



The
University
Of
Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS

Autumn Semester 2018–19

Bayesian Statistics

2 hours

Candidates may bring to the examination a calculator which conforms to University regulations.

Marks will be awarded for your best **three** answers. Total marks 84.

Standard results from the lecture notes may be used without derivation, but must be clearly stated.

**Please leave this exam paper on your desk
Do not remove it from the hall**

Registration number from U-Card (9 digits)
to be completed by student

--	--	--	--	--	--	--	--	--

Blank

1 An astrophysicist studying the expansion of the universe has a set of redshift measurements, each covering the same section of the Milky Way. If current theory is correct, redshift should be close to zero.

- (i) Let x_i be the i -th observation and assume $x_i \sim N(x_i | \mu, 1/\lambda)$, $i = 1, \dots, n$. Write down the joint posterior distribution of (μ, λ) using the conjugate prior,

$$\pi(\mu, \lambda) = N\left(\mu \mid m, \frac{1}{p\lambda}\right) \text{Ga}(\lambda \mid a, b),$$

where $m \in \mathbb{R}$ and $p, a, b > 0$ are known constants, and give explicit expressions for its parameters. (13 marks)

- (ii) The scientist may report a discrepancy (call this decision d_1) or an agreement (decision d_2) with the current theory supported by the data. If a real discrepancy is reported there is a good chance it would be published in a top journal; if the discrepancy is not real, her career would suffer a major drawback. After some consideration, she believes that the following loss function reflects well her preferences:

$$L(d_1, \mu) = \begin{cases} 10 & |\mu| \leq 0.01 \\ 0 & |\mu| > 0.01 \end{cases}, \quad L(d_2, \mu) = \begin{cases} 0 & |\mu| \leq 0.01 \\ 3 & |\mu| > 0.01 \end{cases}.$$

A set of $n = 250$ measures are taken and the following statistics are recorded:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = 0.02, \quad s_d^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = 1/4.$$

After elicitation, the scientist's prior parameters are $m = 0$, $p = 0.01$, $a = 2$ and $b = 0.6$. Should the scientist report a discrepancy?

[HINT: If $T \sim t_{254}$, then $P[T \leq 0.316] = 0.624$, $P[T \leq 0.675] = 0.75$, $P[T \leq 0.947] = 0.828$, $P[T \leq 1.651] = 0.95$ (15 marks)

- 2 A contract research organisation (CRO) have designed a clinical trial to assess the effectiveness of a TB drug. For each individual in the trial, the level of a specific T-cell biomarker is measured before and after administering the drug and registered as effective if the level has decreased and ineffective otherwise.
- (i) Let θ represent the effectiveness of the treatment. Assuming $\pi(\theta) = \text{Be}(\theta | a, b)$, write down the posterior distribution and provide explicit expressions for the posterior parameters. *(4 marks)*
- (ii) During Phase II of the trial, 65 patients were treated with 42 improving. The CRO medical advisor believes that with probability 0.95 the effectiveness of the treatment lies in $(0.51, 0.99)$.
- (a) Provide posterior point estimates of the treatment effectiveness under quadratic and zero-one loss and provide an posterior interval of approximate probability 0.95. *(8 marks)*
- (b) The CRO want to benchmark their analysis and asks you to use Jeffreys' as a minimum-informative prior. Compare the posterior mean and approximate probability intervals and comment on the differences if any. *(9 marks)*
- (iii) The CRO are considering extending this phase of the trial, but would like to have an assessment of how worthwhile this could be. The manager thinks that if the treatment is ineffective on the next three new patients, the trial should not continue. Based on the current data, what would you advise? *(7 marks)*

- 3 Two devices are used to determine the weight of the same atomic particle. Each device has a different accuracy but both provide unbiased measurements. A set of n measurements are obtained from the first device, $\mathbf{x} = \{x_1, \dots, x_n\}$, and m from the second, $\mathbf{y} = \{y_1, \dots, y_m\}$.

It is assumed $x_i \sim N(x_i | \mu, 1/p)$ and $y_i \sim N(y_i | \mu, 1/t)$, where $p = 3$ and $t = 1.5$ are the corresponding known measuring precisions and μ the unknown weight.

- (i) Show that $\pi(\mu) = N(\mu | c, 1/q)$, where $c \in \mathbb{R}$ and $q > 0$ are the prior mean and precision, respectively, is a conjugate prior and provide explicit expressions for the posterior parameters. *(9 marks)*
- (ii) Using the statistics from the experiment: $n = 12$, $\bar{x} = 1.6$, $m = 14$, $\bar{y} = 1.3$ and $c = 7.5$, $q = 0.1$:
- (a) Calculate the Bayes point estimate under the linear loss

$$\mathcal{L}(d, \mu) = \begin{cases} 3(d - \mu) & d \leq \mu \\ (\mu - d) & d > \mu \end{cases}$$

and provide a highest posterior density (HPD) interval of probability 0.9.

[HINT. If $Z \sim N(z | 0, 1)$: $P[Z \leq 0.674] = 0.75$, $P[Z \leq 1.281] = 0.9$, $P[Z \leq 1.645] = 0.95$, $P[Z \leq 1.96] = 0.975$, $P[Z \leq 2.576] = 0.995$] *(13 marks)*

- (b) Does the data provide evidence in favour of the expected weight being smaller than 1.5? *(6 marks)*

4 Consider the regression model,

$$y_i = \mu + \beta x_i + \varepsilon_i ; \quad i = 1, \dots, n$$

with $\varepsilon_i \sim N(\varepsilon_i | 0, 1/\lambda_i)$, independent, and prior structure

$$\begin{aligned} \lambda_i &\sim \text{Ga}(\lambda_i | a, \delta) ; \quad \text{independent for } i = 1, \dots, n \\ \mu &\sim N(\mu | m, 1/p) , \quad \beta \sim N(\beta | b, 1/t) \quad \text{and} \quad \delta \sim \text{Ex}(\delta | d) \end{aligned}$$

where a, m, p, b, t and d are known constants.

- (i) Show that the full conditional of:
- (a) Each of the individual precisions, λ_i , is Gamma and provide explicit expressions for the parameters. *(5 marks)*
 - (b) The intercept, μ , is Gaussian and provide explicit expressions for the parameters. *(5 marks)*
 - (c) The regression slope, β , is Gaussian and provide explicit expressions for the parameters. *(5 marks)*
 - (d) The precision hyperparameter, δ , is Gamma and provide explicit expressions for the parameters. *(5 marks)*
- (ii) Write pseudo-code for an MCMC sampling scheme for exploring the posterior distribution. *(8 marks)*

End of Question Paper

Notation and distributions

Bayesian Statistics 2018–19

Throughout the course it is assumed that the probabilistic behaviour of available data, \mathbf{x} , is described by a parametric model; hence all inferences will be conditional to the selected model.

Each model is composed by a family of probability distributions, indexed by a parameter vector, $\boldsymbol{\theta}$, which in turn can be described by their appropriate **probability density function** (pdf). We will denote a specific model by

$$\mathcal{M} = \{f(\mathbf{x} | \boldsymbol{\theta}), \mathbf{x} \in \mathcal{X}, \boldsymbol{\theta} \in \Theta\},$$

where

$$f(\mathbf{x} | \boldsymbol{\theta}) \geq 0 \quad \text{and} \quad \int_{\mathcal{X}} f(\mathbf{x} | \boldsymbol{\theta}) \, d\mathbf{x} = 1;$$

when there is no risk of confusion, we will refer to a model simply as $f(\mathbf{x} | \boldsymbol{\theta})$. We call \mathcal{X} the **support of the distribution** and Θ the **parameter space**.

We will use $f(\mathbf{x} | \boldsymbol{\phi})$ and $f(\mathbf{y} | \boldsymbol{\psi})$ to refer to probability densities of \mathbf{x} and \mathbf{y} , without necessarily meaning that both quantities share a common distribution. In general, the Greek alphabet is reserved for non-observables (typically, parameters) and the Latin alphabet for observations (data). Bold typeface denotes vector valued quantities.

Specific density functions are referred by appropriate names; e.g. if the observable x follows a Gaussian distribution with mean μ and variance σ^2 , its density is denoted by $N(x | \mu, \sigma^2)$. Tables below present some density functions used throughout the course.

Moments and other descriptive measures of probability distributions are described by appropriate symbols. Thus,

$$\mathbb{E}[\mathbf{x} | \boldsymbol{\theta}] = \int_{\mathcal{X}} \mathbf{x} f(\mathbf{x} | \boldsymbol{\theta}) \, d\mathbf{x},$$

$$\mathbb{V}[\mathbf{x} | \boldsymbol{\theta}] = \int_{\mathcal{X}} (\mathbf{x} - \mathbb{E}[\mathbf{x} | \boldsymbol{\theta}])^2 f(\mathbf{x} | \boldsymbol{\theta}) \, d\mathbf{x},$$

$$\text{Cov}[\mathbf{x} | \boldsymbol{\theta}] = \int_{\mathcal{X}} (\mathbf{x} - \mathbb{E}[\mathbf{x} | \boldsymbol{\theta}])' (\mathbf{x} - \mathbb{E}[\mathbf{x} | \boldsymbol{\theta}]) f(\mathbf{x} | \boldsymbol{\theta}) \, d\mathbf{x},$$

respectively stand for the mean, variance and covariance of the given quantity, while $\text{Med}[\mathbf{x} | \boldsymbol{\theta}]$ and $\text{Mode}[\mathbf{x} | \boldsymbol{\theta}]$ denote the median and mode, respectively. Sums are used instead of integrals when the support of the random quantity is discrete.

We use, $\mathbf{t} = \mathbf{t}(\mathbf{x})$ to denote a generic statistic (typically sufficient) derived from observed data, $\mathbf{x} = \{x_1, \dots, x_n\}$; standard symbols are used for common statistics; thus,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{and} \quad s_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

denote the sample mean and variance, respectively; while $x_{(p)}$ stands for the p^{th} order statistic; in particular $x_{(1)}$ and $x_{(n)}$ respectively denote the minimum and maximum observed values.

DISCRETE DISTRIBUTIONS

Name	Notation	p.f. $p(x \theta)$	$\mathbb{E}[X \theta]$	$\mathbb{V}[X \theta]$	Applications	Comments
Bernoulli	$\text{Ber}(x \theta)$	$p(x) = \theta^x(1 - \theta)^{1-x}$ $\mathcal{X} = \{0, 1\}$ $\Theta = (0, 1)$	θ	$\theta(1 - \theta)$	Coins, trials.	Constituent of more complex distributions. Expt. with binary outcome: success w.p. θ and failure w.p. $1 - \theta$.
Binomial	$\text{Bi}(x n, \theta)$	$p(x) = \binom{n}{x}\theta^x(1 - \theta)^{n-x}$ $\mathcal{X} = \{0, 1, 2, \dots, n\}$ $\Theta = (0, 1)$	$n\theta$	$n\theta(1 - \theta)$	Sampling with replacement	$X \equiv$ no. successes in n ind. $\text{Ber}(x \theta)$ trials. $\text{Bi}(x 1, \theta) \equiv \text{Ber}(x \theta)$
Geometric	$\text{Ge}(x \theta)$	$p(x) = \theta(1 - \theta)^x$ $\mathcal{X} = 0, 1, 2, \dots$ $\Theta = (0, 1)$	$\frac{1 - \theta}{\theta}$	$\frac{1 - \theta}{\theta^2}$	Waiting times (for single events)	$X \equiv$ no. failures until 1st success in sequence of ind. $\text{Ber}(x \theta)$ trials. Alternative formulation in terms of $Y \equiv$ no. of trials to 1st success ($Y = X + 1$)
Negative binomial (Pascal)	$\text{NB}(x m, \theta)$	$p(x) = \binom{m+x-1}{x}\theta^m(1 - \theta)^x$ $\mathcal{X} = 0, 1, 2, \dots$ $\Theta = (0, 1)$	$\frac{m(1 - \theta)}{\theta}$	$\frac{m(1 - \theta)}{\theta^2}$	Waiting times (for compound events)	$X \equiv$ no. failures to m -th success in sequence of ind. $\text{Ber}(x \theta)$ trials. Generalisation of Geometric. $\text{NB}(x 1, \theta) \equiv \text{Ge}(x \theta)$
Hypergeometric	$\text{Hy}(x N, d, n)$ (not standard, esp. order of arguments)	$p(x) = \frac{\binom{d}{x}\binom{N-d}{n-x}}{\binom{N}{n}}$ $\mathcal{X} = \{a, a + 1, \dots, b\}$ $a = \max\{0, n + d - N\},$ $b = \min\{n, d\}$	$\frac{nd}{N}$	$\frac{nd}{N} \frac{N - n}{N - 1} \left(1 - \frac{d}{N}\right)$	Sampling without replacement	$X \equiv$ no. of defectives in sample of size n taken without replacement from population of size N of which d are defective. $\text{Bi}(x n, d/N)$ — a suitable approx if $n/N < 0.1$
Poisson	$\text{Po}(x \lambda)$	$p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$ $\mathcal{X} = 0, 1, 2, \dots$ $\Lambda = \mathbb{R}^+$	λ	λ	Counting (rare) events occurring at random in space or time	Arises empirically or via Poisson Process (PP) for counting events. For PP rate ν the no. of events in time $t \sim \text{Po}(x \nu t)$. Also as an approx. to the Binomial. $\text{Bi}(x n, \theta) \approx \text{Po}(x n\theta)$ if n large, θ small, and $n\theta = c$.

CONTINUOUS DISTRIBUTIONS

Name	Notation	p.d.f. $f(x \theta)$	$\mathbb{E}[X \theta]$	$\mathbb{V}[X \theta]$	Applications	Comments
Uniform	$Un(x \alpha, \beta)$	$f(x) = \frac{1}{\beta - \alpha}$ $\mathcal{X} = [\alpha, \beta]$ $\Theta = \{(\alpha, \beta) \in \mathbb{R}^2 : \alpha < \beta\}$	$\frac{\alpha + \beta}{2}$	$\frac{(\beta - \alpha)^2}{12}$	Rounding errors $Un(x -1/2, 1/2)$. Simulating other distributions from $Un(x 0, 1)$	Used as non-informative prior for parameters with bounded support.
Pareto	$Pa(x \alpha, \beta)$	$f(x) = \alpha\beta^\alpha x^{-(\alpha+1)}$ $\mathcal{X} = (\beta, \infty)$ $\Theta = \{(\alpha, \beta) \in \mathbb{R}^2 : \alpha > 0, \beta > 0\}$	$\frac{\alpha\beta}{\alpha - 1}$ (if $\alpha > 1$)	$\frac{\alpha\beta^2}{(\alpha - 2)(\alpha - 1)^2}$ (if $\alpha > 2$)	Distribution of positive random quantities with heavy tails	Conjugate prior for uniform data with known lower bound
Exponential	$Ex(x \lambda)$	$f(x) = \lambda e^{-\lambda x}$ $\mathcal{X} = \mathbb{R}_+$ $\Lambda = \mathbb{R}_+$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	Inter-event times for Poisson Process. Models lifetimes of non-ageing items.	Also parameterised in terms of $1/\lambda$. $Ga(x 1, \lambda) \equiv Ex(x \lambda)$
Gamma	$Ga(x \alpha, \beta)$	$f(x) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma[\alpha]}$ $\mathcal{X} = \mathbb{R}_+$ $\Theta = \{(\alpha, \beta) \in \mathbb{R}^2 : \alpha > 0, \beta > 0\}$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$	Times between k events for Poisson Process. Lifetimes of ageing items. Conjugate prior for exponential model.	Also parameterised in terms of $1/\beta$ $Ga(x 1, \lambda) \equiv Ex(x \lambda)$, $1/x = y \sim IGa(y \alpha, \beta)$
Beta	$Be(x \alpha, \beta)$	$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$ $\mathcal{X} = (0, 1)$ $\Theta = \{(\alpha, \beta) \in \mathbb{R}^2 : \alpha > 0, \beta > 0\}$	$\mu = \frac{\alpha}{\alpha + \beta}$	$\frac{\mu(1-\mu)}{(\alpha + \beta + 1)}$	Useful model for variables with finite range. Conjugate prior for Binomial model.	$Be(x 1, 1) \equiv Un(x 0, 1)$ Can re-scale $Be(x \alpha, \beta)$ to any finite range (a, b) by $Y = (b - a)X + a$
Gaussian (Normal)	$N(x \mu, \sigma^2)$	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$ $\mathcal{X} = \mathbb{R}$ $\Theta = \{(\mu, \sigma^2) \in \mathbb{R}^2 : \sigma^2 > 0\}$	μ	σ^2	Empirically and theoretically (via CLT) a useful model. Also parameterised in terms of the precision $\lambda = 1/\sigma^2$	$Y = a + bX \sim N(y a + b\mu, b^2\sigma^2)$ $Z = \frac{X-\mu}{\sigma} \sim N(z 0, 1)$ $P[X \in (u, v)] = P\left[Z \in \left(\frac{u-\mu}{\sigma}, \frac{v-\mu}{\sigma}\right)\right]$
Student t	$St(x \mu, \lambda, \nu)$	$f(x) = \frac{\Gamma[(\nu + 1)/2]}{\Gamma[\nu/2]} \left(\frac{\lambda}{\nu\pi}\right)^{1/2} \times$ $\left(1 + \frac{\lambda}{\nu}(x - \mu)^2\right)^{-(\nu+1)/2}$ $\mathcal{X} = \mathbb{R}, \mu \in \mathbb{R}, \lambda, \nu > 0$	μ (if $\nu > 1$)	$\lambda^{-1} \frac{\nu}{\nu - 2}$ (if $\nu > 2$)	Useful alternative to Gaussian for random quantities with heavy tails	If $X \sim N(x 0, 1)$ and $Y \sim \chi^2_\nu(y)$ independent then $\frac{X}{\sqrt{Y/\nu}} \sim t_\nu$. If $Y = \sqrt{\lambda}(x - \mu)$ then $Y \sim t_\nu(y)$ $t_1 \equiv$ Cauchy. $t^2_\nu \equiv F_{1,\nu}$.

MULTIVARIATE DISTRIBUTIONS

Name	Notation	p.d.f. $f(\mathbf{x} \boldsymbol{\theta})$	$\mathbb{E}[X \boldsymbol{\theta}]$	$\mathbb{V}[X \boldsymbol{\theta}]$	Applications	Comments
Multinomial	$\text{Mu}(\mathbf{x} \boldsymbol{\theta}, n)$	$p(\mathbf{x}) = \frac{n!}{\prod_{l=1}^k x_l!} \prod_{l=1}^k \theta_l^{x_l}$ $\mathbf{x} = \{x_1, \dots, x_k\}, \quad x_l = 0, 1, \dots, \sum x_l = n$ $\boldsymbol{\theta} = \{\theta_1, \dots, \theta_k\}, \quad 0 < \theta_l < 1, \sum \theta_l = 1$	$\mathbb{E}[x_i] = n\theta_i$	$\mathbb{V}[x_i] = n\theta_i(1 - \theta_i)$ $\text{Cov}[x_i, x_j] = -n\theta_i\theta_j$	Counts of events with more than two possible outcomes	Generalisation of the Binomial distribution
Dirichlet	$\text{Di}(\mathbf{x} \boldsymbol{\alpha})$	$f(\mathbf{x}) = \frac{\Gamma(\sum \alpha_l)}{\prod \Gamma(\alpha_l)} \prod_{l=1}^k x_l^{\alpha_l - 1}$ $\mathbf{x} = \{x_1, \dots, x_k\}, \quad 0 < x_l < 1, \sum_{l=1}^k x_l = 1$ $\boldsymbol{\alpha} = \{\alpha_1, \dots, \alpha_k\}, \quad 0 < \alpha_l$	$\mathbb{E}[x_i] = \mu_i$ $= \frac{\alpha_i}{\sum \alpha_l}$	$\mathbb{V}[x_i] = \frac{\mu_i(1 - \mu_i)}{1 + \sum \alpha_l}$ $\text{Cov}[x_i, x_j] = -\frac{\mu_i\mu_j}{1 + \sum \alpha_l}$	Distribution of probabilities of exclusive events.	Generalisation of the Beta distribution. Conjugate prior for multinomial data
Normal-Gamma	$\text{NG}(x, y \mu, \kappa, \alpha, \beta)$	$f(x, y) = \text{N}(x \mu, (y\kappa)^{-1}) \text{Ga}(y \alpha, \beta)$ $\mathcal{X} = \{(x, y) : x \in \mathbb{R}, y > 0\}$ $\mu \in \mathbb{R}; \kappa, \alpha, \beta > 0$	$\mathbb{E}[x] = \mu$ $\mathbb{E}[y] = \frac{\alpha}{\beta}$	$\mathbb{V}[x] = \frac{\beta}{\kappa(\alpha - 1)}$ $\mathbb{V}[y] = \frac{\alpha}{\beta^2}$	Conjugate prior for Gaussian data, both parameters unknown	The marginal distribution of x is $\text{St}(x \mu, \kappa\alpha/\beta, 2\alpha)$
(Multivariate) Gaussian	$\text{N}_k(\mathbf{x} \boldsymbol{\mu}, \Lambda)$	$f(\mathbf{x}) = \frac{ \Lambda ^{1/2}}{(2\pi)^{k/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})' \Lambda (\mathbf{x} - \boldsymbol{\mu})\right]$ $\mathcal{X} = \mathbb{R}^k$ $\boldsymbol{\mu} \in \mathbb{R}^k; \Lambda \text{ symmetric positive-definite}$	$\boldsymbol{\mu}$	Λ^{-1}	See univariate case	Usually parameterised in terms of the covariance matrix $\Sigma = \Lambda^{-1}$
(Multivariate) Student	$\text{St}_k(\mathbf{x} \boldsymbol{\mu}, \Lambda, \nu)$	$f(\mathbf{x}) = \frac{ \Lambda ^{1/2} \Gamma((\nu + k)/2)}{(\nu\pi)^{k/2} \Gamma(\nu/2)} \times$ $\left[1 + \frac{1}{\nu}(\mathbf{x} - \boldsymbol{\mu})' \Lambda (\mathbf{x} - \boldsymbol{\mu})\right]^{-(\nu+k)/2}$ $\mathcal{X} = \mathbb{R}^k$ $\boldsymbol{\mu} \in \mathbb{R}^k; \Lambda \text{ symmetric positive-definite}, \nu > 0$	$\boldsymbol{\mu}$ (if $\nu > 1$)	$\frac{\nu}{\nu - 2} \Lambda^{-1}$ (if $\nu > 2$)	See univariate case	Usually parameterised in terms of the covariance matrix $\Sigma = \Lambda^{-1}$
Wishart	$\text{Wi}_k(X \alpha, \Omega)$	$f(X) = \frac{(\pi)^{k(k-1)} \Omega ^\alpha}{\prod_{i=1}^k \Gamma[(2\alpha + 1 - i)/2]} \times$ $ X ^{\alpha - (k+1)/2} \exp[-\text{tr}(\Omega X)]$ $\mathcal{X} = \text{symmetric positive-definite}$ $\alpha > (k - 1)/2; \Omega \text{ symmetric non-singular}$	$\alpha \Omega^{-1}$	$\mathbb{V}[X_{ij}] = \alpha (\omega_{ij}^2 + \omega_{ii}\omega_{jj})$	Conjugate prior for the precision matrix in a Gaussian model	Can also be used for the covariance matrix after the appropriate transformation.