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

Problem Sheet 3: Statistics in Gravitational Wave Astronomy

IMPRS students taking this course should complete the questions in the first part of this sheet and hand them in to be marked. The questions in the second part of the sheet, labelled "Additional questions', are for personal study and do not need to be handed in.

1. This question is focussed on calculating properties of gravitational wave source populations. For two different classes of source you will estimate properties such as the characteristic strains, signal-to-noise ratios and background spectrum generated by a population of such sources. There are two sub-parts to this question. At the end of each there is an optional part which is more open ended and involves some numerical investigation.

   (a) This question is concerned with properties of a background generated by a population of massive black holes. For the first few parts of this question you can assume that these binaries are identical, i.e., they all have the same values for the two masses, m_1 and m_2, and hence the derived quantities of total mass, $M = m_1 + m_2$, reduced mass, $\mu = m_1 m_2/M$, and chirp mass,

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

   In the final part of this question you'll be asked to think about and play around with different assumptions about the mass distribution in the binaries.

   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. Assuming that the binary is Newtonian and circular, derive the following characteristic properties of the emitted gravitational waves [1].

         A. The GW amplitude scales like

         $$h \sim \frac{1}{D} \mathcal{M}_c^{\frac{5}{3}} f^{\frac{2}{3}}.$$

         B. The GW energy loss scales like

         $$\dot{E}_{\rm GW} \sim \mathcal{M}_c^{\frac{10}{3}} f^{\frac{10}{3}}.$$

         C. The rate of change of frequency scales like

         $$\dot{f} \sim \mathcal{M}_c^{\frac{5}{3}} f^{\frac{11}{3}}.$$

[1] You may find it useful to recall that the energy of a Newtonian binary of semi-major axis a is $E = -M\mu/(2a)$ and the frequency is related to the semi-major axis via $2\pi f = \sqrt{M/a^3}$.

D. The Fourier transform of $h(t)$ scales like

$$\tilde{h} \sim \frac{1}{D}\mathcal{M}_c^{\frac{5}{6}} f^{-\frac{7}{6}}.$$

E. The characteristic strain scales like

$$h_c \sim \frac{1}{D}\mathcal{M}_c^{\frac{5}{6}} f^{-\frac{1}{6}}.$$

F. The energy density of a GW background generated by a population of these sources scales like

$$\Omega_{\rm GW}(f) \sim \mathcal{M}_c^{\frac{5}{3}} f^{\frac{2}{3}}$$

In the above f denotes the orbital frequency of the binary and D the distance to the source.

ii. We now suppose that there is an additional process driving the evolution of the binaries, stellar hardening. This leads to an evolution of the semi-major axis, a, of the binary of the form

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{a}\right) = k\frac{\rho_*}{\sigma^3}\frac{m_2}{a}$$

where k is a numerical constant, ρ_* is the stellar density and σ is the velocity dispersion of the stars. Show that this equation implies the energy loss from stellar hardening scales like

$$\dot{E}_{\rm hard} \propto \frac{\rho_* m_2 \mu}{\sigma^3} M^{\frac{2}{3}} f^{\frac{2}{3}}. \tag{1}$$

iii. Derive an expression for the corresponding GW background energy density, $\Omega_{\rm GW}(f)$.

iv. Assuming that the sources are at a common redshift (or the population is dominated by sources at a particular redshift), show that the spectrum is a broken power-law and find the asymptotic slopes at low and high frequency.

v. If a broken power law background were detected, what would it tell you about the relevant physical processes influencing the population? How is that information encoded in the background?

vi. (OPTIONAL) Try varying the assumptions about the population, e.g., the masses of the binary components or the properties of the stellar system or the redshift distribution of the sources. Compute the background from such populations numerically and explore how these assumptions change the results qualitatively and quantitatively.

(b) We now consider a population of burst sources. We will represent these as sine-Gaussians with a four parameter waveform family

$$h(t) = \frac{A}{D}\cos(2\pi f_0 t)\mathrm{e}^{-\frac{1}{2}Qt^2}.$$

i. Find the average waveform power

$$\langle h^2 \rangle = \frac{1}{2T}\int_{-T}^{T} h^2(t)\mathrm{d}t$$

and show that, for a reasonable choice of T, e.g., $T = 2/\sqrt{Q}$, it scales like A^2/D^2 as expected. Why is this a 'reasonable choice' for T?

ii. Compute the Fourier transform of $h(t)$. Hence deduce the bandwidth of the source is $\Delta f \sim \sqrt{Q}/\pi$.

iii. Estimate the SNR that might be possible in a burst search using

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

You may assume white noise, $S_n(f) = \sigma^2$, for simplicity.

iv. Compute the SNR that would be obtained if the source was found by matched filtering and compare it to the burst search SNR. Comment on your answer.

v. Compute the energy distribution for one of these bursts, $\mathrm{d}E/\mathrm{d}f$.

vi. Find an expression for the GW background energy density produced by a population of these bursts, assuming that A and Q are constant and the distribution of f_0 is a power law $\mathrm{d}n/\mathrm{d}f_0 \propto f_0^\alpha$ with $\alpha > -1$.

vii. Assuming the sources are at a common redshift, show that the asymptotic slope of the background is f^3 at low frequency and $f^{3+\alpha}$ at high frequency.

viii. (OPTIONAL) Use simulations on a computer to explore how things change under modifications of these various assumptions.

2. We have data on the winning men's long jump distances (m) from 1900 through 2008, as follows

$$\begin{aligned}\text{Year} = \{&1900, 1904, 1906, 1908, 1912, 1920, 1924, 1928, 1932, 1936, 1948, 1952, 1956,\\ &1960, 1964, 1968, 1972, 1976, 1980, 1984, 1988, 1992, 1996, 2000, 2004, 2008\},\\ \text{Jump} = \{&7.185, 7.340, 7.200, 7.480, 7.600, 7.150, 7.445, 7.730, 7.640, 8.060, 7.825, 7.570,\\ &7.830, 8.120, 8.070, 8.900, 8.240, 8.350, 8.540, 8.540, 8.720, 8.670, 8.500, 8.550,\\ &8.590, 8.370\}.\end{aligned}$$

Fit a linear regression of the distances as a function of Olympic year:

$$Jump = \beta_0 + \beta_1 \text{Year} + \epsilon$$

assuming $\epsilon \sim N(0, \sigma^2 = 1/\tau^2)$. Use the following priors for the three parameters:

$$\begin{aligned}\beta_0, \beta_1 &\sim \text{Normal}\,(\mu_0 = 0, \tau_0 = 0.001)\\ \tau &\sim \text{Gamma}\,(a = 0.1, b = 0.1)\end{aligned}$$

(a) Fit the model using `pystan`. Generate trace plots, posterior distributions and summary statistics. Compute also the autocorrelation function, the effective sample size and the Gelman-Rubin statistic.

(b) Try centring the "Year" covariate, i.e., subtract its average value and use the centred covariate as the explanatory variable for sampling.

(c) Now repeat the analysis using "robust regression" by replacing the Normal distribution with a Student-t distribution. Try fixing the degrees of freedom to $\nu = 3$ and allowing this to be a model parameter with a suitable Gamma prior. How do your results change?

3. A study is conducted to measure the log-concentration of a particular chemical in soil. It can be assumed that the log-concentration follows a Normal distribution with mean μ and variance σ^2. A set, $\{y_i\}$, of $n = 10$ measurements are made of the log-concentration in soils in a certain region of the UK with the following results

$$-0.566, 3.74, 5.55, -1.90, -3.54, 5.16, -1.76, 4.08, 4.62, 0.732.$$

The primary parameter of interest is μ, the unknown mean of the distribution. We assume initially that $\sigma^2 = 30$ is known.

(a) We consult a panel of 10 UK experts and they believe that $\mu \approx 1$, but could take values in the range $[0, 2]$. Construct a suitable conjugate prior based on this information and obtain the posterior distribution from the experimental data. Include a brief justification of your choice of prior.

(b) Suppose now that we had also consulted a panel of 5 US experts, and their opinion was that $\mu \approx 5$ but could be between 3 and 7. Derive a suitable mixture prior that combines the opinions of both sets of experts and obtain the posterior distribution for this new prior.

(c) Now suppose that σ^2 is also unknown. Using a suitable prior on the precision, $\tau = 1/\sigma^2$, and the same mixture prior on μ, obtain samples from the posterior distribution numerically. You may wish to use the various convergence diagnostics that were discussed in lectures to verify the accuracy of your posterior distribution, but you do not need to report the results of such checks in your solution. Obtain estimates of the posterior mean, median, and the lower and upper quartiles from your samples.

(d) Based on your results, what is the posterior probability that $\mu < 1$? If we take 5 additional measurements, what is the probability that at least one of them will return a negative log-concentration?

(e) In building the mixture prior in part (b) you would have used some weighting between the two groups of experts, $p(\mu) = wp_1(\mu) + (1 - w)p_2(\mu)$. We will now include the weight w as an additional model parameter. In the case that $\sigma^2 = 30$ is known, obtain the combined posterior distribution on (μ, w), using a flat prior for $w \in [0, 1]$. Obtain also the marginal distributions on μ and w and comment on the result.

Additional questions

4. **Analysis of binomial data: drug**. Consider the example from lecture 5 where a new drug is being considered for relief of chronic pain, with the success rate θ being the proportion of patients experiencing pain relief. In the past, drugs of this type have shown variable pain relief rates, with a mean of 40% and a standard deviation of 10%. We have seen that these could be translated into a Beta$(9.2, 13.8)$ distribution. This drug had 15 successes out of 20 patients.

   (a) Calculate the posterior distribution of the success rate θ.

   (b) What is the posterior mean and 95% highest posterior density (HPD) interval for the response rate?

   (c) Compute a symmetric 95% credible interval. Compare this to the 95% HPD interval.

   (d) What is the probability that the true success rate is greater than 0.6?

   (e) How is this value affected if a uniform prior is adopted? And how is it affected in the case that Jeffreys' prior is adopted?

   (f) Using the original Beta$(9.2, 13.8)$ prior, suppose 40 more patients were entered into the study. What is the chance that at least 25 of them experience pain relief?

   (g) We might ask whether the observed data is 'compatible' with the expressed prior distribution. One method is to calculate the predictive probability of observing such an extreme number of successes under this prior: this is a standard p-value but where the null hypothesis is a distribution. Use the predictive distribution for 20 future patients to find the probability of getting at least 15 successes (i.e., at least 15 patients experiencing pain relief). Do you think this suggests the data are incompatible with the prior?

   (h) Suppose that most drugs (95%) are assumed to come from the stated Beta$(9.2, 13.8)$ prior, but there is a small chance that the drug might be a 'winner'. 'Winners' are assumed to have a prior distribution with mean 0.8 and standard deviation 0.1.

      (i) What Beta distribution might represent the 'winners' prior? Remember that a Beta(a, b) distribution has mean $a/(a+b)$ and variance $ab/\{(a+b)^2(a+b+1)\}$.

      (ii) Plot the mixture prior.

      (iii) What is now the chance that the response rate is greater than 0.6?
      Hint: You might start by showing that if

$$\theta \sim \pi \text{Beta}(a_1, b_1) + (1-\pi)\text{Beta}(a_2, b_2),$$

      then

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

      where

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

      Here y denotes the number of successes.

(iv) For this mixture prior, repeat the prior/data compatibility test performed previously. Are the data more compatible with this mixture prior?

(i) Repeat the above analysis numerically using `pystan`.

(i) Compute the posterior mean, standard deviation and a 95% credible interval. Compare with the exact results.

(ii) What is the probability that the true success rate is greater than 0.6. Compare with the exact result.

(iii) Suppose 40 more patients were entered into the study. What is the chance that at least 25 of them experience pain relief? Compare with the exact result.

(iv) Conduct the 'prior/data compatibility check', i.e., calculate the predictive probability of observing at least 15 successes under this prior. Compare with the exact result.

(v) For the mixture prior, what is now the chance that the response rate is greater than 0.6? Compare with the exact result.

(vi) Under this mixture prior, what is the posterior predictive probability that at least 25 out of 40 new patients experience pain relief?

(vii) For this mixture prior, repeat the prior/data compatibility test performed previously. Are the data more compatible with this mixture prior? Compare with the exact result.