



The
University
Of
Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS

**Autumn Semester
2015–16**

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 In microscopic imaging it is common to model the number of photons arriving at the lens in each frame, X , as $\text{Po}(x | \lambda)$, where λ is the rate of photon emission per frame. Given a random sample, $\mathbf{x} = \{x_1, \dots, x_n\}$,

- (i) (a) Show that $\pi(\lambda) = \text{Ga}(\lambda | a, b)$ is a conjugate prior and give explicit expressions for the posterior parameters. (5 marks)
- (b) Find the Bayes estimator for λ under 0-1 loss. (3 marks)
- (ii) (a) Calculate the predictive distribution of Y , the number of photons captured by the lens in the next random sample of m frames,

$$Y = \sum_{j=n+1}^{n+m} X_j$$

(11 marks)

- (b) The scientist a priori believes that $\mathbb{E}[\lambda] = 10/3$ and $\mathbb{V}[\lambda] = 50/9$. Calculate the scientist's probability of observing not more than one photon in the next frame if 3 photons were detected in a sample of $n = 10$ frames. (9 marks)

2 Assume

$$X_i \sim N\left(x_i \mid \mu, \frac{1}{a_i \lambda}\right),$$

independent for $i = 1, \dots, n$, where $\mathbf{a} = \{a_1, \dots, a_n\}$ are known constants with

$$0 < a_i < 1 \text{ and } \sum_{i=1}^n a_i = 1.$$

- (i) Show that

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

is a conjugate prior and provide explicit expressions for the posterior parameters. (15 marks)

- (ii) Show that

$$\mathbb{E}[\mu | \mathbf{x}] = w\hat{\mu} + (1 - w)m,$$

where $0 < w < 1$ and $\hat{\mu} = \sum_{i=1}^n a_i x_i$ is the MLE. (5 marks)

- (iii) Find the posterior Bayes estimator of $\sigma^2 = \lambda^{-1}$ under quadratic loss. (8 marks)

- 3 A chemist is interested in the (relative) molecular weight of a new compound. She sends samples to n different labs and collects the measurements $W = \{W_1, \dots, W_n\}$. Given that each lab has a different weighing instrument, she thinks it is sensible to assume $W_i \sim N(w_i | \mu, 1/\lambda_i)$, where μ is the actual weight and $\lambda = \{\lambda_1, \dots, \lambda_n\}$ are the known measuring precisions.

[Additional information: Assume $Z \sim N(z | 0, 1)$ and let $\Phi(x) = P[Z < x]$, then $\Phi(-1.96) = 0.025$, $\Phi(-1.645) = 0.05$, $\Phi(-1.26) = 0.104$ and $\Phi(1.521) = 0.936$.]

- (i) Write down the likelihood and show that

$$\hat{\mu} = \frac{\sum \lambda_i w_i}{\sum \lambda_i}$$

is the MLE. (7 marks)

- (ii) From previous stoichiometry analyses she believes $P[\mu > 0.1] = 0.5$ and $P[0.02 < \mu < 0.18] = 0.9$ and is willing to use a Gaussian distribution to express her uncertainty.

- (a) Use the scientist's prior opinions to elicit her prior. (7 marks)

- (b) The new compound could potentially be used in drug production if its molecular weight is within $(0.1, 0.2)$, but it could be risky to use otherwise. After consultation, she thinks that her preferences can be described by

$$\mathcal{L}(a_1, \mu) = \begin{cases} 0 & \mu \in (0.1, 0.2) \\ 9 & \mu \notin (0.1, 0.2) \end{cases} \quad \mathcal{L}(a_2, \mu) = \begin{cases} 2 & \mu \in (0.1, 0.2) \\ 0 & \mu \notin (0.1, 0.2) \end{cases}$$

where $a_1 = \text{use the compound}$ and $a_2 = \text{do not use the compound}$. Find her optimal decision if $\hat{\mu} = 0.2$ and $\sum \lambda_i = 350$ are recorded from a random sample of size $n = 100$. (14 marks)

- 4 Consider the hierarchical model,

$$\begin{aligned} X_i &\sim \text{Ber}(x_i | \theta_i), \text{ ind. } i = 1, \dots, n \\ \pi(\theta_i) &= \text{Be}(\theta_i | a, a), \text{ ind. } i = 1, \dots, n \\ \pi(a) &= \text{Ga}(a | c, d), \text{ with } \mathbb{E}[a] = \frac{c}{d}. \end{aligned}$$

- (i) Write down the full conditional distributions for $\theta = \{\theta_1, \dots, \theta_n\}$ and a . (13 marks)
- (ii) Write pseudo-code for a Metropolis-within-Gibbs strategy to sample from $\pi(\theta, a | x)$. (15 marks)

End of Question Paper

Notation and distributions

Bayesian Statistics 2015–16

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 density functions. We will denote a specific model by

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

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

$$\begin{aligned}\mathbb{E}[\mathbf{x} | \boldsymbol{\theta}] &= \int_{\mathcal{X}} \mathbf{x} f(\mathbf{x} | \boldsymbol{\theta}) d\mathbf{x}, \\ \mathbb{V}[\mathbf{x} | \boldsymbol{\theta}] &= \int_{\mathcal{X}} (\mathbf{x} - \mathbb{E}[\mathbf{x} | \boldsymbol{\theta}])^2 f(\mathbf{x} | \boldsymbol{\theta}) d\mathbf{x}, \\ \text{Cov}[\mathbf{x} | \boldsymbol{\theta}] &= \int_{\mathcal{X}} (\mathbf{x} - \mathbb{E}[\mathbf{x} | \boldsymbol{\theta}])^t (\mathbf{x} - \mathbb{E}[\mathbf{x} | \boldsymbol{\theta}]) f(\mathbf{x} | \boldsymbol{\theta}) d\mathbf{x},\end{aligned}$$

respectively stand for the expected value, 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.

SOME DISCRETE DISTRIBUTIONS

Name	Context	Notation	p.f. $p(x \theta)$	$\mathbb{E}[X \theta]$	$\mathbb{V}[X \theta]$	Applications	Comments
Uniform	Set of k equally likely outcomes (usually, not necessarily, the integers)	$U(1, \dots, k)$	$p(x) = 1/k$ $\mathcal{X} = \{1, \dots, k\}, \mathcal{K} = \mathbb{Z}_+$	$\frac{k+1}{2}$	$\frac{k^2-1}{12}$	Dice	
Bernoulli	Expt. with two outcomes: 'success' w.p. θ and 'failure' w.p. $1 - \theta$ $X \equiv$ no. successes	$\text{Ber}(x \theta)$	$p(x) = \theta^x(1 - \theta)^{1-x}$ $\mathcal{X} = \{0, 1\}$ $\Theta = (0, 1)$	θ	$\theta(1 - \theta)$	Coins, constituent of more complex distributions	
Binomial	$X \equiv$ no. successes in n ind. $\text{Ber}(x \theta)$ trials	$\text{Bi}(x n, \theta)$	$p(x) = \binom{n}{x}\theta^x(1 - \theta)^{n-x}$ $\mathcal{X} = \{0, 1, 2, \dots, n\}$ $\Theta = (0, 1)$	$n\theta$	$n\theta(1 - \theta)$	Sampling with replacement	$\text{Bi}(x 1, \theta) \equiv \text{Ber}(x \theta)$
Geometric	$X \equiv$ no. failures until 1st success in sequence of ind. $\text{Ber}(x \theta)$ trials	$\text{Ge}(x \theta)$	$p(x) = \theta(1 - \theta)^x$ $\mathcal{X} = 0, 1, 2, \dots$ $\Theta = (0, 1)$	$\frac{1 - \theta}{\theta}$	$\frac{1 - \theta}{\theta^2}$	Waiting times (for single events)	Alternative formulation in terms of $Y \equiv$ no. of trials to 1st success ($Y = X + 1$)
Negative binomial (or Pascal)	$X \equiv$ no. failures to m -th success in sequence of ind. $\text{Ber}(x \theta)$ trials. Generalisation of Geometric	$\text{NB}(x m, \theta)$	$p(x) = \binom{m+x-1}{x}\theta^m(1 - \theta)^x$ $\mathcal{X} = 0, 1, 2, \dots$ $\Theta = (0, 1)$	$\frac{m(1 - \theta)}{\theta}$	$\frac{m(1 - \theta)}{\theta^2}$	Waiting times (for compound events)	$\text{NB}(x 1, \theta) \equiv \text{Ge}(x \theta)$
Poisson	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	$\text{Po}(x \lambda)$	$p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$ $\mathcal{X} = 0, 1, 2, \dots$ $\Lambda = \mathbb{R}^+$	λ	λ	Counting events occurring 'at random' in space or time	$\text{Bi}(x n, \theta) \equiv \text{Po}(x n\theta)$ if n large, θ small

SOME CONTINUOUS DISTRIBUTIONS

Name	Notation	p.d.f. $f(x \theta)$	$\mathbb{E}[X \theta]$	$\mathbb{V}[X \theta]$	Applications	Comments
Uniform	$\text{Un}(x \alpha, \beta)$	$f(x) = \frac{1}{\beta - \alpha}$ $\mathcal{X} = [\alpha, \beta]$ $\Theta = \{(\alpha, \beta) \in \mathbb{R}^2 : \alpha < \beta\}$	$\frac{\alpha + \beta}{2}$	$\frac{(\beta - \alpha)^2}{12}$	Rounding errors $\text{Un}(x -1/2, 1/2)$. Simulating other distributions from $\text{Un}(x 0, 1)$	
Exponential	$\text{Ex}(x \lambda)$	$f(x) = \lambda e^{-\lambda x}$ $\mathcal{X} = \mathbb{R}_+$ $\Lambda = \mathbb{R}_+$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	Inter-event times for Poisson Process. Models lifetimes of non-ageing items.	Also parameterised in terms of $1/\lambda$. $\text{Ga}(x 1, \lambda) \equiv \text{Ex}(x \lambda)$
Gamma	$\text{Ga}(x \alpha, \beta)$	$f(x) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma[\alpha]}$ $\mathcal{X} = \mathbb{R}_+$ $\Theta = \{(\alpha, \beta) \in \mathbb{R}^2 : \alpha > 0, \beta > 0\}$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$	Times between k events for Poisson Process. Lifetimes of ageing items.	Also parameterised in terms of $1/\beta$ $\text{Ga}(x 1, \lambda) \equiv \text{Ex}(x \lambda)$, $\text{Ga}(x \nu/2, 1/2) \equiv \chi_{(\nu)}^2(x)$
Beta	$\text{Be}(x \alpha, \beta)$	$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\text{B}(\alpha, \beta)}$ $\mathcal{X} = (0, 1)$ $\Theta = \{(\alpha, \beta) \in \mathbb{R}^2 : \alpha > 0, \beta > 0\}$	$\frac{\alpha}{\alpha + \beta}$	$\frac{\alpha\beta(\alpha + \beta)^{-2}}{(\alpha + \beta + 1)}$	Useful model for variables with finite range. Also as a Bayesian conjugate prior.	$\text{Be}(x 1, 1) \equiv \text{Un}(x 0, 1)$ $\text{Be}(x \alpha, \beta)$ is reflection about $\frac{1}{2}$ of $\text{Be}(x \beta, \alpha)$. Can re-scale $\text{Be}(x \alpha, \beta)$ to any finite range $[a, b]$ by $Y = (b - a)X + a$
Normal (Gaussian)	$\text{N}(x \mu, \sigma^2)$	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right]$ $\mathcal{X} = \mathbb{R}$ $\Theta = \{(\mu, \sigma^2) \in \mathbb{R}^2 : \sigma^2 > 0\}$	μ	σ^2	Empirically and theoretically (via CLT) a useful model. Often parameterised in terms of the precision $\lambda = 1/\sigma^2$	$Y = aX + b \sim \text{N}(y a\mu + b, a^2\sigma^2)$ $Z = \frac{X - \mu}{\sigma} \sim \text{N}(z 0, 1)$ $\text{P}[X \in (u, v)] = \text{P}\left[Z \in \left(\frac{u - \mu}{\sigma}, \frac{v - \mu}{\sigma}\right)\right]$
Chi-square	$\chi_{(\nu)}^2(x)$	$f(x) = \frac{2^{-\nu/2}}{\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2}$ $\mathcal{X} = \mathbb{R}_+$; $\Theta = \mathbb{R}_+$	ν	2ν	Sum of squares of ν independent standard Gaussians	$\chi_{(\nu)}^2(x) \equiv \text{Ga}(x \nu/2, 1/2)$
Student t	$\text{St}(x \mu, \lambda, \nu)$	$f(x) = \frac{\Gamma[(\nu+1)/2]}{\Gamma[\nu/2]} \left(\frac{\lambda}{\nu\pi}\right)^{1/2} \times$ $(1 + \lambda(x - \mu)^2/\nu)^{-(\nu+1)/2}$ $\mathcal{X} = \mathbb{R}, \mu \in \mathbb{R}, \lambda, \nu > 0$	μ (if $\nu > 1$)	$\lambda^{-1} \frac{\nu}{\nu - 2}$ (if $\nu > 2$)	Useful alternative to Gaussian for variables with heavy tails.	If $X \sim \text{N}(x 0, 1)$ and $Y \sim \chi_{(\nu)}^2(y)$ independent then $\frac{X}{\sqrt{Y/\nu}} \sim t_\nu$. If $Y = \sqrt{\lambda}(x - \mu)$ then $Y \sim t_\nu(y)$ $t_1 \equiv \text{Cauchy}$. $t_\nu^2 \equiv F_{1,\nu}$.

SOME MULTIVARIATE DISTRIBUTIONS

Name	Notation	p.d.f. $f(x \theta)$	$\mathbb{E}[X \theta]$	$\mathbb{V}[X \theta]$	Applications	Comments
Multinomial	$\text{Mu}(\mathbf{x} \boldsymbol{\theta}, n)$	$p(\mathbf{x}) = \frac{n!}{\prod_{l=1}^k x_l!} \prod_{l=1}^k \theta_l^{x_l}$ $\mathbf{x} = \{x_1, \dots, x_k\}, x_l = 0, 1, \dots, \sum x_l = n$ $\boldsymbol{\theta} = \{\theta_1, \dots, \theta_k\}, 0 < \theta_l < 1, \sum \theta_l = 1$	$\mathbb{E}[x_i] = n\theta_i$	$\mathbb{V}[x_i] = n\theta_i(1 - \theta_i)$ $\text{Cov}[x_i, x_j] = -n\theta_i\theta_j$	Counts of events with more than two possible outcomes	Generalisation of the Binomial distribution
Dirichlet	$\text{Di}(\mathbf{x} \boldsymbol{\alpha})$	$f(\mathbf{x}) = \frac{\Gamma(\sum \alpha_l)}{\prod \Gamma(\alpha_l)} \prod x_l^{\alpha_l - 1}$ $\mathbf{x} = \{x_1, \dots, x_k\}, 0 < x_l < 1, \sum_{l=1}^k x_l = 1$ $\boldsymbol{\alpha} = \{\alpha_1, \dots, \alpha_k\}, 0 < \alpha_l$	$\mathbb{E}[x_i] = \mu_i = \frac{\alpha_i}{\sum \alpha_l}$	$\mathbb{V}[x_i] = \frac{\mu_i(1 - \mu_i)}{1 + \sum \alpha_l}$ $\text{Cov}[x_i, x_j] = -\frac{\mu_i\mu_j}{1 + \sum \alpha_l}$	Distribution of points in a simplex	Generalisation of the Beta distribution
Normal-Gamma	$\text{NG}(x, y \mu, \lambda, \alpha, \beta)$	$f(x, y) = \text{N}(x \mu, (y\lambda)^{-1})\text{Ga}(y \alpha, \beta)$ $\mathcal{X} = \{(x, y) : x \in \mathbb{R}, y > 0\}$ $\mu \in \mathbb{R}; \lambda, \alpha, \beta > 0$	$\mathbb{E}[x] = \mu$ $\mathbb{E}[y] = \alpha\beta^{-1}$	$\mathbb{V}[x] = \frac{\beta}{\lambda(\alpha - 1)}$ $\mathbb{V}[y] = \alpha\beta^{-2}$	Conjugate prior for Gaussian data	$f(x) = \text{St}(x \mu, \lambda\alpha\beta^{-1}, 2\alpha)$
Gaussian	$\text{N}_k(\mathbf{x} \boldsymbol{\mu}, \Lambda)$	$f(\mathbf{x}) = \frac{ \Lambda ^{1/2}}{(2\pi)^{k/2}} \exp[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})' \Lambda (\mathbf{x} - \boldsymbol{\mu})]$ $\mathcal{X} = \mathbf{x} \in \mathbb{R}^k$ $\boldsymbol{\mu} \in \mathbb{R}^k; \Lambda$ symmetric positive-definite	$\boldsymbol{\mu}$	Λ^{-1}	See univariate case	Usually parameterised in terms of the covariance matrix $\Sigma = \Lambda^{-1}$
Student	$\text{St}_k(\mathbf{x} \boldsymbol{\mu}, \Lambda, \nu)$	$f(\mathbf{x}) = \frac{ \Lambda ^{1/2} \Gamma((\nu + k)/2)}{(\nu\pi)^{k/2} \Gamma(\nu/2)} \times$ $\left[1 + \frac{1}{\nu} (\mathbf{x} - \boldsymbol{\mu})' \Lambda (\mathbf{x} - \boldsymbol{\mu}) \right]^{-(\nu+k)/2}$ $\mathcal{X} = \mathbf{x} \in \mathbb{R}^k$ $\boldsymbol{\mu} \in \mathbb{R}^k; \Lambda$ 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 $\Sigma = \Lambda^{-1}$