estimator δ^π of θ under the loss function $\ell(\theta, \delta)$.

   (b) Let $\ell(\theta, \delta) = w(\theta)(\theta - \delta)^2$ be the weighted quadratic loss function, where $w(\theta)$ is a non-negative function, and $\theta \in \Theta \subset \mathbb{R}$. Prove that the Bayes estimator $\delta^\pi(x_{1:n})$ under the weighted quadratic loss function is

$$\delta^\pi(x_{1:n}) = \frac{\mathrm{E}^\pi(w(\theta)\theta|x_{1:n})}{\mathrm{E}^\pi(w(\theta)|x_{1:n})} \tag{1}$$

   (c) Derive $w(\theta)$ for $\theta \in \Theta \subset \mathbb{R}$, such that the mean squared error of the Bayesian estimator δ^π in (1) is minimised. Justify your answer.

   (d) State the duality principle that the Bayes estimator (1) exhibits. Justify your answer.

3. (a) Describe how the 'Frequentist', the 'Subjective Bayesian', and the 'Objective Bayesian' schools of statistics interpret probability.

   (b) Recall the Representation theorem for $0-1$ random quantities.

   Theorem: If $x_1, x_2, ...$ is an infinitely exchangeable sequence of $0-1$ random quantities with probability measure P, there exists a distribution function Π such that the joint mass function $p(x_1, ..., x_n)$ for $x_1, ..., x_n$ has the form

$$p(x_1, ..., x_n) = \int_0^1 \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} \mathrm{d}\Pi(\theta)$$

   i. Present the quantities $\Pi(\theta)$ and θ with respect to $\{x_i\}_{i\geq 1}$. (You are not required to prove the theorem.)

   ii. From the Subjective probability perspective, provide an interpretation of the aforesaid theorem regarding: the generation of the sequence of random variables $x_1, ..., x_n$; the quantity $\Pi(\theta)$; and the quantity θ.

4. Laplace's method gives an approximation to integrals of the form

$$I = \int_{-\infty}^{\infty} \exp\{-E(x)\}dx \tag{2}$$

by the formula

$$I \approx \exp\{-E(\tilde{x})\}\sqrt{\frac{2\pi}{E''(\tilde{x})}} \tag{3}$$

where $E(x)$ is a convex and two times differentiable function with a minimum at $x = \tilde{x}$.

(a) Show how this approximation formula can be obtained.

(b) Derive the Laplace approximation for the normalization constant for the following non-normalized density

$$f(x) \propto e^{-\beta x} e^{-\alpha e^{-x}}$$

where $x \in \mathbb{R}$, $\alpha > 0$, and $\beta > 0$.

5. (a) Given n samples $\{x_i\}_{i\in[1..n]}$ from a probability distribution on $\mathbb{R}$ with density g, write down the formula for the importance sampling estimate of the expectation $\mathbb{E}_f[m]$ of a function $m : \mathbb{R} \to \mathbb{R}$ with respect to a distribution on $\mathbb{R}$ with density f, using g as the importance function.

(b) What condition must g satisfy to be suitable for this approximation?

(c) Show that this is an unbiased estimator of the expectation.

(d) Assume we wish to compute $\mathbb{E}[|X|]$ for X from a t-distribution with 3 degrees of freedom, $X \sim \mathrm{t}_3$, using a Monte Carlo method. Consider the choices of importance function g corresponding to the following distributions:

i. t_3;

ii. t_1, the t-distribution with 1 degree of freedom;

iii. $\mathbf{N}(0, 1)$, the standard Normal distribution.

For each of these choices, explain whether you would use it, and why.

6. ***There is no question 6 on this paper.***

SECTION B

7. (a) Consider the Bayesian model

$$\begin{cases} x_i|\theta & \overset{\text{iid}}{\sim} f(\mathrm{d}\cdot|\theta),\ \forall i = 1, ..., n \\ \theta & \sim \pi(\mathrm{d}\theta) \end{cases} \tag{4}$$

with $\theta \in \Theta$ and $\{x_i \in \mathcal{X}\}_{i=1}^n$, and consider a loss function $\ell(\theta, \delta)$.

i. Give a definition of the term Frequentist risk (or average loss) $R(\theta, \delta)$ of the decision rule δ for parameter θ.

ii. Prove the following theorem: If a prior distribution π is strictly positive on Θ, with finite Bayes risk and the risk function, $R(\theta, \delta)$, is a continuous function of θ for every δ, the Bayes estimator δ^π is admissible.

(b) Consider the Bayesian model

$$\begin{cases} x|\theta & \sim \mathrm{N}(\theta, 1) \\ \theta & \sim \mathrm{N}(0, 1) \end{cases} \tag{5}$$

where $\theta \in \mathbb{R}$, and consider that only one observation has been collected.

i. Show that the posterior distribution of θ is $\theta|x \sim \mathrm{N}(\frac{1}{2}x, \frac{1}{2})$.

ii. Show that the Bayes point estimator under the loss function

$$\ell(\theta, \delta) = \exp(\frac{3}{4}\theta^2)(\theta - \delta)^2$$

is $\delta^\pi(x) = 2x$.

iii. Compute the Frequentist risk for decision rule $\delta_c(x) = cx$, where $c > 0$.

iv. Show that the Bayes point estimator $\delta^\pi(x) = 2x$ is inadmissible.

v. Examine and report what has caused the Bayes point estimator $\delta^\pi(x)$ to be inadmissible? Justify your answer.

8. Consider a sequence of exchangeable observables $x_{1:n} = (x_1, ..., x_n)$, where $x_i \in \mathcal{X}_k$, for $i = 1, ..., n$, and $\mathcal{X}_k = \{x \in \{0,1\}^k | \sum_{j=1}^k x_j = 1\}$. In words, x_i is a k-dimensional vector all of whose elements are equal to 0 except for one which is equal to 1, for $i = 1, ..., n$. Consider the Bayesian model

$$\begin{cases} x_i|\theta & \overset{\text{IID}}{\sim} \text{Mu}_k(\theta) \\ \theta & \sim \pi(\text{d}\cdot) \end{cases}$$

where $\theta \in \Theta$, with $\Theta = \{\theta \in (0,1)^k | \sum_{j=1}^k \theta_j = 1\}$. Here, Mu_k denotes the Multinomial probability distribution with PMF

$$\text{Mu}_k(x|\theta) = \begin{cases} \prod_{j=1}^k \theta_j^{x_j} & \text{, if } x \in \mathcal{X}_k \\ 0 & \text{, otherwise} \end{cases} \tag{6}$$

(a) Show that the parametric model (6) is a member of the $k-1$ exponential family.

(b) Compute the likelihood $f(x_{1:n}|\theta)$, and find the sufficient statistic $t_n := t_n(x_{1:n})$.

(c) Derive the conjugate prior distribution for θ, and then show that it is a Dirichlet distribution.

You may use the fact that the Dirichlet distribution $\text{Di}_k(\text{d}\theta|a)$ with parameter $a = (a_1, ..., a_k)$, where $\{a_j > 0;\ j = 1, ..., k\}$ has PDF

$$\text{Di}_k(\theta|a) = \begin{cases} \frac{\Gamma(\sum_{j=1}^k a_j)}{\prod_{j=1}^k \Gamma(a_j)} \prod_{j=1}^k \theta_j^{a_j - 1} & \text{, if } \theta \in \Theta \\ 0 & \text{, otherwise} \end{cases}$$

(d) Compute the posterior distribution. State the name of the distribution, and express its parameters with respect to the observations and the hyper-parameters of the prior. Justify your answer.

(e) Compute the probability mass function of the predictive distribution for a future observation $y = x_{n+1}$ in closed form.

Hint $\Gamma(x) = (x-1)\Gamma(x-1)$.

(f) Suppose you are interested in checking if a k-sided die is fair or not. You collect n observations $\{x_i\}_{i=1}^n$, where $x_i \in \mathcal{X}_k$, according to the following experiment. You throw the die n times; at the i-th throw, you record the result as $x_{i,j} = 1$ if the result is the j-th side and as $x_{i,j} = 0$ if the result is otherwise for $j = 1, ..., k$.

i. Set up the hypothesis test, by stating explicitly the pair of hypotheses, and computing the Bayes factor in closed form.

ii. Suppose that it is a 4-sided die, you throw it 6 times, and it comes up '1', 4 times; '2', 0 times; '3', 1 time; and '4', 1 time. Perform the Bayesian test to check whether the dice is fair or not. State your decision based on Jeffreys' scale rule.

9. (a) i. State the general algorithm for rejection sampling from a probability density function $\rho(x)$ given only an unnormalized version $\tilde{\rho}(x)$ and an envelope function $\tilde{q}(x) = kq(x)$, where q is another probability density function from which we can easily sample.

   ii. What assumption is being made about the relationship between $\tilde{\rho}$ and $\tilde{q}$?

   (b) Consider a random quantity X having Weibull distribution $\mathrm{We}(2, 1)$ with probability density function (pdf)

$$f(x) = 2x \exp\left\{-x^2\right\}$$

   i. Show that $f(x)$ is log-concave with respect to x.

   ii. Consider the problem of sampling from $\mathrm{We}(2, 1)$ using rejection sampling and using a proposal function based on two tangents to the logarithm of the pdf where the tangents are at $x = 4/5$ and $x = 3/2$.
   Derive the numerical algorithm for this sampling problem in order to produce two proposals of random draws from $\mathrm{We}(2, 1)$ and state whether such proposals are accepted or not. Base your calculations on the availability of the following sequence of uniformly distributed random numbers between 0 and 1:

0.279	0.624	0.118	0.441	0.132	0.910
0.923	0.538	0.586	0.258	0.400	0.111

   iii. Calculate the acceptance rate of the algorithm considered in part (ii).

   iv. Build the minimum curvature normal proposal for the pdf of $\mathrm{We}(2, 1)$ and compute the acceptance rate for the corresponding acceptance/rejection algorithm.

10. (a) State and prove the Markov blanket theorem concerning dependence in a Bayesian network, and explain the connection between the theorem and moralization.

(b) The development of road accident statistics is being studied for M similarly sized towns over the course of N years. Let A_{ij} be the number of accidents in year i and town j, and S_{ij} the number of serious accidents in year i and town j. Expert judgement is that for a given year, the variability between towns in number of accidents should be modelled well by a Poisson distribution with uncertain parameter λ_i, and the variation of the λ_i's in turn by a Gamma distribution with fixed parameters α and β. S_{ij}, the number of serious accidents in year i and town j, is assumed to be a fraction of all accidents A_{ij} in the respective year and town, with a long-term proportion θ_j specific for each town but the same for all years. The proportions θ_j, $j = 1, \ldots, M$, are assumed to be independent of the λ_i's, and follow a Beta distribution with fixed parameters a and b common to all towns.

i. Draw a directed acyclic graph for the Bayesian network describing the joint distribution of the $\{A_{ij}\}$, $\{S_{ij}\}$, $\{\lambda_i\}$ and $\{\theta_j\}$ based on the above, and add vertices for the further parameters that are assumed to be fixed. (Hint: use two intersecting plates.)

ii. Specify the distributions for the vertices $\{A_{ij}\}$, $\{S_{ij}\}$, $\{\lambda_i\}$ and $\{\theta_j\}$ given their respective parents.

iii. Derive (up to multiplicative constants) the conditional distributions for λ_i and θ_j given that, in each case, all other variables have been observed. (Note therefore that the conditioning variables are different for λ_i and for θ_j.) Describe briefly how random samples could be obtained from these distributions. You may refer to standard functions in R, or name standard algorithms. For any algorithm you name, show that preconditions (if any) for its application are met.

iv. On closer inspection, the experts are uneasy about the assumption of conditional independence of the λ_i's given α and β, and hypothesize a consistent trend over time in expected number of accidents instead. Make a suggestion on how the Bayesian network, i.e., the graph and the node distributions, could be modified to account for this.

The Poisson distribution for $x \in \{0, 1, \ldots\}$ with parameter λ takes the form:

$$P[x|\lambda] = e^{-\lambda}\frac{\lambda^x}{x!}$$

The pdf of a Gamma-distributed random quantity x with parameters u and v is:

$$f(x|u, v) = \frac{v^u}{\Gamma(u)}x^{u-1}e^{-vx}$$

The pdf of a Beta-distributed random quantity x with parameters a and b is:

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

SECTION C

11. Consider a collection of paired data $\{(y_i, x_i^{\mathrm{T}}),\, i = 1, ..., n\}$, where each pair (y_i, x_i^{T}) is independently collected, $y_i \in \mathbb{R}$, $x_i = (x_{i,1}, ..., x_{i,p})^{\mathrm{T}} \in \mathbb{R}^p$ $p \geq 1$, and $n \geq 1$. Consider the Normal linear model

$$y = X\beta + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma^2)$$

where the vector of responses is $y = (y_{1,...,}y_n)^{\mathrm{T}} \in \mathbb{R}^n$, the design matrix is $X = (x_1, ..., x_n)^{\mathrm{T}} \in \mathbb{R}^{n\times p}$, and the uncertain parameters are $\beta \in \mathbb{R}^p$ and $\sigma^2 > 0$.

(a) Compute the likelihood function $f(y_{1:n}|\beta, \sigma^2)$. For simplicity, you may suppress the conditioning on x_i's in the notation.

(b) Show that the parametric sufficient statistic is $t_n = (n, X^{\mathrm{T}}X, X^{\mathrm{T}}y)$ and justify your answer.

Hint If you use the following identity, you should prove it.

$$(y - X\beta)^{\mathrm{T}}(y - X\beta) = (\beta - \hat{\beta})^{\mathrm{T}}(X^{\mathrm{T}}X)(\beta - \hat{\beta}) + S$$

where $\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$, and $S = (y - X\hat{\beta})^{\mathrm{T}}(y - X\hat{\beta})$

(c) Derive the prior distribution for (β, σ^2) which is conjugate to this likelihood function. Show that the prior distribution satisfies

$$\pi(\mathrm{d}\beta, \mathrm{d}\sigma^2) = \mathrm{N}(\mathrm{d}\beta|m, \sigma^2 V)\, \mathrm{IG}(\mathrm{d}\sigma^2|a, d)$$

with $V > 0$, $a > 0$, $d > 0$.
Here, the PDF of (k-dimensional) Normal distribution is

$$\mathrm{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{k}{2}}\sqrt{\det(\Sigma)}} \exp(-\frac{1}{2}(x - \mu)^{\mathrm{T}}\Sigma^{-1}(x - \mu)) \text{ , for } x \in \mathbb{R}^k$$

Here, the PDF of the inverse Gamma is

$$\mathrm{IG}(x|a, d) = \begin{cases} \frac{(\frac{a}{2})^{\frac{d}{2}}}{\Gamma(\frac{d}{2})} x^{-\frac{d+2}{2}} \exp(-\frac{a}{2}\frac{1}{x}) & \text{, if } x \in (0, +\infty) \\ 0 & \text{, otherwise} \end{cases}$$

(d) The posterior distribution of (β, σ^2) is

$$\pi(\mathrm{d}\beta, \mathrm{d}\sigma^2|y_{1:n}) = \mathrm{N}(\mathrm{d}\beta|m^*, \sigma^2 V^*)\, \mathrm{IG}(\mathrm{d}\sigma^2|a^*, d^*)$$

with

$$m^* = (V^{-1} + X^{\mathrm{T}}X)^{-1}(V^{-1}m + X^{\mathrm{T}}y)\,; \qquad V^* = (V^{-1} + X^{\mathrm{T}}X)^{-1}$$
$$a^* = a + m^{\mathrm{T}}V^{-1}m + y^{\mathrm{T}}y - m^{*,\mathrm{T}}V^{*,-1}m\,; \qquad d^* = d + n$$

Show that $\mathrm{E}(\beta|y_{1:n}) = m^*$ and that $\mathrm{Var}(\beta|y_{1:n}) = V^*\frac{a^*}{d^*-2}$.

(e) Compute the marginal posterior distribution of β, and recognize that it is a multivariate Student-t distribution $\pi(\mathrm{d}\beta|y_{1:n}) = \mathrm{St}(\mathrm{d}\beta|m^*, a^*V^*, d^*)$.

Here, the PDF of (k-dimensional) Student-t distribution is

$$\mathrm{St}(x|m,V,d) = \frac{\Gamma(\frac{d+2}{2})}{\det(V)^{\frac{1}{2}}\pi^{\frac{k}{2}}\Gamma(\frac{d}{2})}\left(1+(x-m)^{\mathrm{T}}V^{-1}(x-m)\right)^{-\frac{d+k}{2}}, \text{ for } x \in \mathbb{R}^k$$

(f) Construct the $(1-a)100\%$ highest probability density credible set for β, and show that it is

$$C_a = \{\beta \in \mathbb{R}^p \,|\, (\beta - m^*)^{\mathrm{T}}(V^*)^{-1}(\beta - m^*) \leq \frac{pa^*}{d^*}F_{p,d+n,1-a}\}$$

where $F_{p,d+n,1-a}$ is the upper $1-a$ quantile of distribution F with degrees of freedom p and d.

Hint-1 If $x \sim \chi^2_{d_1}$ and $y \sim \chi^2_{d_2}$then $z = \frac{x/d_1}{y/d_2}$ with $z \sim \mathrm{F}_{d_1,d_2}$

Hint-2 $y \sim \mathrm{IG}(a,b)$ if and only if $z = ay^{-1}$ with $z \sim \chi^2_b$