

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

Problem Sheet 3: Statistics in Gravitational Wave Astronomy

1. (a) As in the question description we denote the two masses by m_1 and m_2 , the total mass by $M = m_1 + m_2$, the reduced mass by $\mu = m_1 m_2 / M$, and the chirp mass by

$$\mathcal{M}_c = \frac{m_1^{\frac{3}{5}} m_2^{\frac{3}{5}}}{M^{\frac{1}{5}}}.$$

We will use geometric units throughout, i.e., we set $c = G = 1$ so we don't need to worry about keeping track of these factors.

- i. For a Newtonian binary, the motion is equivalent to that of a body of mass μ orbiting in a fixed Newtonian potential with mass M . Denoting the orbital radius by a (it is also the semi-major axis for a circular binary), the orbital frequency is given by

$$2\pi f = \sqrt{\frac{M}{a^3}}$$

and the total energy of the binary is

$$E = -\frac{M\mu}{2a}.$$

- A. The GW amplitude is determined by the quadrupole moment of the spacetime

$$h \sim \frac{\ddot{I}_{jk}}{D}, \quad I_{jk} = \int \rho x_i x_j dV.$$

For a binary, the density is only non-zero at the location of the objects. Using the effective-one-body analogy we deduce

$$I \sim \mu a^2 \exp(2\pi i f t)$$

where the frequency is now twice the orbital frequency because we are taking squares of positions, which vary at that frequency. It follows that

$$h \sim \frac{1}{D} f^2 \mu a^2 \sim \frac{1}{D} f^2 \mu \left(\frac{M}{f^2} \right)^{\frac{2}{3}} = \frac{1}{D} f^{\frac{2}{3}} \frac{m_1 m_2}{M^{\frac{1}{3}}} = \frac{1}{D} \mathcal{M}_c^{\frac{5}{3}} f^{\frac{2}{3}}.$$

- B. The GW energy loss is determined by

$$\dot{E}_{\text{GW}} \sim D^2 \dot{h}^2 = \ddot{I}^2 \sim \mu^2 a^4 f^6 \sim \mu^2 f^6 \left(\frac{M}{f^2} \right)^{\frac{4}{3}} = \mu^2 M^{\frac{4}{3}} f^{\frac{10}{3}} = \mathcal{M}_c^{\frac{10}{3}} f^{\frac{10}{3}}.$$

C. The rate of change of frequency is given by

$$\dot{f} \sim \sqrt{\frac{M}{a}} \frac{d}{dt} \left(\frac{1}{a} \right) \sim \frac{1}{M\mu} \sqrt{\frac{M}{a}} \dot{E} \sim \mu M^{\frac{1}{3}} (Mf)^{\frac{1}{3}} f^{\frac{10}{3}} = \mu M^{\frac{2}{3}} f^{\frac{11}{3}} = \mathcal{M}_c^{\frac{5}{3}} f^{\frac{11}{3}}.$$

D. The Fourier transform of $h(t)$ is given approximately by

$$\tilde{h} \sim \frac{h}{\sqrt{\dot{f}}} \sim \frac{1}{D} \frac{\mathcal{M}_c^{\frac{5}{3}} f^{\frac{2}{3}}}{\mathcal{M}_c^{\frac{5}{6}} f^{\frac{11}{6}}} = \frac{1}{D} \mathcal{M}_c^{\frac{5}{6}} f^{-\frac{7}{6}}.$$

E. The characteristic strain is given by

$$h_c \sim h \sqrt{\frac{f^2}{\dot{f}}} \sim \frac{1}{D} \mathcal{M}_c^{\frac{5}{3}} f^{\frac{2}{3}} \frac{f}{\mathcal{M}_c^{\frac{5}{6}} f^{\frac{11}{6}}} = \frac{1}{D} \mathcal{M}_c^{\frac{5}{6}} f^{-\frac{1}{6}}.$$

F. The energy density of a GW background generated by a population of these sources is given by

$$\rho_c \Omega_{\text{GW}}(f) = \int_0^\infty \frac{N(z)}{1+z} \left(f_r \frac{dE}{df_r} \right)_{f_r=f(1+z)} dz.$$

For the inspiraling binaries the previous results give

$$f \frac{dE}{df} \sim f \frac{\dot{E}}{\dot{f}} \sim \mathcal{M}_c^{\frac{5}{3}} f^{\frac{2}{3}}$$

and so we find

$$\Omega_{\text{GW}}(f) \sim \mathcal{M}_c^{\frac{5}{3}} f^{\frac{2}{3}} \int_0^\infty \frac{N(z)}{(1+z)^{\frac{1}{3}}} dz.$$

ii. The energy of the binary is proportional to $1/a$, hence we have

$$\dot{E}_{\text{hard}} \propto \mu M \frac{d}{dt} \left(\frac{1}{a} \right) = k \mu M \frac{\rho_*}{\sigma^3} \frac{m_2}{a} \propto k \frac{\rho_* m_2 \mu}{\sigma^3} (Mf)^{\frac{2}{3}} = k \frac{\rho m_2 \mu}{\sigma^3} M^{\frac{2}{3}} f^{\frac{2}{3}}.$$

iii. The previous derivation of the background energy density assumed that all of the energy loss driving the frequency evolution was due to GW emission. If there are other processes driving energy loss and hence frequency evolution, the background is suppressed because not all of the orbital energy lost is emitted as gravitational waves. In general we have $f = f(E)$ and hence $\dot{f} = (df/dE) \dot{E}$ and therefore

$$\frac{dE_{\text{GW}}}{df} = \frac{\dot{E}_{\text{GW}}}{(df/dE)[\dot{E}_{\text{GW}} + \dot{E}_{\text{other}}]} = \frac{\dot{E}_{\text{GW}}}{\dot{E}_{\text{GW}} + \dot{E}_{\text{other}}} \left(\frac{dE_{\text{GW}}}{df} \right)_{\text{pure GW}}.$$

The final bracketed expression denotes the background energy density in the pure GW-driven evolution case. In the case of stellar hardening we therefore find a modified expression for the GW background energy density

$$\rho_c \Omega_{\text{GW}}(f) = \mathcal{M}_c^{\frac{5}{3}} f^{\frac{2}{3}} \int_0^\infty \frac{N(z)}{(1+z)^{\frac{1}{3}}} \frac{\mathcal{M}_c^{\frac{10}{3}}}{\mathcal{M}_c^{\frac{10}{3}} + k(\rho m_2 \mu / \sigma^3) M^{\frac{2}{3}} f^{-\frac{8}{3}} (1+z)^{-\frac{8}{3}}} dz.$$

This can be simplified a bit more — for example, we notice that the factor $\mu M^{\frac{2}{3}}$ in the hardening term is just $\mathcal{M}_c^{\frac{5}{3}}$ — but the above result is all we need to answer the next few questions.

- iv. If the sources are at a common redshift, z_0 , we can replace $N(z)$ by a delta function, $\delta(z - z_0)$, and do the integral explicitly. It is then clear that we have

$$\Omega_{\text{GW}}(f) \sim \frac{f^{\frac{2}{3}}}{1 + \lambda f^{-\frac{8}{3}}}$$

where

$$\lambda = k(\rho m_2/\sigma^3) \mathcal{M}_c^{-\frac{5}{3}} (1 + z_0)^{-\frac{8}{3}}.$$

This is a broken power-law, as required. For $f \ll 1$ the term $f^{-\frac{8}{3}}$ dominates in the denominator and we have $\Omega_{\text{GW}} \sim f^{\frac{10}{3}}$. This is the stellar hardening dominated regime. For $f \gg 1$ the constant term dominates in the denominator and we find $\Omega_{\text{GW}} \sim f^{\frac{2}{3}}$. This is the GW dominated regime and this is the standard result for GW backgrounds.

- v. If a broken power law background were detected, it tells us about the processes that drive the inspiral of the binary. In this example the power at low frequencies (where hardening dominates) is suppressed relative to that of a pure GW background (see Figure 1). The low frequency slope is characteristic of whatever process drove the early evolution of the binaries — a measurement of this tells you which physical process was important at that time. The high frequency slope tells us about the late evolution of the binary, and in this case the value $f^{\frac{2}{3}}$ is consistent with GW-driven inspiral. The turn over point tells us about the relative efficiencies of the two processes. In this example it occurs where $f \approx \lambda^{\frac{3}{8}}$ and so a measurement of that value tells us about the parameters that go into λ , such as σ , ρ and the typical source redshift, z_0 .
- vi. (OPTIONAL) No results here. If there is a distribution over masses, then the background energy density involves an integral over the mass distribution as well as the redshift. Try playing around with different choices. Try also including some dependence of ρ and σ on the binary properties. The GW background in the PTA regime may well be suppressed by stellar processes of the type described here. If we see that suppression we will want to be able to interpret it in the context of models of the binary population.
- (b) i. The average waveform power is

$$\langle h^2 \rangle = \frac{1}{2T} \int_{-T}^T h^2(t) dt = \frac{1}{2\sqrt{Q}T} \frac{A^2}{D^2} \int_{-\sqrt{Q}T}^{\sqrt{Q}T} \cos^2\left(\frac{2\pi f_0}{\sqrt{Q}}u\right) e^{-u^2} du.$$

We see that beyond $\sqrt{Q}T \sim \text{few}$, the waveform is exponentially suppressed. Hence, the duration of the signal is order $\sim 1/\sqrt{Q}$. We take $|\sqrt{Q}T| \lesssim 2$ as a reasonable approximation.

For this choice, we find

$$\langle h^2 \rangle = \frac{A^2}{D^2} \frac{\sqrt{\pi}}{8} \left(\text{erf}(2) + e^{-\left(\frac{2\pi f_0}{\sqrt{Q}}\right)^2} \text{Re} \left[\text{erf} \left(2 + i \frac{2\pi f_0}{\sqrt{Q}} \right) \right] \right) \sim \frac{A^2}{D^2}$$

with a pre-factor that is order 0.few.

- ii. Using standard results for Fourier transforms, $\mathcal{F}[g] = \tilde{g}(f)$, including $\mathcal{F}[\exp(-t^2)] = \sqrt{\pi} \exp(-\pi^2 f^2)$, $\mathcal{F}[g(\alpha t)] = \tilde{g}(f/\alpha)/|\alpha|$ and $\mathcal{F}[\exp(2\pi i f_0 t)g(t)] = \tilde{g}(f - f_0)$, we find

$$\tilde{h}(f) = \frac{A}{2D} \sqrt{\frac{\pi}{Q}} \left(e^{-\frac{\pi^2}{Q}(f-f_0)^2} + e^{-\frac{\pi^2}{Q}(f+f_0)^2} \right).$$

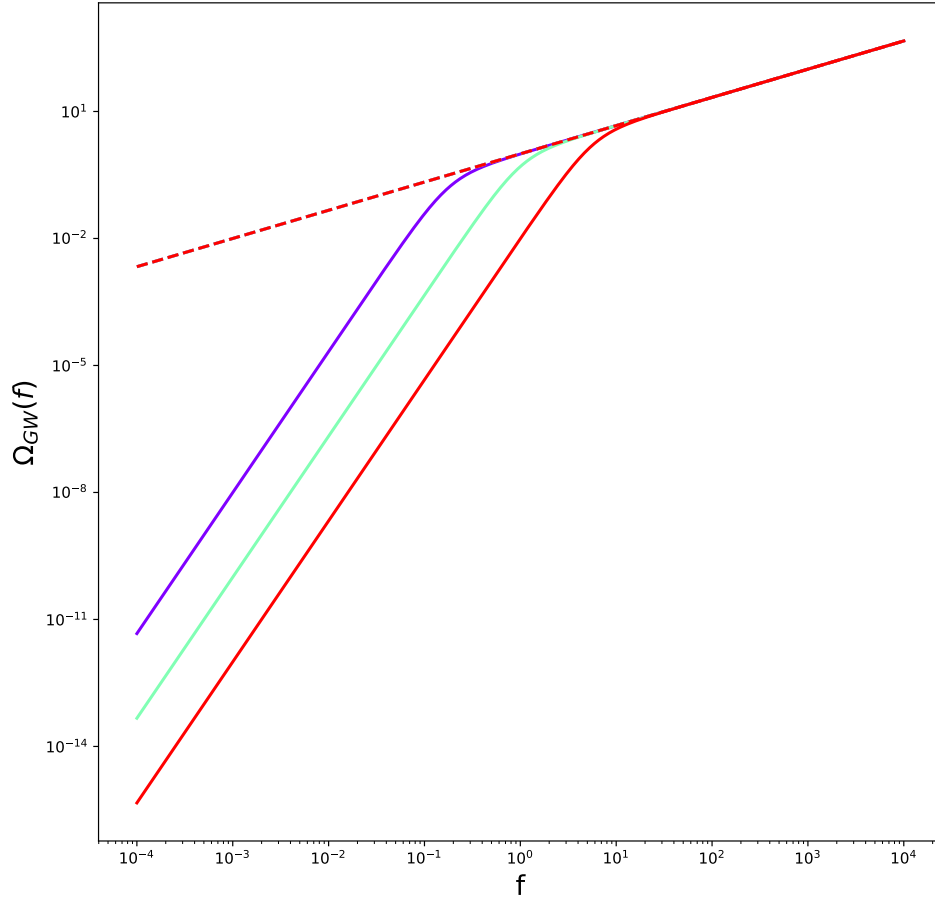


Figure 1: Example backgrounds. We show $\Omega_{\text{GW}}(f)$ as a function of frequency for $\lambda = 0.01$ (purple), $\lambda = 1$ (green) and $\lambda = 100$ (red). Also shown, as a dashed red line, is the background in the absence of stellar hardening.

We can use the fact that the time series is real to wrap onto only positive frequencies and then we have

$$\tilde{h}(f) = \frac{A}{D} \sqrt{\frac{\pi}{Q}} e^{-\frac{\pi^2}{Q}(f-f_0)^2}.$$

We see that the Fourier transform is also proportional to a Gaussian which goes to zero exponentially when $\pi^2(f-f_0)^2/Q \sim \text{few}$. Hence the bandwidth is $\Delta f \sim \sqrt{Q}/\pi$.

iii. Using the power ratio formula

$$\left(\frac{S}{N}\right)^2 \approx \frac{\langle h^2 \rangle}{\Delta f S_n(f)}$$

and assuming white noise, $S_n(f) = \sigma^2$, we have

$$\left(\frac{S}{N}\right)^2 \approx k \frac{A^2}{D^2 \sqrt{Q} \sigma^2}$$

where k is a constant of order unity. This SNR could be achieved by windowing the data (to the time range $|\sqrt{Q}T| \lesssim \text{a few}$) and bandpassing it (to the frequency range $\pi|f-f_0|/\sqrt{Q} \lesssim \text{a few}$) and then comparing the signal power to the average off-source noise power.

iv. Using the Fourier transform obtained above, the matched filtering SNR is

$$\left(\frac{S}{N}\right)^2 = 4 \int_0^\infty \frac{|\tilde{h}(f)|^2}{S_n(f)} df = \frac{4}{\sigma^2} \frac{A^2 \pi}{4D^2 Q} e^{-\frac{2\pi^2}{Q}(f-f_0)^2} df \approx \frac{A^2}{2D^2 \sigma^2 \sqrt{Q}} \int_{-\infty}^\infty e^{-\frac{x^2}{2}} dx$$

which is also equal to $A^2/(D^2 \sigma^2 \sqrt{Q})$ times a constant of order unity.

We have found that the matched filtering SNR is essentially the same as the burst search SNR, so we are not gaining anything by doing matched filtering. We argued in lectures that matched filtering gained over a burst search by a factor of the square root of the number of cycles spent near a particular frequency. These sine-Gaussian sources are peculiar in that as Q decreases so that the source spends more time near frequency f_0 , the bandwidth also decreases so the burst power is increasingly concentrated — we effectively have only ‘1 cycle’ in the vicinity of each relevant frequency. This result does not necessarily mean matched filtering is no better than a burst search — the SNR does not directly translate to a false alarm probability. There may be many instrumental artefacts that could give broadband power in the frequency domain which looks burst like, but those artefacts would look nothing like the specific sine-Gaussian form of the matched filter. Nonetheless, this problem illustrates why excess power searches are quite effective for sources that are burst-like, even if models are available.

v. The energy distribution can be found from

$$\int \frac{dE}{df} df = \int_{-\infty}^\infty D^2 \dot{h}^2(t) dt = \int_{-\infty}^\infty D^2 f^2 \tilde{h}^2(f) df.$$

We find

$$\frac{dE}{df} = A^2 \frac{f^2 \pi}{2Q} \exp\left(-\frac{\pi^2}{Q}(f-f_0)^2\right).$$

- vi. Assuming the number of objects per unit comoving volume with redshift between z and $z + dz$ and with f_0 between f_0 and $f_0 + df_0$ is $N(z)df_0dz$, the background energy density is

$$\rho_c \Omega_{\text{GW}}(f) = \int_0^\infty \int_0^\infty N(z)(1+z)^2 A^2 \frac{f^3 \pi}{2Q} \exp\left(-\frac{\pi^2}{Q}(f(1+z) - f_0)^2\right) f_0^\alpha df_0 dz.$$

- vii. The common redshift assumption allows us to replace the integral over z by evaluation of the integrand at z_0 as before. We then have

$$\rho_c \Omega_{\text{GW}}(f) = N_0(1+z_0)^2 A^2 \frac{\pi}{2Q} f^3 \int_0^\infty \exp\left(-\frac{\pi^2}{Q}(f(1+z_0) - f_0)^2\right) f_0^\alpha df_0 dz.$$

The integral over f_0 takes the form

$$\int_0^\infty x^\alpha \exp[-(x - \lambda f)^2] dx$$

where $\lambda = \pi(1+z_0)/\sqrt{Q}$. This integral can be written down as a combination of hypergeometric functions

$$\begin{aligned} \int_0^\infty x^\alpha \exp[-(x - \lambda f)^2] dx &= \frac{1}{2} e^{-\lambda^2 f^2} \left[\alpha \lambda f \Gamma\left(\frac{\alpha}{2}\right) {}_1F_1\left(\frac{\alpha}{2} + 1; \frac{3}{2}; \lambda^2 f^2\right) \right. \\ &\quad \left. + \Gamma\left(\frac{\alpha+1}{2}\right) {}_1F_1\left(\frac{\alpha}{2} + 1; \frac{1}{2}; \lambda^2 f^2\right) \right]. \end{aligned}$$

The exact background computed from this expression is shown in Figure 2, but we can also find analytic approximations for the low and high frequency behaviour. If $f \ll 1$, then the integral is approximately

$$\int_0^\infty x^\alpha \exp[-x^2] dx = \frac{1}{2} \Gamma\left(\frac{\alpha+1}{2}\right)$$

with corrections of order λf . Hence, the dominant behaviour is a constant and $\Omega_{\text{GW}}(f) \sim f^3$ due to the factor out the front of the expression.

For $f \gg 1$ we can make a change of variable in the integral

$$\begin{aligned} \int_0^\infty x^\alpha \exp[-(x - \lambda f)^2] dx &= \int_{-\lambda f}^\infty (u + \lambda f)^\alpha \exp[-u^2] du \\ &\approx \lambda^\alpha f^\alpha \int_{-\infty}^\infty \left(1 + \frac{u}{\lambda f}\right)^\alpha \exp[-u^2] du \\ &= \sqrt{\pi} \lambda^\alpha f^\alpha \left(1 + O\left(\frac{1}{f}\right)\right). \end{aligned}$$

So we deduce $\Omega_{\text{GW}} \sim f^{3+\alpha}$.

- viii. (OPTIONAL) No results here again, but things to explore would be how the introduction of a redshift distribution modifies things, what happens if the distribution of f_0 is changed, e.g., by introducing a cut-off in the frequency range, what happens if we add a distribution for Q etc.

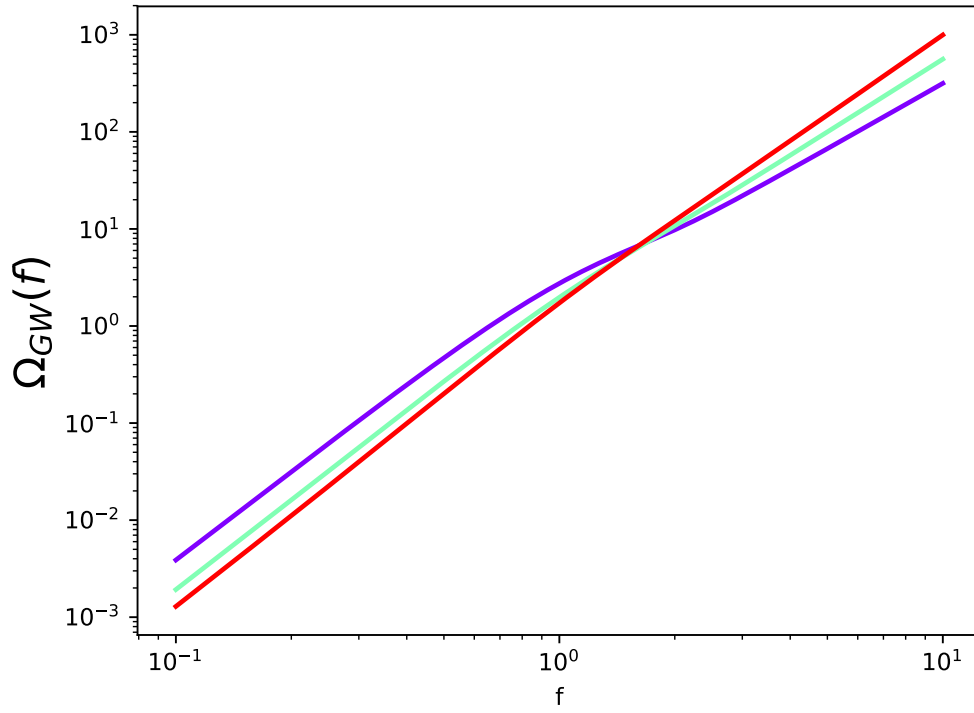


Figure 2: Example backgrounds for the burst population model. We show $\Omega_{\text{GW}}(f)$ as a function of frequency for $\lambda = 1$ and three choices of α : $\alpha = -0.75$ (purple), $\alpha = -0.5$ (green) and $\alpha = -0.25$ (red).

2. (a) The pymc3 model definition for this problem is

with `lin_model`:

```
beta = pm.Normal("beta", mu=mu0, sigma=np.sqrt(var0), shape=2)
tau = pm.Gamma("tau", alpha=a, beta=b)

mu = beta[0] + beta[1] * year

Y_obs = pm.Normal("Y_obs", mu=mu, sigma=1./np.sqrt(tau), observed=jump)
```

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, computed using the `display(az.summary())` command, are shown in Figure 4.

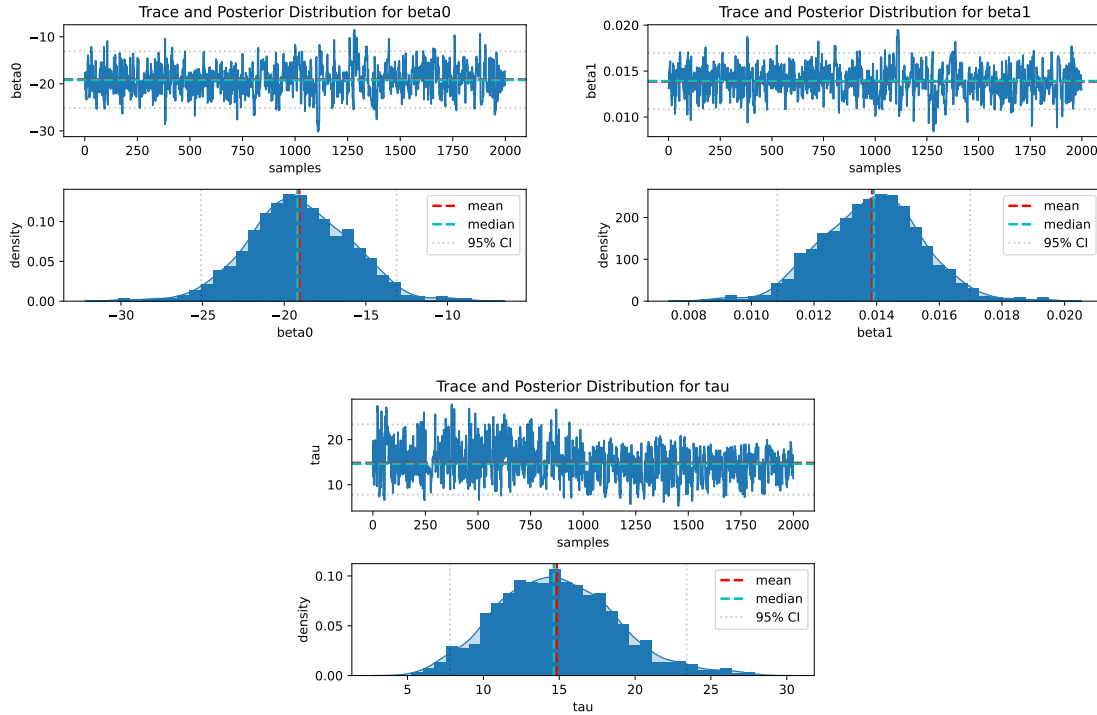


Figure 3: Trace plots and posterior distributions for the linear model fit to the long jump data, for parameters β_0 (top left), β_1 (top right) and $\tau = 1/\sigma^2$ (bottom).

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
beta[0]	-19.07551	3.14669	-24.47930	-13.08199	0.15245	0.10940	431.21715	426.90707	1.01377
beta[1]	0.01386	0.00161	0.01085	0.01666	0.00008	0.00006	431.64718	445.31158	1.01316
tau	14.84132	3.91074	7.19898	21.79108	0.36471	0.27834	131.30007	139.89747	1.04303

Figure 4: Summary table for the linear model fit to the long jump data.

(b) `pymc3` 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 `pymc3` 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 shown in Figure 5. There are a larger number of effective samples in this run, indicating that it is

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
beta[0]	8.00854	0.05312	7.91346	8.10571	0.00116	0.00082	2078.00263	1308.81131	1.00242
beta[1]	0.01409	0.00158	0.01129	0.01724	0.00004	0.00003	1951.82128	1232.09654	1.00217
tau	15.63327	4.47560	7.51565	24.09624	0.09896	0.07386	2108.45621	1124.68092	1.00075

Figure 5: Summary table for the linear model fit to the centred long jump data.

easier to sample from. The result is consistent, with $\hat{\beta}_1 = 0.0141$ compared to $\hat{\beta}_1 = 0.0139$ in the non-centred case.

- (c) The `pymc3` model for robust regression with fixed student-t degrees of freedom is

```
robust_model = pm.Model()

mu0=0.
var0=1000.
a=0.1
b=0.1
nu=3

with robust_model:

    beta = pm.Normal("beta", mu=mu0, sigma=np.sqrt(var0), shape=2)
    tau = pm.Gamma("tau", alpha=a, beta=b)

    mu = beta[0] + beta[1] * year_cent

    Y_obs = pm.StudentT("Y_obs", nu=nu, mu=mu, sigma=1./np.sqrt(tau), observed=jump)
```

and the summary table from fitting this model with $\nu = 3$ is shown in Figure 6. The new estimate of the slope coefficient is $\hat{\beta}_1 = 0.01394$.

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
beta[0]	7.99801	0.04776	7.90466	8.08585	0.00105	0.00074	2096.23516	1419.18743	1.00138
beta[1]	0.01394	0.00139	0.01153	0.01677	0.00003	0.00002	2571.04073	1481.04443	1.00057
tau	25.19306	8.82742	9.30647	40.23682	0.19984	0.15264	2114.42835	1343.20320	1.00319

Figure 6: Summary table for the robust linear model fit to the centred long jump data, with fixed degrees of freedom, $\nu = 3$.

To allow the degrees of freedom to vary we use the pymc3 model

```
robust_model_B = pm.Model()

mu0=0.
var0=1000.
a=0.1
b=0.1
c=0.1
d=0.1

with robust_model_B:

    beta = pm.Normal("beta", mu=mu0, sigma=np.sqrt(var0), shape=2)
    tau = pm.Gamma("tau", alpha=a, beta=b)
    nu = pm.Gamma("nu", alpha=c, beta=d)

    mu = beta[0] + beta[1] * year_cent

    Y_obs = pm.StudentT("Y_obs", nu=nu, mu=mu, sigma=1./np.sqrt(tau), observed=jump)
```

and the result table from fitting this model is shown in Figure 7. The new

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
beta[0]	8.00013	0.05174	7.90109	8.09468	0.00110	0.00078	2236.86928	1412.53794	1.00317
beta[1]	0.01394	0.00137	0.01173	0.01688	0.00003	0.00002	1731.33443	1343.89161	1.00103
tau	22.28845	8.39863	7.80234	37.21323	0.22789	0.16885	1489.62043	1209.27159	1.00026
nu	7.83173	6.55233	1.04216	18.71523	0.19047	0.13472	1551.14647	1331.78205	0.99920

Figure 7: Summary table for the robust linear model fit to the centred long jump data, with variable degrees of freedom.

estimate of the slope coefficient is now $\hat{\beta}_1 = 0.01394$. 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.

3. (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 $N(1, 1/3)$ 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

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

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

- (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 $N(5, 4/3)$. 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 UK 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

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

where

$$w' = \frac{k_1w}{k_1w + k_2(1-w)}, \quad k_i = \frac{1}{\sqrt{\sigma^2 + n\sigma_i^2}} \exp\left[-\frac{1}{2} \left(\frac{n(\bar{y} - \mu_i)^2}{\sigma^2 + n\sigma_i^2}\right)\right]. \quad (2)$$

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 `pymc3` model definition is as follows

```
npts=10
y=np.array([-0.566,  3.74,  5.55, -1.90, -3.54, 5.16, -1.76, 4.08, 4.62, 0.732])

# Specify prior hyperparameters
# Mixture prior on linear model coefficients.
mup=np.array([1.,5.0])
taup=np.array([3.0,0.75])
wt=np.array([2./3.,1./3.])

# Prior on precision
a=0.01
b=0.01

# Define pymc3 model
mixture_model = pm.Model()

with mixture_model:

    mu = pm.NormalMixture("mu", w=wt, mu=mup, tau=taup)
    tau = pm.Gamma("tau", alpha=a, beta=b)

    Y_obs = pm.Normal("Y_obs", mu=mu, sigma=1./np.sqrt(tau), observed=y)
```

The output table after fitting the model is given in Figure ?? and the resulting posteriors and trace plots are shown in Figure 9. Note that you will not get exactly these values due to sampling error, but your values should be close to these.

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
mu	1.26053	0.75868	-0.07840	2.41142	0.00907	0.00751	13268.36775	5062.13866	1.00047
tau	0.09221	0.04248	0.02367	0.17227	0.00029	0.00021	20246.06095	19667.50197	1.00015

Figure 8: Summary table for the Normal model fit with the Normal mixture prior.

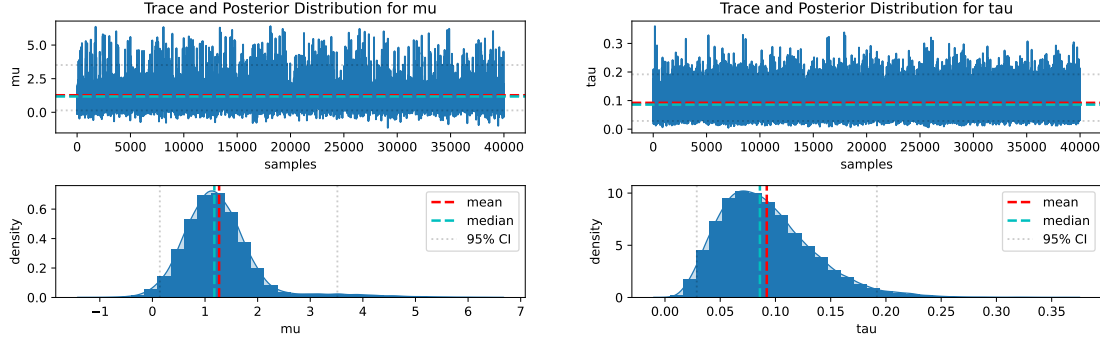


Figure 9: Posterior distributions and trace plots for the mean μ (left) and precision $\tau = 1/\sigma^2$ (right) of the log concentration of the chemical.

- (d) The probability that $\mu < 1$ can be found by integrating the posterior for μ from $-\infty$ to 1. This can be done by computing the fraction of posterior samples that have $mu < 1$. We obtain an estimate $p = 0.370$.

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 probability of obtaining a negative measurement in a single future observation. This is accomplished by adding this line to the model definition:

```
ypred = pm.Normal("ypred",mu=mu,sigma=1./np.sqrt(tau))
```

and then computing the fraction of posterior samples with $y_{\text{pred}} < 0$. This gives $p_{-,1} \approx 0.359$.

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 $p_{-,1} = 0.359$ we find $p_{-,5} \approx 0.892$.

- (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. (1) above, but with w' and $(1 - w')$ replaced by

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

with k_i as defined in Eq. (2). In this case we find the joint posterior is

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

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

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

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.

Additional questions

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

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

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

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

Taking $a = 9.2$, $b = 13.8$, $n = 20$, and $y = 15$, results in a $\text{Beta}(24.2, 18.8)$ 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 $\text{Beta}(1, 1)$ 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

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

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

- (g) We start by calculating the prior predictive distribution

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

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 Beta(12, 3) prior.
 ii. The mixture prior $\theta \sim \pi \text{Beta}(a_1, b_1) + (1 - \pi) \text{Beta}(a_2, b_2)$ is plotted in Figure 10.

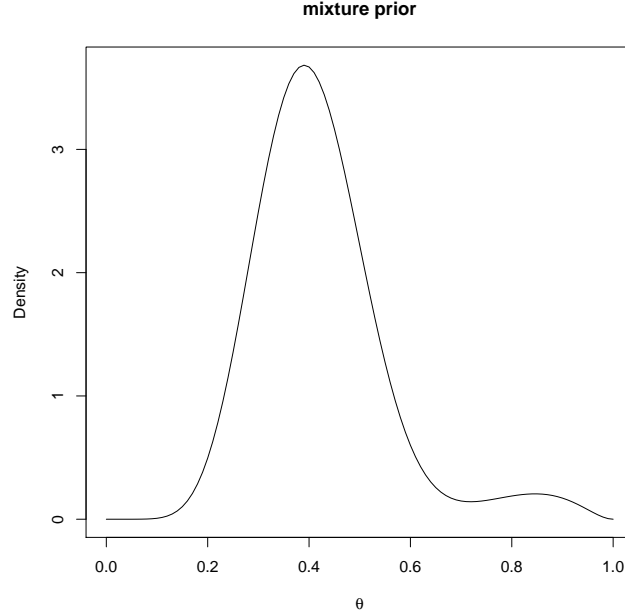


Figure 10: The mixture prior for question 4(h)(ii).

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

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

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

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

with

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

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,

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

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 11.

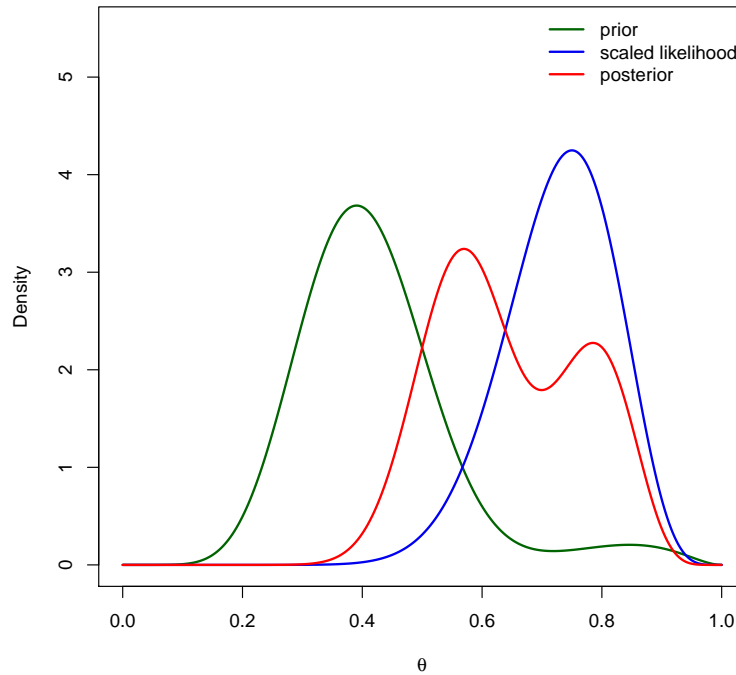


Figure 11: Comparison of prior, likelihood and posterior for question 4(h)(v).