Max Planck Institute for Gravitational Physics
IMPRS Lecture Series

Making sense of data: introduction to statistics for gravitational wave astronomy

Problem Sheet 2: Bayesian Statistics

1. The probability that he chose route i given the observation that the journey took less than 1 hour is given by Bayes' Theorem

$$p(i|T<1\text{ hr}) = \frac{p(T<1\text{ hr}|i)p(i)}{\sum_j p(T<1\text{ hr}|j)p(j)}.$$

He chooses one of the four routes at random, so $p_i = 0.25$ for $i=1,\ldots 4$. Hence

$$\begin{aligned}
p(1|T<1\text{ hr}) &= \frac{0.2}{0.2+0.5+0.8+0.9} = 0.083\\
p(2|T<1\text{ hr}) &= \frac{0.5}{0.2+0.5+0.8+0.9} = 0.208\\
p(3|T<1\text{ hr}) &= \frac{0.6}{0.2+0.5+0.8+0.9} = 0.333\\
p(4|T<1\text{ hr}) &= \frac{0.9}{0.2+0.5+0.8+0.9} = 0.375.
\end{aligned}$$

2. (a) The log-likelihood is

$$l(\mu|\mathbf{x},\sigma) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{n}{2}\ln(2\pi\sigma^2).$$

The second derivative of the low-likelihood with respect to μ is therefore

$$\frac{\partial^2 l}{\partial\mu^2} = -\frac{n}{\sigma^2}$$

and the Fisher matrix is

$$I_\mu = -\mathbb{E}\left[\frac{\partial^2 l}{\partial\mu^2}\right] = \frac{n}{\sigma^2} \propto 1$$

hence the Jeffreys prior, $p(\mu)\propto\sqrt{I_\mu}\propto 1$, as required.

(b) The posterior distribution, using the Jeffreys prior is

$$\begin{aligned}
p(\mu|\mathbf{x}) &\propto \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_i-\mu)^2\right]\\
&\propto \exp\left[-\frac{n}{2\sigma^2}\left(\mu-\frac{1}{n}\sum_{i=1}^{n}x_i\right)^2\right] \qquad (1)
\end{aligned}$$

where we have dropped factors that are independent of μ. The μ-dependence in the above is the μ dependence of a Normal distribution and so we deduce

$$p(\mu|\mathbf{x}) \sim N\left(\frac{1}{n}\sum_{i=1}^{n}x_i, \frac{\sigma^2}{n}\right).$$

(c) Using the previous result, the posterior is $N(10.1, 0.1)$. A 95% HPD confidence interval is

$$[10.1 - 1.96\sqrt{0.1}, 10.1 + 1.96\sqrt{0.1}] = [9.480, 10.720]$$

as required.

3. (a) The posterior distribution is proportional to

$$p(\theta|\mathbf{x}) \propto \begin{cases} \frac{aX^a}{\theta^{a+n+1}} & \text{for } \theta \geq X \\ 0 & \text{otherwise} \end{cases}$$

where $X = \max\{x_0, x_1, \ldots, x_n\}$. Hence, the posterior is a Pareto distribution with parameters $A = a + n$ and X.

(b) Based on this observed data the posterior is a Pareto distribution with parameters $a = 5$ and $x_0 = 17$. The posterior mean is 21.25, compared to the prior mean of 0.2. The posterior median is 19.53 compared to the prior median of 1.414. The posterior variance is $ax_0^2/((a-1)^2(a-2)) = 30.1$ compared to the prior variance which is divergent.

(c) This prior is incompatible with the observed data, since it implies $\theta \leq 15$ and therefore no data values should be observed with $x \geq 15$. The observation $x_3 = 17$ violates this condition. Observing this data would tell the chemist that they were too restrictive in their prior specification and so they should revise it.

4. (a) From a simple application of Bayes Theorem, the posterior is

$$\mathbb{P}(H_0|\mathbf{x}) = \frac{\mathbb{P}(\mathbf{x}|H_0)p_0}{\mathbb{P}(\mathbf{x}|H_0)p_0 + \mathbb{P}(\mathbf{x}|H_1)p_1} = \frac{p_0}{p_0 + [\mathbb{P}(\mathbf{x}|H_1)/\mathbb{P}(\mathbf{x}|H_0)]p_1} = \frac{p_0}{p_0 + p_1/B_{01}}$$

where $B_{01} = \mathbb{P}(\mathbf{x}|H_0)/\mathbb{P}(\mathbf{x}|H_1)$ is the Bayes factor in favour of H_0 over H_1.

(b) The likelihood under hypothesis H_i is

$$\begin{aligned} \mathbb{P}(\mathbf{x}|H_i) &= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j - \mu_i)^2\right] \\ &= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{j=1}^{n} x_j^2 - 2\mu_i \sum_{j=1}^{n} x_j + n\mu_i^2\right)\right]. \end{aligned} \tag{2}$$

Hence, denoting $\bar{x} = \sum_{j=1}^{n} x_j/n$ as usual, we deduce

$$\begin{aligned} B_{01} = \frac{\mathbb{P}(\mathbf{x}|H_0)}{\mathbb{P}(\mathbf{x}|H_1)} &= \exp\left[-\frac{n}{2\sigma^2}\left(-2(\mu_0 - \mu_1)\bar{x} + \mu_0^2 - \mu_1^2\right)\right] \\ &= \exp\left[-\frac{n}{2\sigma^2}(\mu_0 - \mu_1)(\mu_0 + \mu_1 - 2\bar{x})\right] \end{aligned} \tag{3}$$

as required. Setting $\mu_0 = 0$, $\mu_1 = 1$, $\sigma^2 = 1$, $n = 9$ and $\bar{x} = 0.645$ we find

$$B_{01} = \exp(-4.5 \times (-1) \times (-0.29)) = \exp(-1.305) = 0.271.$$

There is weak evidence against the null hypothesis, with the posterior probability for H_0 given the observed data and equal prior weights on the two hypotheses, of 21%. As n increases, with all other values fixed, the evidence against H_0 increases. For $n \geq 21$ $\mathbb{P}(H_0|\mathbf{x}) < 0.05$ and so you would reject the null hypothesis at the 5% level.

(c) We need to recalculate $\mathbb{P}(\mathbf{x}|H_1)$, which is done as follows

$$\begin{aligned}
\mathbb{P}(\mathbf{x}|H_1) &= (2\pi\sigma^2)^{-n/2}(2\pi\tau^2)^{-1/2}\int_{-\infty}^{\infty}\exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j-\mu)^2-\frac{1}{2\tau^2}\mu^2\right]\mathrm{d}\mu \\
&= (2\pi\sigma^2)^{-n/2}(2\pi\tau^2)^{-1/2}\exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{n}x_j^2+\frac{n^2\Sigma^2\bar{x}^2}{2\sigma^4}\right]\times \\
&\qquad\times\int_{-\infty}^{\infty}\exp\left[-\frac{1}{2\Sigma^2}\left(\mu-\frac{n\Sigma^2}{\sigma^2}\bar{x}\right)^2\right]\mathrm{d}\mu \\
&= (2\pi\sigma^2)^{-n/2}\frac{\Sigma}{\tau}\exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{n}x_j^2+\frac{n^2\Sigma^2\bar{x}^2}{2\sigma^4}\right] \qquad (4)
\end{aligned}$$

where $\Sigma^{-2} = n\sigma^{-2}+\tau^{-2}$. The Bayes factor then becomes

$$B_{01} = \frac{\tau}{\Sigma}\exp\left[-\frac{n}{2\sigma^2}(\mu_0^2-2\mu_0\bar{x})-\frac{n^2\Sigma^2\bar{x}^2}{2\sigma^4}\right].$$

In the limit $\tau\to\infty$ we have $\Sigma^2\to\sigma^2/n$ and

$$B_{01}\to\frac{\tau}{\Sigma}\exp\left[-\frac{n}{2\sigma^2}(\mu_0-\bar{x})^2\right]\to\infty.$$

In the limit as $\tau\to\infty$, there is a lot of prior weight to arbitrarily large means. Any finite $\bar{x}$ favours means close to $\bar{x}$, so for large τ, such means are more consistent with the null hypothesis than the alternative and we never reject H_0. The moral is — don't be too generic in your prior specification!

5. (a) The posterior takes the form

$$p(\mathbf{p}|\mathbf{x}) \propto p(\mathbf{x}|\mathbf{p})p(\mathbf{p}) \propto \prod_{i=1}^{m}p_i^{x_i}\prod_{j=1}^{m}p_j^{\alpha_j-1} \propto \prod_{i=1}^{m}p_i^{\alpha_i+x_i-1}$$

and we deduce $p(\mathbf{p}|\mathbf{x}) \sim \text{Dir}(\alpha_1+x_1,\alpha_2+x_2,\ldots,\alpha_m+x_m)$.

(b) We showed in lectures that the Bayes estimator with a quadratic error loss is the posterior mean. For the Dirichlet distribution this is $\alpha_i/\sum_j\alpha_j$ and so in this case the Bayes estimate for the parameters is

$$\hat{p}_i = \frac{\alpha_i+x_i}{N+\sum_{j=1}^{m}\alpha_j}.$$

(c) The posterior means, and hence Bayes estimate with quadratic loss, are

$$\begin{aligned}
\hat{p}_1 &= \frac{11}{66} = \frac{1}{6} = 0.167, \qquad \hat{p}_2 = \frac{13}{66} = 0.197, \qquad \hat{p}_3 = \frac{13}{66} = 0.197, \\
\hat{p}_4 &= \frac{9}{66} = 0.136, \qquad \hat{p}_5 = \frac{8}{66} = 0.121, \qquad \hat{p}_6 = \frac{12}{66} = 0.182. \qquad (5)
\end{aligned}$$

6. (a) We have

$$p(\sigma) = \begin{cases} \frac{1}{T} & \text{for } 0\le\sigma\le T \\ 0 & \text{otherwise} \end{cases}.$$

Under a change of variables to $S = S(\sigma)$ we must have

$$p(S)\,\mathrm{d}S = p(\sigma)\,\mathrm{d}\sigma, \qquad \Rightarrow \qquad p(S) = p(\sigma(S))\frac{\mathrm{d}\sigma}{\mathrm{d}S}.$$

In this case $S = \sigma^2$ and we deduce

$$p(\sigma^2) = \left\{ \begin{array}{ll} \frac{1}{2T\sigma} & \text{for } 0 \leq \sigma^2 \leq T^2 \\ 0 & \text{otherwise} \end{array} \right. .$$

(b) If we assume σ^2 is fixed, then this is a standard Normal-Normal model and so using results from the lecture notes, we deduce

$$p(\mu|\mathbf{x}, \sigma^2) \sim N\left(\frac{s^2\sum_{i=1}^n x_i}{ns^2+\sigma^2}, \frac{\sigma^2 s^2}{ns^2+\sigma^2}\right).$$

If μ is fixed, the posterior on σ^2 is

$$p(\sigma^2|\mathbf{x},\mu) \propto \sigma^{-(n+1)} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\right] \mathbb{I}[0 \leq \sigma^2 \leq T^2]$$

$$\Rightarrow \quad p(\sigma^2|\mathbf{x},\mu) = \frac{A^{\frac{n-1}{2}}}{\Gamma\left(\frac{n-1}{2}\right) - \Gamma\left(\frac{1}{T^2}; \frac{n-1}{2}\right)} \sigma^{-(n+1)} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\right] \times \mathbb{I}[0 \leq \sigma^2 \leq T^2] \qquad (6)$$

where

$$A = \frac{1}{2}\sum_{i=1}^n (x_i-\mu)^2, \qquad \Gamma(x;n) = \int_0^x x^{n-1}\mathrm{e}^{-x}\,\mathrm{d}x$$

is the incomplete Gamma function, and $\Gamma(n) = \Gamma(\infty; n)$ is the complete Gamma function. This is a truncated inverse-Gamma distribution (i.e., $\tau = 1/\sigma^2$ follows a Gamma distribution). In the limit $T \to \infty$ the distribution is no longer truncated and we deduce

$$p(\tau|\mathbf{x},\mu) \sim \text{Gamma}\left(\frac{n}{2} - \frac{1}{2}, \frac{1}{2}\sum_{i=1}^n (x_i-\mu)^2\right)$$

as in lecture notes.

7. The cumulative density function of the Pareto distribution can be found to be

$$P(\theta \leq \Theta) = \int_{x_0}^{\Theta} \frac{a x_0^a}{\theta^{a+1}}\,\mathrm{d}\theta = \left\{ \begin{array}{ll} 1 - \left(\frac{x_0}{\Theta}\right)^a & \text{for } \Theta \geq x_0 \\ 0 & \text{otherwise} \end{array} \right. .$$

The CDF follows a $U[0,1]$ distribution and the inverse CDF is

$$F^{-1}(u) = \frac{x_0}{(1-u)^{\frac{1}{a}}}.$$

Hence we can draw samples from the Pareto distribution by simulating u_i $U[0,1]$ and then computing $\theta_i = F^{-1}(u_i)$.

8. (a) The posterior is

$$p(\phi|\mathbf{x}) \propto p(\mathbf{x}|\phi)p(\phi) \propto \phi^{\alpha-1}(1-\phi)^{\beta-1}\prod_{t=1}^{T+1} p_t^{x_t}$$
$$= \phi^{\alpha-1}(1-\phi)^{\beta-1}(1-\phi)^{\sum_{t=1}^{T} x_t}\phi^{\sum_{t=2}^{T+1}(t-1)x_t}$$
$$= \phi^{\alpha-1+\sum_{t=1}^{T+1}(t-1)x_t}(1-\phi)^{\beta-1+\sum_{t=1}^{T} x_t} \qquad (7)$$

which is the kernel of a Beta$(\alpha+\sum_{t=1}^{T+1}(t-1)x_t, \beta+\sum_{t=1}^{T} x_t)$ distribution.

(b) The mode of a Beta(a,b) distribution is at $x=(a-1)/(a+b-2)$ (provided $a>1$ and $b>1$, but if either of these conditions is violated it is not possible to use rejection sampling from a uniform distribution to obtain samples from Beta(a,b)). Hence, if we define

$$A = \frac{(a-1)^{a-1}(b-1)^{b-1}}{(a+b-2)^{a+b-2}}$$

we can generate samples from the Beta(a,b) distribution using the following simple rejection sampling algorithm

i. Draw $u_1 \sim U[0,1]$ and $u_2 \sim U[0,A]$.

ii. If

$$u_2 \le u_1^{a-1}(1-u_1)^{b-1}$$

then set $x_i = u_1$ and increment $i \to i+1$. Otherwise return to step i.

(c) The 95% HPD interval has width of 0.117, while the 95% symmetric credible interval has width of 0.111. Since the Beta distribution is unimodal the HPD interval must be the shortest 95% interval and therefore something is wrong in these results. Checking the quoted values using the properties of the Beta(91,9) distribution we find that everything is correct except the HPD interval. The pdf at the two ends of this interval is not equal, so it can't be HPD, and the probability contained is 92.4% so it is not even a 95% interval. The true HPD interval is $(0.853, 0.962)$.

9. The posterior on (ϕ_1, ϕ_a) is

$$p(\phi_1,\phi_a|\mathbf{x}) \propto \prod_{i=1}^{T} p_i^{x_i} = (1-\phi_1)^{x_1}\phi_1^{\sum_{t=2}^{T+1} x_t}(1-\phi_a)^{\sum_{t=2}^{T} x_t}\phi_a^{\sum_{j=3}^{T+1}(t-2)x_t}.$$

The conditional distributions can thus be seen to be

$$\phi_1|\phi_a,\mathbf{x} \sim \text{Beta}\left(1+\sum_{t=2}^{T+1} x_t, 1+x_1\right)$$
$$\phi_a|\phi_1,\mathbf{x} \sim \text{Beta}\left(1+\sum_{j=3}^{T+1}(t-2)x_t, 1+\sum_{t=2}^{T} x_t\right) \qquad (8)$$

A Gibbs sampling algorithm would work as follows

(a) Draw initial parameter values, (ϕ_1^0, ϕ_a^0), e.g., from the prior $U[0,1]$.

(b) At step $i = 1, \ldots, N$:

- Draw

$$\phi_1^i \sim \text{Beta}\left(1 + \sum_{t=2}^{T+1} x_t, 1 + x_1\right)$$

- Draw

$$\phi_a^i \sim \text{Beta}\left(1 + \sum_{j=3}^{T+1}(t-2)x_t, 1 + \sum_{t=2}^{T} x_t\right)$$

- Increment $i \to i + 1$.

(c) Discard the first M samples as burn-in. The remaining $N - M$ samples are a sample from the posterior.

The algorithm we have described is a standard Gibbs sampling algorithm. However, in this case the conditional distribution of ϕ_1 does not depend on ϕ_a and vice-versa. Thus we can draw samples directly from the posterior and there is no need to do MCMC. The Gibbs sampling algorithm above is providing direct samples from the posterior for all iterations $i \geq 1$.

10. (a) The acceptance probability for a move from x to y is

$$\alpha(x, y) = \min\left(1, \frac{q(y, x)\pi(y)}{q(x, y)\pi(x)}\right)$$

where $q(x, y)$ is the probability that a move from x to y would be proposed by the chosen proposal distribution. In this case we have

$$\pi(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{x^2}{2\sigma^2}\right]$$

and

$$q(x, y) = \frac{1}{\sqrt{2\pi}\,\tau} \exp\left[-\frac{(y - ax)^2}{2\tau^2}\right].$$

Therefore we have

$$\begin{aligned}
\alpha(x, y) &= \min\left(1, \frac{q(y, x)\pi(y)}{q(x, y)\pi(x)}\right) \\
&= \min\left(1, \exp\left[\frac{(x^2 - y^2)}{2\sigma^2} + \frac{[(y - ax)^2 - (x - ay)^2]}{2\tau^2}\right]\right) \\
&= \min\left(1, \exp\left[(x^2 - y^2)\left(\frac{1}{2\sigma^2} + \frac{(a^2 - 1)}{2\tau^2}\right)\right]\right). \qquad (9)
\end{aligned}$$

(b) The condition that the acceptance probability $\alpha(x, y) = 1$ for all x, y is that the argument of the exponential is 0, i.e.,

$$\frac{1}{\sigma^2} + \frac{(a^2 - 1)}{\tau^2} = 0, \qquad \Rightarrow \qquad \tau^2 = \sigma^2(1 - a^2).$$

(c) If $a = 0$ then the proposal distribution, $q(x, y)$, is independent of the current point, x, so this describes an independence sampler. Additionally setting $\tau = \sigma$ the acceptance probability is again always 1, but in this case we are proposing samples directly from the posterior and so we don't need to use MCMC.

11. The Markov chain is reversible if there exists a distribution $\pi(x)$ such that

$$\pi(x)\mathcal{K}(x,y) = \pi(y)\mathcal{K}(y,x)$$

where $\mathcal{K}(x,y)$ is the probability of moving from point x to point y. For a Markov Chain constructed by the Metropolis-Hastings algorithm we have $\mathcal{K}(x,y) = q(x,y)\alpha(x,y)$ using the notation of the previous question. Therefore

$$\begin{aligned}\pi(x)\mathcal{K}(x,y) &= \pi(x)q(x,y)\min\left(\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, 1\right)\\ &= \min\left(\pi(y)q(y,x), \pi(x)q(x,y)\right)\\ &= \min\left(1, \frac{\pi(x)q(x,y)}{\pi(y)q(y,x)}\right)\pi(y)q(y,x) = \mathcal{K}(y,x)\pi(y). \qquad (10)\end{aligned}$$

As required. Integrating this equation we find

$$\int \pi(x)\mathcal{K}(x,y)\mathrm{d}x = \int \pi(y)\mathcal{K}(y,x)\mathrm{d}x = \pi(y)\int \mathcal{K}(y,x)\mathrm{d}x = \pi(y) \qquad (11)$$

as required. The last equality follows from the fact that $\mathcal{K}(y,x)$ is a probability distribution over x and therefore must integrate to 1.

12. (a) The posterior distribution of the success rate is

$$\begin{aligned}p(\theta \mid y) &\propto f(y \mid \theta)p(\theta)\\ &= \binom{n}{y}\theta^{y}(1-\theta)^{n-y}\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}\\ &\propto \theta^{a+y-1}(1-\theta)^{b+n-y-1},\end{aligned}$$

which we recognise as the kernel of a beta distribution with parameters $a+y$ and $b+n-y$. Therefore,

$$\theta \mid y \sim \text{Beta}(a+y, b+n-y).$$

Taking $a = 9.2$, $b = 13.8$, $n = 20$, and $y = 15$, results in a Beta$(24.2, 18.8)$ distribution.

(b) The posterior mean is $24.2/(24.2 + 18.8) = 0.563$. The HPD interval is $(0.416, 0.708)$.

(c) By computing the 2.5% and 97.5% percentiles of the posterior distribution, we obtain the symmetric credible interval $(0.414, 0.706)$. The two intervals (HPD and credible) are basically the same because in this case the posterior distribution is unimodal (and also practically symmetric around the mean).

(d) The probability that the true success rate is greater than 0.6 is 0.316.

(e) Under a uniform prior, i.e., with a Beta$(1,1)$ prior distribution, the above probability changes to 0.904. With a Jeffreys' prior, it is 0.918.

(f) Let z denotes the number of positive responses in further $m = 40$ patients. We must first calculate the posterior predictive distribution

$$
\begin{aligned}
f(z \mid y) &= \int_{\Theta} f(z \mid \theta) p(\theta \mid y) \mathrm{d}\theta \\
&= \int_0^1 \binom{m}{z} \theta^z (1-\theta)^{m-z} \frac{1}{B(a+y, b+n-y)} \theta^{a+y-1} (1-\theta)^{b+n-y-1} \mathrm{d}\theta \\
&= \binom{m}{z} \frac{1}{B(a+y, b+n-y)} \int_0^1 \theta^{a+y+z-1} (1-\theta)^{b+n-y+m-z-1} \mathrm{d}\theta \\
&= \binom{m}{z} \frac{B(a+y+z, b+n-y+m-z)}{B(a+y, b+n-y)} \\
&\qquad \times \int_0^1 \frac{1}{B(a+y+z, b+n-y+m-z)} \theta^{a+y+z-1} (1-\theta)^{b+n-y+m-z-1} \mathrm{d}\theta \\
&= \binom{m}{z} \frac{B(a+y+z, b+n-y+m-z)}{B(a+y, b+n-y)}
\end{aligned}
$$

It is now straightforward to find that $\Pr(z \geq 25) = 0.329$.

(g) We start by calculating the prior predictive distribution

$$
\begin{aligned}
f(y) &= \int_{\Theta} f(y \mid \theta) p(\theta) \mathrm{d}\theta \\
&= \int_0^1 \binom{n}{y} \theta^y (1-\theta)^{n-y} \frac{1}{B(a, b)} \theta^{a-1} (1-\theta)^{b-1} \mathrm{d}\theta \\
&= \binom{n}{y} \frac{1}{B(a, b)} \int_0^1 \theta^{a+y-1} (1-\theta)^{b+n-y-1} \mathrm{d}\theta \\
&= \binom{n}{y} \frac{B(a+y, b+n-y)}{B(a, b)}
\end{aligned}
$$

The prior predictive probability of observing at least 15 positive responses can then be computed from the last expression and it is 0.01526. This suggests some evidence that the data and the prior are incompatible.

(h) i. Solving for a and b gives a $\mathrm{Beta}(12, 3)$ prior.

ii. The mixture prior $\theta \sim \pi \mathrm{Beta}(a_1, b_1) + (1-\pi) \mathrm{Beta}(a_2, b_2)$ is plotted in Figure 1.

iii. We will start by finding the posterior distribution of θ.

$$
\begin{aligned}
p(\theta \mid y) &\propto \binom{n}{y} \theta^y (1-\theta)^{n-y} \left\{ \pi \frac{1}{B(a_1, b_1)} \theta^{a_1-1} (1-\theta)^{b_1-1} + (1-\pi) \frac{1}{B(a_2, b_2)} \theta^{a_2-1} (1-\theta)^{b_2-1} \right\} \\
&\propto \pi \frac{1}{B(a_1, b_1)} \theta^{a_1+y-1} (1-\theta)^{b_1+n-y-1} + (1-\pi) \frac{1}{B(a_2, b_2)} \theta^{a_2+y-1} (1-\theta)^{b_2+n-y-1} \\
&= \pi \frac{B(a_1+y, b_1+n-y)}{B(a_1, b_1)} \frac{1}{B(a_1+y, b_1+n-y)} \theta^{a_1+y-1} (1-\theta)^{b_1+n-y-1} \\
&\quad + (1-\pi) \frac{B(a_2+y, b_2+n-y)}{B(a_2, b_2)} \frac{1}{B(a_2+y, b_2+n-y)} \theta^{a_2+y-1} (1-\theta)^{b_2+n-y-1} \\
&= \pi \frac{B(a_1+y, b_1+n-y)}{B(a_1, b_1)} \mathrm{Beta}(\theta \mid a_1+y, b_1+n-y) \\
&\quad + (1-\pi) \frac{B(a_2+y, b_2+n-y)}{B(a_2, b_2)} \mathrm{Beta}(\theta \mid a_2+y, b_2+n-y).
\end{aligned}
$$

We are almost there, but note that the 'weights' $\pi \frac{B(a_1+y, b_1+n-y)}{B(a_1, b_1)}$ and $(1-\pi) \frac{B(a_2+y, b_2+n-y)}{B(a_2, b_2)}$ do not sum up to one. Renormalising, we finally obtain that

$$
\theta \mid y \sim \omega_1 \mathrm{Beta}(\theta \mid a_1+y, b_1+n-y) + (1-\omega_1) \mathrm{Beta}(\theta \mid a_2+y, b_2+n-y)
$$

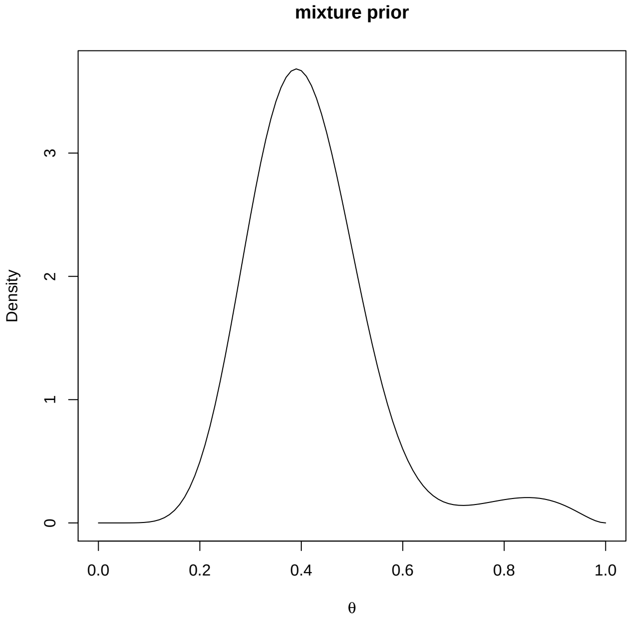

Figure 1: The mixture prior for question 12(h)(ii).

with

$$\omega_1 = \pi \frac{B(a_1+y, b_1+n-y)}{B(a_1,b_1)} \left(\pi \frac{B(a_1+y, b_1+n-y)}{B(a_1,b_1)} + \right. \\ \left. +(1-\pi)\frac{B(a_2+y, b_2+n-y)}{B(a_2,b_2)} \right)^{-1}$$

We are now ready to compute the required probability, which turns out to be 0.58062.

iv. The procedure is similar to the one in part (g), the only difference is the computation of the prior predictive distribution. In this case,

$$\begin{aligned} f(y) &= \int_\Theta f(y \mid \theta) p(\theta) \mathrm{d}\theta \\ &= \int_0^1 \binom{n}{y} \theta^y (1-\theta)^{n-y} \left\{ \pi \frac{1}{B(a_1,b_1)} \theta^{a_1-1}(1-\theta)^{b_1-1} \right. \\ &\qquad \left. +(1-\pi)\frac{1}{B(a_2,b_2)}\theta^{a_2-1}(1-\theta)^{b_2-1} \right\} \mathrm{d}\theta \\ &= \pi \binom{n}{y} \frac{1}{B(a_1,b_1)} \int_0^1 \theta^{a_1+y-1}(1-\theta)^{b_1+n-y-1}\mathrm{d}\theta \\ &\qquad + (1-\pi)\binom{n}{y}\frac{1}{B(a_2,b_2)}\int_0^1 \theta^{a_2+y-1}(1-\theta)^{b_2+n-y-1}\mathrm{d}\theta \\ &= \pi \binom{n}{y} \frac{B(a_1+y, b_1+n-y)}{B(a_1,b_1)} + (1-\pi)\binom{n}{y}\frac{B(a_2+y,b_2+n-y)}{B(a_2,b_2)} \end{aligned}$$

The prior predictive probability of observing at least 15 positive responses is now 0.0514, which does not provide strong evidence of incompatibility.

v. The prior/likelihood/posterior plot is shown in Figure 2.

13. (a) The `pystan` model definition for this problem is

```
lin_model_def = """
```

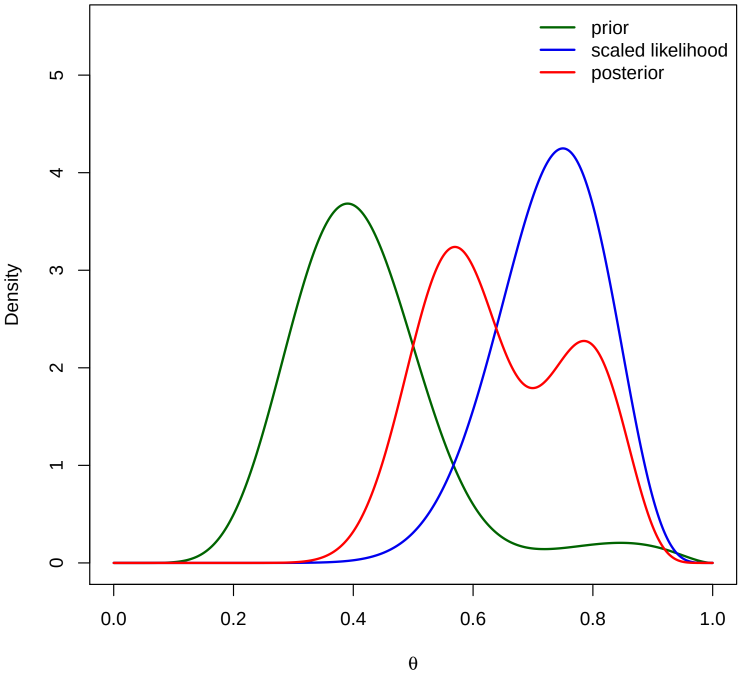

Figure 2: Comparison of prior, likelihood and posterior for question 12(h)(v).

```
data {
  int npts;
  vector[npts] year;
  vector[npts] jump;

  real mu0;
  real var0;
  real a;
  real b;
}

parameters {
  real beta0;
  real beta1;
  real<lower=0> v;
}

model {
  for (i in 1:npts) {
        target+=normal_lpdf(jump[i] | beta0+beta1*year[i],sqrt(v));
  }
  target += normal_lpdf(beta0 | mu0, sqrt(var0));
  target += normal_lpdf(beta1 | mu0, sqrt(var0));
  target += inv_gamma_lpdf(v | a,b);
}
"""
```

Fitting this model gives the traceplots and posterior distributions shown in Figure 3. Autocorrelation plots show no evidence of autocorrelation, with coefficients close to 0 for all lags greater than 0. Summary statistics, effective number of samples and Gelman-Rubin statistics can be read off from the table below.

```
        mean se_mean      sd    2.5%     25%     50%     75%  97.5%  n_eff    Rhat
beta0 -19.32    0.03    2.92  -25.13  -21.23  -19.32  -17.38  -13.6  11107     1.0
beta1   0.01  1.4e-5  1.5e-3    0.01    0.01    0.01    0.01   0.02  11114     1.0
v       0.07  2.2e-4    0.02    0.04    0.05    0.07    0.08   0.12  10274     1.0
```

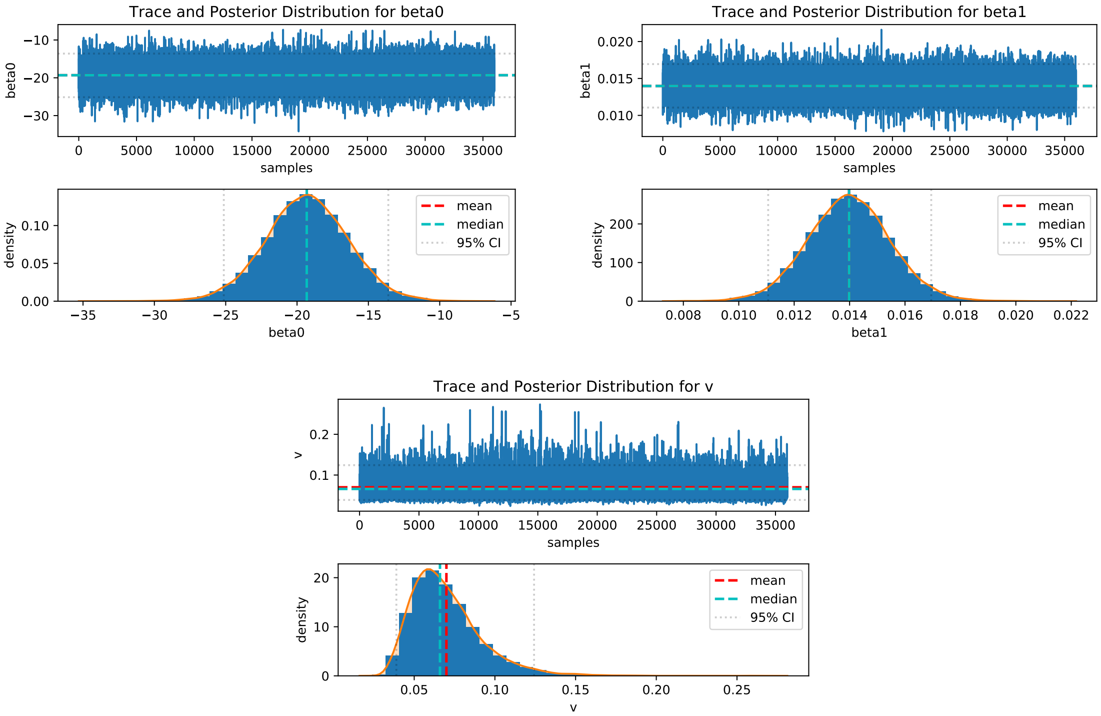

Figure 3: Trace plots and posterior distributions for the linear model fit to the long jump data, for parameters β_0 (top left), β_1 (top right) and σ^2 (bottom).

(b) `pystan` is sampling well for this model, although trying the same fit using `rjags` gives quite poor sampling. Centring of covariates often helps improve sampling, while leaving the posterior on the slope of the regression line, which is the key parameter, unchanged. In this case we do not need to change the `pystan` model, but just need to change the `year` data array as follows

```
year_cent=year-np.mean(year)
```

Sampling from this model we obtain the summary table

```
        mean se_mean      sd    2.5%     25%     50%     75%  97.5%  n_eff    Rhat
beta0   8.01  3.7e-4    0.05    7.91    7.97    8.01    8.04   8.11  19145     1.0
beta1   0.01  7.1e-6  1.5e-3    0.01    0.01    0.01    0.02   0.02  45858     1.0
v       0.07  1.7e-4    0.02    0.04    0.05    0.07    0.08   0.12  17484     1.0
```

There are a larger number of effective samples in this run, indicating that it is easier to sample from. The result is consistent, with $\hat{\beta}_1 = 0.0141$ compared to $\hat{\beta}_1 = 0.0140$ in the non-centred case.

(c) The `pystan` model for robust regression with fixed student-t degrees of freedom is

```
lin_model_robust_def = """
data {
  int npts;
  vector[npts] year;
  vector[npts] jump;
  real mu0;
  real var0;
  real a;
  real b;
  real nu;
}

parameters {
  real beta0;
  real beta1;
  real<lower=0> v;
}

model {
  for (i in 1:npts) {
        target+=student_t_lpdf(jump[i] | nu, beta0+beta1*year[i],sqrt(v));
  }
  target += normal_lpdf(beta0 | mu0, sqrt(var0));
  target += normal_lpdf(beta1 | mu0, sqrt(var0));
  target += inv_gamma_lpdf(v | a,b);
}
"""
```

and the summary table from fitting this model with $\nu = 3$ is

```
        mean se_mean     sd   2.5%    25%    50%    75%  97.5%  n_eff   Rhat
beta0    8.0  3.5e-4   0.05   7.91   7.97    8.0   8.03   8.09  18204    1.0
beta1   0.01  6.4e-6 1.3e-3   0.01   0.01   0.01   0.01   0.02  44428    1.0
v       0.04  1.3e-4   0.02   0.02   0.03   0.04   0.05   0.09  17361    1.0
```

The new estimate of the slope coefficient is $\hat{\beta}_1 = 0.01393$. To allow the degrees of freedom to vary we use the `pystan` model

```
lin_model_robustB_def = """
data {
  int npts;
  vector[npts] year;
  vector[npts] jump;

  real mu0;
  real var0;
  real a;
  real b;
  real c;
  real d;
}

parameters {
  real beta0;
  real beta1;
  real<lower=0> v;
  real nu;
}

model {
  for (i in 1:npts) {
        target+=student_t_lpdf(jump[i] | nu, beta0+beta1*year[i],sqrt(v));
  }
  target += normal_lpdf(beta0 | mu0, sqrt(var0));
  target += normal_lpdf(beta1 | mu0, sqrt(var0));
  target += inv_gamma_lpdf(v | a,b);
  target += gamma_lpdf(nu |c,d);
}
"""
```

and the result table from fitting this model with $c = d = 0.1$ is

```
        mean se_mean      sd    2.5%     25%     50%     75%  97.5%  n_eff    Rhat
beta0    8.0  4.5e-4    0.05    7.91    7.97     8.0    8.03    8.1  11482     1.0
beta1   0.01  7.5e-6  1.4e-3    0.01    0.01    0.01    0.01   0.02  35027     1.0
v       0.05  2.1e-4    0.02    0.02    0.04    0.05    0.06    0.1   9020     1.0
nu      7.92    0.11    6.62    1.67    3.66     5.9    9.84  26.08   3514     1.0
```

The new estimate of the slope coefficient is now $\hat{\beta}_1 = 0.01396$. To fit both of these latter two models, we used the centred "year" covariate. Inspection of the data shows that the year 1968 is an outlier. This data point could be removed from the data before analysing, which makes some difference to the results. Robust regression is more immune to the presence of the outlier and so favours somewhat shallower slopes than the first fits.

14. (a) The conjugate prior to a Normal distribution is a Normal distribution. The expert prior could be interpreted as a uniform distribution on $[0, 2]$, which has mean 1 and variance 1/3. The Normal distribution with this mean and variance is $\mathrm{N}(1, 1/3)$ and so that is a good choice of prior. It is not the only choice. Anything of the form $N(1, k)$ with $k \sim 1$, e.g., $k = 0.5, 1, 2$ is OK since the expert opinion is vague. However a prior with $k \ll 1$ or $k \gg 1$ would not respect the expert opinion and a truncated distribution would not be conjugate. The posterior for a Normal-Normal model with known measurement variance σ^2 and prior $N(\mu_0, \sigma_0^2)$ is

$$\mathrm{N}\left(\frac{n\bar{y}\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2}, \frac{\sigma^2\sigma_0^2}{n\sigma_0^2 + \sigma^2}\right).$$

This data has $n = 10$, $\sigma^2 = 30$ and $\bar{y} = 1.6116$ so for the $\mathrm{N}(1, 1/3)$ prior the posterior is $\mathrm{N}(1.06, 0.3)$.

(b) As in part (a) there are several ways to interpret the US expert's information. Following the procedure above the US expert prior can be interpreted as $\mathrm{N}(5, 4/3)$. A suitable mixture prior is of the form $p(\mu) = wp_1(\mu) + (1 - w)p_2(\mu)$ where $p_1(\mu)$ and $p_2(\mu)$ are the prior from the UK and US experts respectively and w is the weight for prior $p_1(\mu)$. A suitable choice is $w = 2/3$ since there are twice as many US experts. In this case we have $p_1(\mu) = N(\mu_1, \sigma_1^2)$ and $p_2(\mu) = N(\mu_2, \sigma_2^2)$. The posterior can be found to be

$$w'\mathrm{N}\left(\frac{n\bar{y}\sigma_1^2 + \mu_1\sigma^2}{n\sigma_1^2 + \sigma^2}, \frac{\sigma^2\sigma_1^2}{n\sigma_1^2 + \sigma^2}\right) + (1 - w')\mathrm{N}\left(\frac{n\bar{y}\sigma_2^2 + \mu_2\sigma^2}{n\sigma_2^2 + \sigma^2}, \frac{\sigma^2\sigma_2^2}{n\sigma_2^2 + \sigma^2}\right), \tag{12}$$

where

$$w' = \frac{k_1 w}{k_1 w + k_2(1 - w)}, \qquad k_i = \frac{1}{\sqrt{\sigma^2 + n\sigma_i^2}} \exp\left[-\frac{1}{2}\left(\frac{n(\bar{y} - \mu_i)^2}{\sigma^2 + n\sigma_i^2}\right)\right]. \tag{13}$$

In this case we find $w' = 0.890$ and the posterior is $0.890N(1.06, 0.3)+0.110N(3.95, 0.923)$.

(c) We need to choose a suitable prior on the precision $\tau = 1/\sigma^2$ and we use $\Gamma(0.01, 0.01)$. The `pystan` model definition is as follows

```
normal_model_def = """
data {
  int npts;
  vector[npts] y;
  vector[2] wt;

  real mu1;
  real var1;
  real mu2;
  real var2;
  real a;
  real b;
}

parameters {
  real mu;
  real<lower=0> v;
}

model {
  for (i in 1:npts) {
        target+=normal_lpdf(y[i] | mu,sqrt(v));
  }
  target += log(exp(log(wt[1])+normal_lpdf(mu | mu1,sqrt(var1)))
                         +exp(log(wt[2])+normal_lpdf(mu | mu2,sqrt(var2))));
  target += inv_gamma_lpdf(v | a,b);
}

generated quantities {
   real sigma;
   sigma = sqrt(v);
}

"""
```

The output table after fitting the model is

```
        mean se_mean     sd   2.5%    25%    50%    75%  97.5%  n_eff   Rhat
mu      1.24  7.4e-3   0.71   0.16   0.82   1.17   1.54   3.21   9114    1.0
v       13.7    0.07   8.29   5.28   8.69  11.66  16.18  34.44  13376    1.0
sigma   3.58  7.7e-3   0.93    2.3   2.95   3.41   4.02   5.87  14605    1.0
```

and the resulting posteriors and trace plots are shown in Figure 4. Note that you will not get exactly these values due to sampling error, but your values should be close to these.

(d) The probability that $\mu < 1$ can be found by integrating the posterior for μ from $-\infty$ to 1. This can be done by including a line like

```
real frac;
if (mu < 1)
    frac=1;
else
    frac=0;
```

in the "generated quantities" section of the pystan model definition and looking at the posterior mean of the new variable "frac". We obtain an estimate $p = 0.375$.

To compute the probability that a single future measurement will yield a negative log-concentration, we first need to compute $p_{-,1}$, the posterior predictive

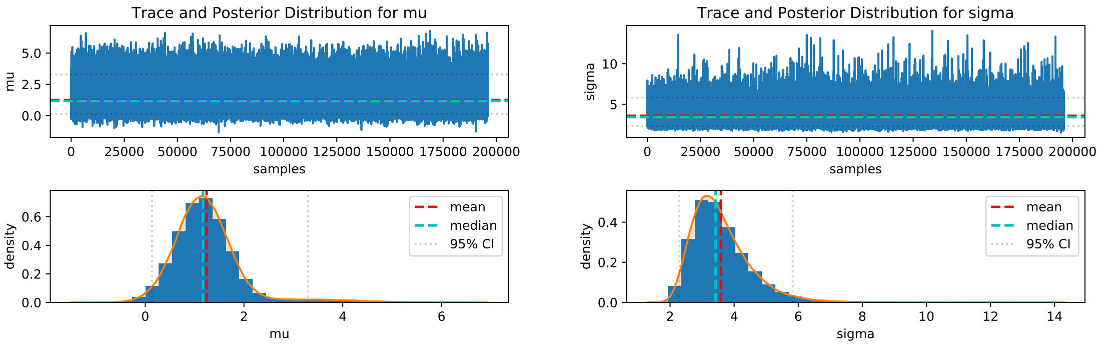

Figure 4: Posterior distributions and trace plots for the mean μ (left) and standard deviation σ (right) of the log concentration of the chemical.

probability of obtaining a negative measurement in a single future observation. This is accomplished by adding these lines to the "generated quantities" part of the model definition

```
real ynew;
real prob;
ynew=normal_rng(mu,sigma);
if (ynew < 0)
    prob=1;
else
    prob=0;
```

and looking at the posterior mean for "prob". This gives $p_{-,1} \approx 0.360$.

The probability that at least one of N future measurements yields a value less than 0 is one minus the probability that none of them yield a value less than 0 which can be calculated as $p_{-,5} = 1 - (1 - p_{-,1})^N$. For $N = 5$ and $p_{-,1} = 0.360$ we find $p_{-,5} \approx 0.893$.

(e) If we include w as a parameter with a flat prior in the range $[0, 1]$ the posterior on (μ, w) is given by Eq. (12) above, but with w' and $(1 - w')$ replaced by

$$w' \to \frac{2k_1 w}{k_1 + k_2}, \qquad 1 - w' \to \frac{2k_2(1 - w)}{k_1 + k_2},$$

with k_i as defined in Eq. (13). In this case we find the joint posterior is

$$1.561 w p_G(\mu; 1.06, 0.3) + 0.439(1 - w) p_G(\mu; 3.95, 0.923),$$

where $p_G(x; \mu, \sigma^2)$ denotes the pdf of an $N(\mu, \sigma^2)$ distribution.

The marginal distribution on μ is found by integrating over w

$$p(\mu|\mathbf{d}) = 0.781 p_G(\mu; 1.06, 0.3) + 0.219 p_G(\mu; 3.95, 0.923).$$

The marginal distribution on w is found by integrating over μ

$$p(w|\mathbf{d}) = 0.439 + 1.122w.$$

The marginalisation distribution on μ is the same distribution that would be obtained using equal weights on the two priors in the mixture, i.e., $w = 1/2$.

This is because $w = 1/2$ is the prior expectation value for a $U[0,1]$ and the w prior is a hyperprior, i.e., the prior on a parameter that describes a prior on other parameters. The marginal on w is a straight line. It is rising, meaning that the mode of the posterior is $w = 1$, i.e., we favour the prior from the UK experts. We have weak evidence to suggest the UK experts are better at predicting than the US experts, but this is perhaps unsurprising given that the data is being collected in the UK. A straight line posterior does not indicate a strong constraint on the parameter. This is because the w parameter only enters once, as a prior on the mean that is common to all the subsequent observations. As we make more observations we expect to measure μ better and better, but there will be no strong change in our ability to measure w, since it only enters once. If we imagine a scenario in which we collect sets of data in multiple different sites, and we suppose the mean at each site is different, drawn from the prior described by w, then as we add more and more sites we would start to see a concentration in the w prior and stronger evidence that one set of experts is correct.