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

Problem Sheet 2: Bayesian Statistics

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. A motorist travels regularly from Berlin to Golm. On each occasion he chooses a route at random from four possible routes. From experience, the probabilities of completing the journey in under 1 hour via these routes labelled 1 to 4 are 0.2, 0.5, 0.8 and 0.9 respectively. Given that on a certain occasion they complete the journey in under 1 hour, calculate the probability that they travelled on each of the possible routes.

2. (a) Consider the general hypothesis testing problem

$$H_0 : \theta \in \Theta_0 \quad \text{vs} \quad H_1 : \theta \in \Theta_1,$$

such that the union of Θ_0 and Θ_1 is the whole of the parameter space Θ. Letting p_0 and p_1 denote the prior probabilities for the null hypothesis and the alternative hypothesis respectively, show that the posterior probability of H_0 is given by

$$\mathbb{P}(H_0|\mathbf{x}) = \mathbb{P}(\theta \in \Theta_0|\mathbf{x}) = \frac{p_0}{p_0 + p_1/B_{01}},$$

where B_{01} denotes the Bayes factor of H_0 to H_1.

(b) Now suppose that we observe data $\mathbf{x} = \{x_1, \ldots, x_n\}$, such that

$$X_i \sim^{iid} N(\mu, \sigma^2),$$

where σ^2 is known. We wish to test

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu = \mu_1.$$

Show that the Bayes factor is given by

$$B_{01} = \exp\left(-\frac{n(\mu_0 - \mu_1)(\mu_0 + \mu_1 - 2\bar{x})}{2\sigma^2}\right).$$

Calculate the Bayes factor for H_0 against H_1 when $\mu_0 = 0$, $\mu_1 = 1$, $\sigma^2 = 1$, $n = 9$ and $\bar{x} = 0.645$. What is your conclusion? What happens as we increase n, with all other values fixed?

(c) Finally, suppose that we observe data $\mathbf{x} = \{x_1, \ldots, x_n\}$, such that

$$X_i \sim^{iid} N(\mu, \sigma^2),$$

as before. Now we want to test the hypotheses:

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0.$$

We specify $p(\mu|H_1) \sim N(0, \tau^2)$. Calculate the Bayes factor for H_0 against H_1. Comment on the limiting case were we make the prior on μ under H_1 increasingly vague, i.e., $\tau^2 \to \infty$.

3. We observe data $\mathbf{x} = \{x_1, \ldots, x_m\}$ from a multinomial distribution, $\mathbf{X} \sim$MN$(N, \mathbf{p})$, and wish to make inference on the parameters $\mathbf{p} = \{p_1, \ldots, p_m\}$ (note that $\sum_i p_i = 1$). We set a prior on the unknown parameters $\mathbf{p}$ of the form

$$\mathbf{p} \sim \text{Dir}(\alpha_1, \ldots, \alpha_m).$$

   (a) Determine the corresponding posterior distribution for the parameters $\mathbf{p}$.

   (b) Calculate the Bayes estimate for the parameters, assuming a quadratic loss function. [Hint: the Bayes estimate with quadratic loss function is the posterior mean.]

   (c) We throw a 6-sided die 60 times and record the number of times that we observed the number $i = 1, \ldots, 6$, which we denote by x_i. Let p_i denote the associated probability of throwing the number $i = 1, \ldots, 6$ and set $\mathbf{p} = \{p_1, \ldots, p_6\}$. We observe the data $\mathbf{x} = \{10, 12, 12, 8, 7, 11\}$ and specify a Uniform prior on $\mathbf{p}$, which is Dir$(\alpha_1, \ldots, \alpha_6)$ for $\alpha_i = 1$ for $i = 1, \ldots, 6$. Determine the posterior mean for each p_i.

4. Suppose that we wish to simulate observations from the Pareto distribution with pdf

$$p(\theta) = \begin{cases} \frac{a x_o^a}{\theta^{a+1}} & \text{for } \theta \geq x_0; \\ 0 & \text{otherwise,} \end{cases}.$$

Derive the cumulative distribution function for θ and hence describe an algorithmic procedure for sampling random variables from $p(\theta)$ using the method of inversion.

5. A biologist is interested in estimating the annual survival probability of a given species of deer, denoted by ϕ. Data are collected via a radio-tagging experiment which initially places radio-tags on a total of N animals in year 0. Let p_t denote the probability an individual dies within the interval $(t-1, t]$ for $t = 1, \ldots, T$ and p_{T+1} the probability that they survive until the end of the study. Assuming survivability in each year is independent, and animals are independent of each other, we have

$$p_t = \begin{cases} (1-\phi) & t = 1 \\ (1-\phi)\phi^{t-1} & t = 2, \ldots, T \\ \phi^T & t = T+1 \end{cases}.$$

Let X_t denote the number of individuals that die in the interval $(t-1, t]$ for $t = 1, \ldots, T$ and X_{T+1} the number of individuals that survive until the end of the study. The corresponding likelihood is

$$p(\mathbf{x}|\phi) = \frac{N!}{\prod_{t=1}^{T+1} x_t!} \prod_{t=1}^{T+1} p_t^{x_t}.$$

Finally we specify the prior $\phi \sim$Beta(α, β).

(a) Show that the posterior distribution for the survival probability is of the form

$$p(\phi|\mathbf{x}) \sim \text{Beta}\left(\alpha + \sum_{t=1}^{T+1}(t-1)x_t, \beta + \sum_{t=1}^{T} x_t\right).$$

(b) To obtain a set of sampled realisations from the Beta distribution of interest the biologist intends to implement a rejection sampling algorithm. However, due to their limited computational skills they are only able to simulate random deviates from a $U[a, b]$ distribution. Describe a rejection sampling algorithm that the biologist can implement using their limited computational skills.

(c) We specify the prior $\phi \sim$Beta$(1, 1)$ and observe data such that the posterior $p(\phi|\mathbf{x}) \sim$Beta$(91, 9)$. We obtain the following posterior summary statistics for ϕ: posterior mean $\mathbb{E}(\phi) = 0.910$, posterior standard deviation 0.028, posterior median 0.913, 95% symmetric credible interval of $(0.847, 0.958)$ and a 95% highest posterior density interval of $(0.832, 0.949)$. Without conducting an analysis state why at least one of these summary statistics must be incorrect.

6. Suppose that we wish to use the Metropolis-Hastings algorithm to generate a sample from $N(0, \sigma^2)$, and that we use the proposal $q(x, y) = N(ax, \tau^2)$ for $-1 < a < 1$.

   (a) What is the corresponding acceptance probability $\alpha(x, y)$?

   (b) For what value of τ^2 would this particular sampler never reject the candidate value?

   (c) What happens if $a = 0$?

7. Consider a simplified model of PSD estimation. We assume that we have made N observations of noise, n_i, drawn from a zero mean Gaussian with common, but unknown, standard deviation σ. The $N + 1$'th observation comprises noise, drawn from the same distribution, and a signal, represented by a non-zero value of the mean, A. Write down the combined posterior distribution for σ and A, using flat priors. Marginalise over σ to obtain the posterior on A and comment on the result.

8. The Laser Interferometer Ground Observatory (LIGO) detected gravitational waves for the first time in September 2015. LIGO has now completed three observing runs. The first run (O1) lasted 3 months, during which time 3 signals from binary black hole mergers were observed. The second observing run (O2) lasted 6 months. In the first 5 months, one additional merger was observed, and then in the last months 5 further signals were detected. We may assume that the events are distributed according to a homogeneous Poisson distribution with parameter λ with units of yr^{-1}. We are interested in inferring the value of λ. Prior to O1 the value of λ was poorly constrained, with estimated values ranging from 0.01 to 1000.

   (a) Consider the information available prior to the first observing run and construct a conjugate prior for the rate parameter. Briefly justify your reasons for constructing a prior in this way.

   (b) Derive the posterior distribution for λ using the O1 observations. Report the posterior mean, standard deviation, a 95% symmetric confidence interval and plot the posterior distribution.

   (c) What is the posterior probability that the rate is > 15?

(d) Re-analyse the O1 data using a Jeffreys prior; how do your results in (b) and (c) change?

(e) Based on the posterior from the O1 data, what is the posterior predictive probability that we would see 6 or more events in O2? What is the posterior predictive probability that we would see 1 or fewer events in the first 5 months of O2? What is the posterior predictive probability that we would see 5 or more events in the last month of O2/during at least one month of O2? How are these results affected by the choice of prior? Some authors have claimed that the LIGO results provide evidence that the rate is not homogeneous in time. Based on these results, do you agree?

(f) The third observing run, O3, started in April 2019 and lasted for one year. Prior to the start of O3, you update the posterior distribution to use all of the events from O1 and O2 (using one of the previous prior choices). Obtain the corresponding posterior predictive distribution for the difference, $|n_2 - n_1|$, in the number of events that will be observed in the first 6 months, n_1, and the last 6 months, n_2, of O3. How large would the observed difference have to be to provide evidence that the rate is changing with time? The actual observed difference is 4. What do you conclude? Discuss other possible ways to address the question 'is the rate changing with time?' within a Bayesian framework.

Additional questions

9. Let $X_1, \ldots, X_n$ be independent and identically distributed random variables such that $X_i \sim N(\mu, \sigma^2)$ for $i = 1, \ldots, n$, where σ^2 is known.

   (a) Show that Jeffreys' prior for μ is of the form $p(\mu) \propto 1$ for $-\infty < \mu < \infty$.

   (b) Hence show that the posterior distribution for μ is also normal, with mean and variance to be specified.

   (c) Suppose that we observe data $x_1, \ldots, x_{10}$ such that $\bar{x} = 10.1$. Assuming that $\sigma^2 = 1$, show that a 95% highest posterior density interval for μ is $(9.480, 10.720)$.

10. A chemist is interested in the maximum possible yield produced by a certain process. Due to the large variability in the data, they assume that, given a scalar θ, each yield x_i, $i = 1, \ldots, n$, is independent of the other yields and follows a uniform distribution $U[0, \theta]$, so that

$$p(x_i|\theta) = \frac{1}{\theta}, \qquad \text{for } 0 < x_i < \theta.$$

   Before the chemist sees any data, they assume a Pareto prior distribution for θ, so that

$$p(\theta) = \begin{cases} \frac{a x_o^a}{\theta^{a+1}} & \text{for } \theta \geq x_0; \\ 0 & \text{otherwise,} \end{cases}$$

   where $a > 0$ and $x_o > 0$ are known parameters for the prior Pareto distribution, specified by the chemist. Note that the mean of a Pareto distribution is given by $a x_0/(a-1)$, for $a > 1$, whilst the median if $x_0 2^{1/a}$.

   (a) Calculate the posterior distribution of θ.

   (b) Suppose that the chemist specifies a Pareto prior distribution with $a = 2$, $x_0 = 0.1$. Consider observed data $\mathbf{x} = \{x_1, x_2, x_3\} = \{3, 10, 17\}$. Obtain the posterior distribution and indicate how the expert's beliefs have changed after observing the data, using point summary statistics.

   (c) Suppose instead that the chemist specified the alternative prior $\theta \sim U(0, 15)$. What are the implications for the given observed data?

11. Suppose that $X_1, \ldots, X_n \sim^{iid} N(\mu, \sigma^2)$ where both μ and σ^2 are unknown. We specify the priors

$$\mu \sim N(0, s^2), \qquad \text{and } \sigma \sim U[0, T],$$

   where T is "large".

   (a) Using a transformation of variables, calculate the corresponding prior on σ^2.

   (b) Calculate the posterior conditional distribution of μ and σ^2 (i.e., the posterior distribution for μ, treating σ^2 as fixed, and the posterior distribution of σ^2, treating μ as fixed).

12. Radio-tagging data involves placing a radio-tag on a number of individuals and (assuming no radio failures) recording the number of deaths that occur at a series of successive "capture" times. We assume that only a single radio-tagging event occurs where a total of n lambs are "tagged". We let x_t denote the number of sheep that are subsequently recorded as having died within the interval $(t-1, t]$ (assuming tagging

occurs at time 0), for $t = 1, \ldots, T$. We let x_{T+1} denote the number of individuals that survive until time T (i.e. the end of the study). The corresponding likelihood function is a function of the survival probabilities of the sheep. We assume two distinct survival probabilities: ϕ_1 corresponding to first-year survival probability and ϕ_a the "adult" survival probability (i.e. older than first-years). The likelihood is given by

$$p(\mathbf{x}|\phi_1, \phi_a) \propto \prod_{i=1}^{T} p_i^{x_i}$$

where

$$p_i = \begin{cases} 1 - \phi_1 & i = 1 \\ \phi_1(1 - \phi_a) & i = 2 \\ \phi_1 \phi_a^{i-2}(1 - \phi_a) & i = 3, \ldots, T \\ \phi_1 \phi_a^{T-1} & i = T + 1 \end{cases}.$$

Without any prior information on ϕ_1 or ϕ_a we set priors $\phi_1 \sim U[0, 1]$ and $\phi_a \sim U[0, 1]$ independently. Describe a Gibbs sampling algorithm for obtaining a sample from the posterior distribution. Comment on the result.

13. Show that the Metropolis-Hastings algorithm for target distribution $\pi(x)$ generates a reversible Markov chain, such that for $x \neq y$

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

where $\mathcal{K}_H(x, y) = q(y|x)\alpha(x, y)$ and $\alpha(x, y)$ is the acceptance probability.

Hence show that

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