The Lighthouse Problem

Christopher J. Moore

Jan 2024

The coursework will be submitted via a GitLab repository which we will create for you. You should place all your code and your report in this repository. You will be provided access to the repository until 11:59pm on Thursday the 28th of March. After this you will lose access which will constitute submission of your work.

The report (limit 3000 words) should be in PDF format in a folder called `report`*. Your report should contain answers to all the numbered parts of the problem below. To guide you, the approximate number of words that should be dedicated to each part of the question is shown like this,* ***[200]****.*

The code associated with the coursework should be written in Python and follow best software development practice as defined by the Research Computing module. This should include:

- *Writing clear readable code that is compliant with a common style guide and uses suitable build management tools.*
- *Providing appropriate documentation that is compatible with auto-documentation tools like Doxygen/Sphinx.*
- *Providing an easy to use* `main.py` *script that can be run from the command line and produces all the figures and results contained in your report.*
- *The project must be well structured (sensible folder structure, README.md, licence etc..) following standard best practice. The data file* `lighthouse_flash_data.txt` *should also be stored in your repository.*

Parts i) to v) of this problem are based on an old Cambridge problem sheet question by S. Gull and is discussed in the text book by D. S. Sivia, "Data Analysis: A Bayesian Tutorial".

A lighthouse is at position α along a straight coastline and a distance β out to sea. The lighthouse rotates and emits flashes at uniformly-distributed random angles θ; the light beams are narrow and (if $-\pi/2 < \theta < \pi/2$) intersect the coastline at a single point. An array of detectors spread along the coastline record the locations x_k (where $k = 1, 2, \ldots, N$) where N flashes are received; the detectors only record that a flash has occurred, not the direction from which it was received. Your task is to find the location of the lighthouse. The setup is illustrated in Fig. 1.

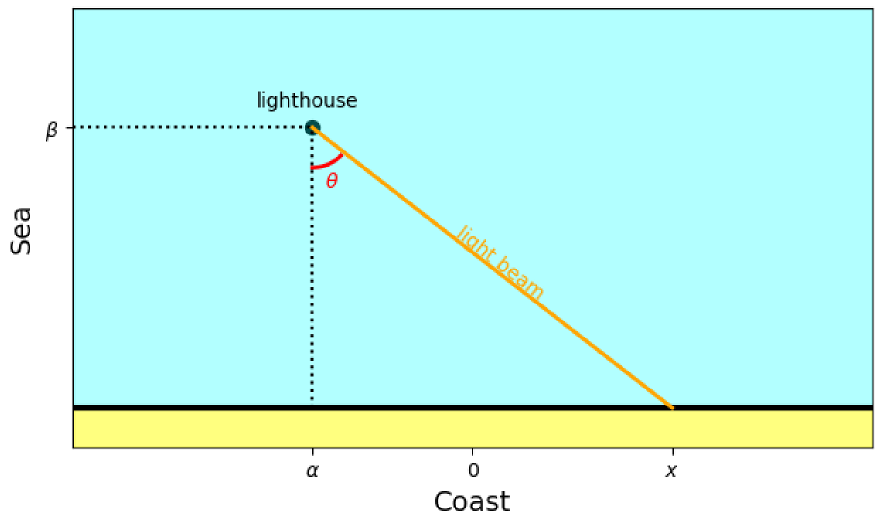

Figure 1: Diagram of the setup of the lighthouse problem.

i) Using the geometry of the problem, find a trigonometric relationship between α, β, θ and x. **[100]**

ii) Using the fact that θ is uniformly distributed, show by a suitable transformation of the PDF that the likelihood for the location of a single flash x is given by **[250]**

$$\mathcal{L}_x(x|\alpha, \beta) = \frac{\beta}{\pi\left(\beta^2 + (x-\alpha)^2\right)}.$$

This is known as a *Cauchy*, or *Lorentzian* distribution. Because all the flashes are emitted at independent angles, the likelihood function for a set of flash locations is given by the product $\mathcal{L}_x(\{x_k\}|\alpha, \beta) = \prod_k \mathcal{L}_x(x_k|\alpha, \beta)$.

iii) A frequentist colleague points out that the most likely location for any flash to be received is $\hat{x} = \alpha$. Based on this, your colleague suggests using the sample mean $(1/N)\sum_k x_k$ as an estimator for the parameter α. Show that while your colleague is correct about the most likely location for a flash, discuss why this is NOT a good estimator for α. *[Hint. You might compare the sample mean with the maximum likelihood estimator for α, which is a good estimator.]* **[600]**

iv) Choose suitable priors for the unknown parameters α and β. Justify your choice. **[50]**

v) $N = 20$ flash locations are stored in the first column of the file `lighthouse_flash_data.txt`. Using this data, draw stochastic samples from the posterior distribution $P(\alpha, \beta | \{x_k\})$. You can use any algorithm and software implementation you choose. Your report should include a brief description of the algorithm you used along with references to any software packages and papers describing the algorithm and implementation. Your report should include (i) a 2-dimensional histogram showing the joint posterior on α and β, (ii) 1-dimensional histograms of the marginalised posteriors on both parameters, (iii) measurements of both parameters quoted in the form mean $\pm$ standard deviation, and (iv) suitable convergence diagnostic information for your chosen sampling algorithm. **[1000]**

You are now told that the array of detectors also measures the intensity of each flash, with some known measurement uncertainty. The measured intensities I_k, for $k = 1, 2, \ldots, N$, are stored in the second column of the data file. Intensity measurements are independent and follow log-normal distributions with an uncertainty $\sigma = 1$. Flashes viewed from further away appear (on average) less intense according to an inverse square law; the expectation of the log-intensity is $\mu = \log(I_0/d^2)$ where I_0 is the absolute intensity of the lighthouse and $d^2 = \beta^2 + (x - \alpha)^2$ is the lighthouse-detector distance. The likelihood for each intensity measurement is

$$\mathcal{L}_I(\log I | \alpha, \beta, I_0) = \frac{\exp\left(\frac{-(\log I - \mu)^2}{2\sigma^2}\right)}{\sqrt{2\pi\sigma^2}},$$

The location and intensity measurements are independent, so the likelihood for a combined location and intensity measurement is

$$\mathcal{L}_x(x|\alpha, \beta)\mathcal{L}_I(\log I | \alpha, \beta, I_0).$$

vi) Choose a suitable prior for the unknown parameter I_0. Justify your choice. **[50]**

vii) Repeat part v) but this time including both the flash location and intensity measurements and exploring the 3-dimensional posterior $P(\alpha, \beta, I_0 | \{x_k\}, \{I_k\})$. Obtain a new measurement for α in the form mean $\pm$ standard deviation. **[800]**

viii) Compare your measurements for α from parts v) and vii). Hopefully including the intensity data has lead to an improved measurement of α. **[150]**