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) 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

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

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

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

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

$$\begin{aligned}\mathbb{P}(\mathbf{x}|H_1) &= (2\pi\sigma^2)^{-n/2}(2\pi\tau^2)^{-1/2}\int_{-\infty}^{\infty}\exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j-\mu)^2-\frac{1}{2\tau^2}\mu^2\right]\mathrm{d}\mu \\ &= (2\pi\sigma^2)^{-n/2}(2\pi\tau^2)^{-1/2}\exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{n}x_j^2+\frac{n^2\Sigma^2\bar{x}^2}{2\sigma^4}\right]\times \\ &\qquad\times\int_{-\infty}^{\infty}\exp\left[-\frac{1}{2\Sigma^2}\left(\mu-\frac{n\Sigma^2}{\sigma^2}\bar{x}\right)^2\right]\mathrm{d}\mu \\ &= (2\pi\sigma^2)^{-n/2}\frac{\Sigma}{\tau}\exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{n}x_j^2+\frac{n^2\Sigma^2\bar{x}^2}{2\sigma^4}\right] \qquad (3)\end{aligned}$$

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!

3. (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 $p(\mathbf{p}|\mathbf{x})\sim\mathrm{Dir}(\alpha_1+x_1,\alpha_2+x_2,\ldots,\alpha_m+x_m)$.

(b) We showed in the lecture notes 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

$$\begin{aligned}&\hat{p}_1=\frac{11}{66}=\frac{1}{6}=0.167, \qquad \hat{p}_2=\frac{13}{66}=0.197, \qquad \hat{p}_3=\frac{13}{66}=0.197, \\ &\hat{p}_4=\frac{9}{66}=0.136, \qquad \hat{p}_5=\frac{8}{66}=0.121, \qquad \hat{p}_6=\frac{12}{66}=0.182. \qquad (4)\end{aligned}$$

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

$$P(\theta\le\Theta)=\int_{x_0}^{\Theta}\frac{ax_0^a}{\theta^{a+1}}\,\mathrm{d}\theta=\begin{cases}1-\left(\frac{x_0}{\Theta}\right)^a & \text{for } \Theta\ge x_0\\ 0 & \text{otherwise}\end{cases}.$$

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 $\theta_i = F^{-1}(u_i)$.

5. (a) The posterior is

$$\begin{aligned} p(\phi|\mathbf{x}) \propto p(\mathbf{x}|\phi)p(\phi) &\propto \phi^{\alpha-1}(1-\phi)^{\beta-1}\prod_{t=1}^{T+1} p_t^{x_t} \\ &= \phi^{\alpha-1}(1-\phi)^{\beta-1}(1-\phi)^{\sum_{t=1}^{T} x_t}\phi^{\sum_{t=2}^{T+1}(t-1)x_t} \\ &= \phi^{\alpha-1+\sum_{t=1}^{T+1}(t-1)x_t}(1-\phi)^{\beta-1+\sum_{t=1}^{T} x_t} \end{aligned} \tag{5}$$

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

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

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 $(0.853, 0.962)$.

6. (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

$$\begin{aligned}\alpha(x,y) &= \min\left(1, \frac{q(y,x)\pi(y)}{q(x,y)\pi(x)}\right)\\ &= \min\left(1, \exp\left[\frac{(x^2-y^2)}{2\sigma^2} + \frac{[(y-ax)^2-(x-ay)^2}{2\tau^2}\right]\right)\\ &= \min\left(1, \exp\left[(x^2-y^2)\left(\frac{1}{2\sigma^2}+\frac{(a^2-1)}{2\tau^2}\right)\right]\right). \qquad (6)\end{aligned}$$

(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.

7. We denote the observed data in frequency bin i by s_i. The full data set, denoted D, takes the form

$$s_i = n_i \quad \forall \quad i \in [1,N], \qquad s_{N+1} = n_{N+1} + A, \qquad n_i \sim N(0,\sigma) \quad \forall \quad i \quad (7)$$

with corresponding likelihood

$$p(D|\sigma,A) \propto \frac{1}{\sigma^{N+1}}\left(\prod_{i=1}^{N}\exp\left[-\frac{s_i^2}{2\sigma^2}\right]\right)\exp\left[-\frac{(s_{N+1}-A)^2}{2\sigma^2}\right]. \qquad (8)$$

For flat priors, this is also the posterior, $p(\sigma, A|D)$. We are not so interested in the value of σ, but the value of A, so we can marginalise the posterior over σ. If we write

$$X(A) = (s_{N+1}-A)^2 + \sum_{i=1}^{N} s_i^2 \qquad (9)$$

then we find

$$p(A|D) \quad \propto \quad \frac{1}{\sigma^{N+1}}\exp\left(-X(A)/2\sigma^2\right) \Rightarrow \int_0^\infty p(\sigma,A|D)\mathrm{d}\sigma \propto \left(\frac{2}{X(A)}\right)^{\frac{N}{2}}\Gamma\left(\frac{N}{2}\right)$$

where $\Gamma(x)$ is the gamma function, with $\Gamma(n+1) = n!$. Note that we are assuming a flat prior in σ here, but other priors could be included straightforwardly.

We now recall that the posterior density for the student-t distribution with n degrees of freedom is

$$p_{t,n}(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}.$$

Hence we see that

$$p(A|D) \propto p_{t,N-1}\left(\frac{A - s_{N+1}}{\sqrt{\frac{1}{N-1}\sum_{i=1}^{N} s_i^2}}\right) = p_{t,N-1}\left(\frac{A - s_{N+1}}{\hat{\sigma}}\right)$$

where $\hat{\sigma} = \sqrt{\sum s_i^2/(N-1)}$ is the usual unbiased estimate of the variance. We can compare this to the standard likelihood used in parameter estimation for gravitational wave detectors, which is

$$p(A|D) \propto p_{N(0,1)}\left(\frac{A - s_{N+1}}{\hat{\sigma}}\right)$$

where

$$p_{N(0,1)}(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{x^2}{2}\right]$$

is the pdf of a standard Normal distribution. We see that this marginalisation over the PSD uncertainty is equivalent to replacing the Normal distribution by a student-t distribution. This is the same procedure that we argued could be used for robust regression.

If we take the limit that $N \to \infty$ and write $\sum_{i=1}^{N} s_i^2 = N\sigma_{\rm est}^2$ we find

$$X(A)^{-\frac{N}{2}} = \left(\frac{1}{N\sigma_{\rm est}^2}\right)^{\frac{N}{2}} \left(1 + \frac{(s_{N+1} - A)^2}{2(N/2)\sigma_{\rm est}^2}\right)^{-\frac{N}{2}} \approx B \exp\left(-\frac{(s_{N+1} - A)^2}{2\sigma_{\rm est}^2}\right) \quad (10)$$

where the last part follows from the standard result

$$\left(1 + \frac{k}{n}\right)^n \sim \mathrm{e}^k \qquad \text{as } n \to \infty \quad (11)$$

For large N we also expect $\sigma_{\rm est}^2 = \sigma^2 + O(1/N)$ and so we recover the standard Normal likelihood.

8. (a) The information available before O1 indicates that the rate is uncertain over orders of magnitude. Under these circumstances it is reasonable to suppose that the the log of the rate is uniform in some range. So, we represent the prior as

$$\log_{10}(\lambda) \sim U[-2, 3].$$

This prior has an expectation value of $999.99/\ln(10^5) = 86.858$ and variance of $999999.9999/2\ln(10^5) - 86.86^2 = 35885.13$. The conjugate distribution to a Poisson model is a Gamma distribution, $\Gamma(a, b)$, for which the mean and variance are a/b and a/b^2 respectively. Matching the mean and variance we find $b = 1/413.15$ and $a = 0.210$. We use a conjugate distribution since we then know the posterior will also be in the conjugate family and so it is computationally convenient. [**Note:** any reasonable prior choice is fine, provided it is justified. It must be wide and flat over several decades and make some use of the prior information.]

(b) We note first that all of the observation runs are different lengths. The rate λ was quoted in units of yr^{-1}. Poisson processes are additive, i.e., if the rate in

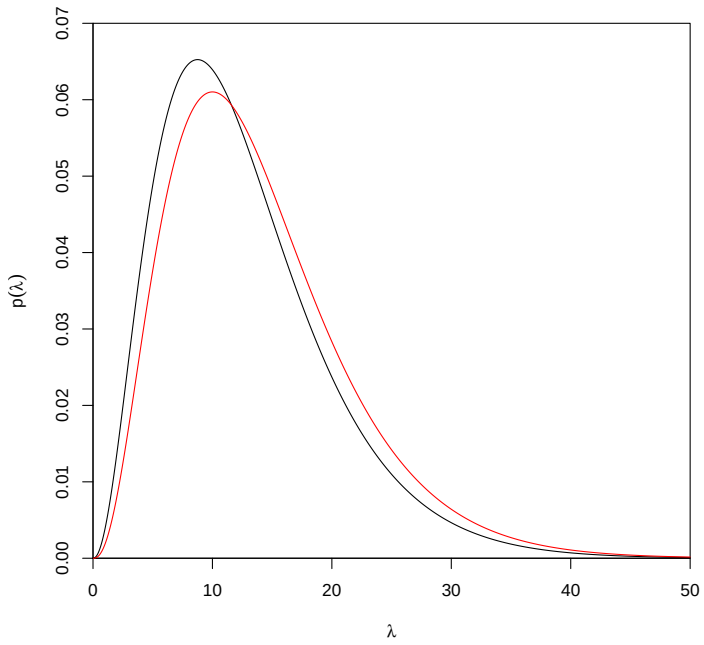

Figure 1: Posterior distribution for the rate per year, λ, after observing the O1 data for the conjugate prior (black line) and the Jeffrey's prior (red line).

time period T is λ, the rate in time period kT is $\tilde{\lambda} = k\lambda$. If the prior on λ is $\Gamma(a, b)$ then we have

$$p(\tilde{\lambda}) = k^{-1}\frac{b^a}{\Gamma(a)}\left(\frac{\tilde{\lambda}}{k}\right)^{a-1} \mathrm{e}^{-b\tilde{\lambda}/k} = \frac{(b/k)^a}{\Gamma(a)}\tilde{\lambda}^{a-1}\mathrm{e}^{-(b/k)\tilde{\lambda}},$$

i.e., the prior on $k\lambda$ is $\Gamma(a, b/k)$. As three events are observed in O1, we can write down the posterior distribution on $\tilde{\lambda}$ as $\Gamma(a+3, b/k+1)$ and the posterior distribution on λ is $\Gamma(a+3, b+k)$. In this case using the conjugate prior derived above we have $\Gamma(3.21, 0.252)$. The posterior mean and standard deviation are 12.717 and 7.098 respectively, a 95% symmetric confidence interval is $(2.817, 29.934)$ and the posterior distribution is shown in Figure 1.

(c) The probability that the rate exceeds 15 can be computed from the cumulative density function of the gamma distribution. This is $\gamma(\alpha, \beta x)/\Gamma(\alpha)$, where $\gamma(\alpha, x)$ is the incomplete gamma function. In this case we need $1 - \gamma(3.21, 3.78)/\Gamma(3.21) = 0.312$.

(d) The Jeffrey's prior for the Poisson distribution is the improper prior $p(\lambda) \propto \lambda^{-1/2}$. This can be approximated by a $\Gamma(1/2, \beta)$ distribution with $\beta \to 0$. The posterior on λ from the O1 data with the Jeffrey's prior is therefore $\Gamma(3.5, 0.25)$. The posterior is shown as a red line in Figure 1, the posterior mean and standard deviation are 14 and 7.48 and a 95% symmetric confidence interval is given by $(3.38, 32.0)$. The probability that the rate exceeds 15 is now 0.379. So, the results change a little bit and in particular the Jefffrey's prior favours slightly higher rates than the conjugate prior, but there is not a very big difference between the two.

(e) The second science run, O2, lasts 6 months, so the Poisson rate is 0.5λ. The posterior for O1 therefore gives a prior on the rate in O2 of $\Gamma(3.21, 0.504)$

using the conjugate prior. For a posterior of the form $\Gamma(\alpha, \beta)$, the posterior predictive probability of seeing n events in O2 is therefore

$$p(n|d_{O1}) = \int_0^\infty \frac{\lambda^n \mathrm{e}^{-\lambda}}{n!} \frac{\beta^\alpha \lambda^{\alpha-1} \mathrm{e}^{-\beta\lambda}}{\Gamma(\alpha)} \mathrm{d}\lambda = \frac{1}{n!} \frac{\beta^\alpha}{(\beta+1)^{\alpha+n}} \frac{\Gamma(\alpha+n)}{\Gamma(\alpha)}. \tag{12}$$

This can be recognised as a negative binomial distribution. We compute the probability of seeing 6 or more events in O2 as $1 - \sum_{n=0}^{5} p(n|d_{O1}) = 0.504$. The posterior for the rate in the first 5 months of O2 is $\Gamma(3.21, 0.608)$ and the probability of seeing 1 or fewer events in 5 months is given by computing $p(0|d_{O1}) + p(1|d_{O1}) = 0.131$ using this α and β in Eq. (12). The posterior probability for the rate in 1 month is $\Gamma(3.21, 3.024)$, from which we compute the probability of seeing 5 or more events in one month of O2 as $p_1 = 1 - \sum_{n=0}^{4} p(n|d_{O1}) = 0.015$. The probability of seeing 5 or more events in at least one month of O2 is given by $1 - (1 - p_1)^6 = 0.086$. With the Jeffrey's prior the posterior distributions on the rate in 6 months, 5 months and 1 month are $\Gamma(3.5, 0.5)$, $\Gamma(3.5, 0.6)$ and $\Gamma(3.5, 3.0)$ respectively. The probabilities of seeing 6 or more events in O2, 1 or fewer in 5 months of O2, 5 or more in the last month of O2 and seeing 5 or more in at least 1 month of O2 are 0.564, 0.103, 0.019 and 0.110 respectively. The probability that the last month would contain the number of events that were seen is significantly small (at a 2% confidence level). However, there is no reason to single out the last month *a priori* and the probability that one month would be at least this exceptional is only around 10%, which is small but not sufficiently significant to be a cause for concern. The choice of prior does not significantly influence this, indicating that we are data dominated and the conclusion is robust. So, based on O2 we cannot conclude the rate is inhomogeneous in time, but the significance is high enough that we should collect more data and see if the next science run shows any evidence for a time-dependent rate.

(f) In total over O1 and O2 we see 9 events and the total observing time is 0.75 years. Therefore the combined posterior is $\Gamma(9.21, 0.752)$ using the conjugate prior derived in (a) or $\Gamma(9.5, 0.75)$ using the Jeffrey's prior. The posterior predictive distribution for the rate in a given 6 month period of O3 is $\Gamma(9.21, 1.504)$ or $\Gamma(9.5, 1.5)$ respectively. The distribution of the difference $r = |n_1 - n_2|$ of the number of events observed in two independent samples from a Poisson distribution with rate θ is given by the *Skellam distribution* with pmf

$$p(r|\theta) = \begin{cases} \mathrm{e}^{-2\theta} I_0(2\theta) & r = 0 \\ 2\mathrm{e}^{-2\theta} I_r(2\theta) & r = 1, 2 \ldots, \end{cases}$$

where $I_k(x)$ is the modified Bessel function of the first kind. Hence the posterior predictive distribution on r is

$$p(r|d_{1+2}) = \int_0^\infty \mathrm{e}^{-2\lambda} I_r(2\lambda) \frac{\beta^\alpha \lambda^{\alpha-1} \mathrm{e}^{-\beta\lambda}}{\Gamma(\alpha)} \mathrm{d}\lambda$$

for a $\Gamma(\alpha, \beta)$ posterior distribution on the rate. Note that this can also be written as the difference between two independent negative binomial variables, which follow a *generalised discrete Laplace distribution*, but the expressions that must be evaluated are no easier than this integral.

R	$p(r = R\|d_{1+2})$	$p(r \leq R\|d_{1+2})$	R	$p(r = R\|d_{1+2})$	$p(r \leq R\|d_{1+2})$
0	0.123	0.123	6	0.048	0.935
1	0.229	0.352	7	0.029	0.965
2	0.194	0.546	8	0.018	0.983
3	0.158	0.704	9	0.009	0.992
4	0.109	0.813	10	0.003	0.995
5	0.075	0.888	11	0.002	0.997

Table 1: Posterior predictive probability of the absolute difference in the number of events detected in the first and second 6 month periods of the O3 science run. The columns give the difference in the number of events, the posterior probability of observing that difference and the cumulative posterior probability of observing a difference less than or equal to that value.

Table 1 lists the cumulative posterior density for the difference r. We see that there is less than 5% probability of seeing a difference of 7 or more events. Therefore a difference of this size or larger would be significant at a 5% level.

The observed difference in the number of events, 4, is below this threshold and so we conclude that O3 did not give any evidence for the rate varying with time. This value would be significant at the 20% level, but not at 15%. In fact, the result is even less significant, because the sensitivity of the detectors increased and the rate of observed events thus went up by a factor of $4-5$ from ~ 1 per month to ~ 1 per week. The above analysis assumed that the rate was the same in O3 as in O1 and O2. The observed value of 4 is equivalent to a difference of about 1 in the previous calculation, so it is not at all significant.

There are a number of other ways in which this question could be addressed. For example, we could look at the number of events in each month and set a threshold, based on the posterior predictive distribution, on the difference between the largest and smallest monthly count. Alternatively, we could model the rates in each one month period as being potentially different, with λ_i denoting the rate in month i. These rates can be connected by a hyperprior, e.g., $\lambda_i \sim \Gamma(\alpha, \beta)$, and the parameters of that hyperprior constrained from the data. Alternatively, the rates can be modelled parametrically, e.g., $\lambda_i = a + bi$, and the parameters of the parametric model constrained from the data. If the posterior on the slope parameter, b, is inconsistent with 0 there is evidence for an evolving rate. Similarly if the parameters of the hyperprior are inconsistent with a constant rate there is evidence for evolution. The advantage of these kind of approaches is that the results of the analysis give an estimate of the nature and size of the effect, not just the presence of the effect.

Additional questions

9. (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 log-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

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

$$\propto \exp\left[-\frac{n}{2\sigma^2}\left(\mu - \frac{1}{n}\sum_{i=1}^{n}x_i\right)^2\right] \qquad (13)$$

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.

10. (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.

11. (a) We have

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

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

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

(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

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

$$\Rightarrow \quad p(\sigma^2|\mathbf{x}, \mu) = \frac{A^{\frac{n-1}{2}}}{\Gamma\left(\frac{n-1}{2}\right) - \Gamma\left(\frac{1}{T^2}; \frac{n-1}{2}\right)} \sigma^{-(n+1)} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\right] \times \mathbb{I}[0 \leq \sigma^2 \leq T^2] \qquad (14)$$

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.

12. 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

$$\phi_1|\phi_a, \mathbf{x} \sim \text{Beta}\left(1 + \sum_{t=2}^{T+1} x_t, 1 + x_1\right)$$
$$\phi_a|\phi_1, \mathbf{x} \sim \text{Beta}\left(1 + \sum_{j=3}^{T+1}(t-2)x_t, 1 + \sum_{t=2}^{T} x_t\right) \tag{15}$$

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 $i \geq 1$.

13. 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 question (6). Therefore

$$\begin{aligned}\pi(x)\mathcal{K}(x,y) &= \pi(x)q(x,y)\min\left(\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, 1\right) \\ &= \min\left(\pi(y)q(y,x), \pi(x)q(x,y)\right) \\ &= \min\left(1, \frac{\pi(x)q(x,y)}{\pi(y)q(y,x)}\right)\pi(y)q(y,x) = \mathcal{K}(y,x)\pi(y). \end{aligned} \tag{16}$$

As required. Integrating this equation we find

$$\int \pi(x)\mathcal{K}(x,y)\mathrm{d}x = \int \pi(y)\mathcal{K}(y,x)\mathrm{d}x = \pi(y)\int \mathcal{K}(y,x)\mathrm{d}x = \pi(y) \tag{17}$$

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.