

S2 Statistics for Data Science

CRSiD: tmb76

University of Cambridge

Contents

The Lighthouse Problem

The Setup

A lighthouse that is at a distance β from the coast is at position α along it. The lighthouse emits flashes at random angles θ , following a uniform distribution. The flashes can be considered narrow and, provided $\pi/2 < \theta < 3\pi/2$, intersect the coastline at a single point. Detectors on the coastline record only the flashes' locations x_k (where $k = 1, 2, \dots, N$) for N flashes received. The setup is illustrated in Figure 1.

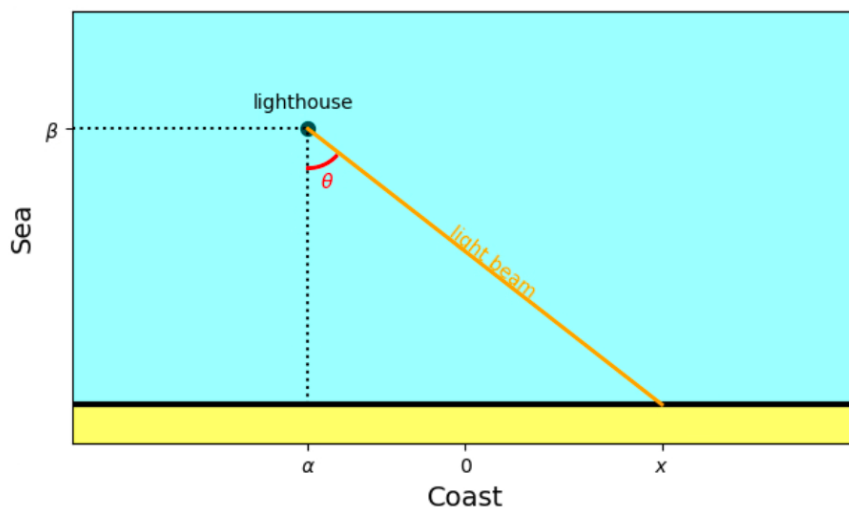


Figure 1: The lighthouse problem setup

The goal of this project is to find the location of the lighthouse from the recorded data x_k .

(i) The Geometry of the Problem

As stated above, only the distribution of the flashes' angle, θ , is known. However, the observed data obtained is the location of the flashes on the coastline, x_k . Thus,

it is worth considering what the relationship is between the angle of the flash and the location of the flash on the coastline, expressed in terms of the unknown lighthouse location parameters α and β . Using trigonometry, the following relationship can be derived:

$$x = \alpha + \beta \tan(\theta) \iff \theta = \tan^{-1} \left(\frac{x - \alpha}{\beta} \right) \quad (1)$$

(ii) The Likelihood Function

In addition to the relationship above, the probability distribution of the flashes' angles is known to be uniform. Thus, the probability of a single flash to have the angle θ is given by:

$$\theta \sim U(-\pi/2, \pi/2) \quad (2)$$

$$P(\theta) = \mathbb{1}_{(-\pi/2, \pi/2)}(\theta) \times \frac{1}{\pi} = \begin{cases} \frac{1}{\pi}, & \text{if } -\frac{\pi}{2} < \theta < \frac{\pi}{2} \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

Now, using the trigonometric relationship between θ and x derived in (1), a change of variable is conducted to find the probability distribution of the flashes' locations on the coastline, x . Here, this will be considered as the likelihood of a single flash to be observed at location x given the lighthouse location parameters α and β . For this 1-D case, the change of variable can be written as:

$$\mathcal{L}_x(x|\alpha, \beta) = \mathcal{L}_\theta(\theta) \times \left| \frac{d\theta}{dx} \right| \quad (4)$$

Where $\mathcal{L}_\theta(\theta)$ is the probability distribution of the flashe's angles, which is given in Eq (3) as a uniform distribution. Since the angle θ must be such that $-\pi/2 < \theta < \pi/2$ for the flash to be observed on the coastline, $\mathcal{L}_\theta(\theta)$ can be set to $1/\pi$. The derivative term is given by:

$$\frac{d\theta}{dx} = \frac{d}{dx} \left(\tan^{-1} \left(\frac{x - \alpha}{\beta} \right) \right) \quad (5)$$

Using the chain rule, the derivative can be found to be:

$$\frac{d\theta}{dx} = \frac{1}{1 + \left(\frac{x - \alpha}{\beta} \right)^2} \times \frac{1}{\beta} \quad (6)$$

With some algebraic manipulation, this can be re-arranged as:

$$\frac{d\theta}{dx} = \frac{\beta}{\beta^2 + (x - \alpha)^2} \quad (7)$$

Thus the likelihood for a single flash to be observed at location x given the lighthouse location parameters α and β is:

$$\mathcal{L}_x(x|\alpha, \beta) = \frac{1}{\pi} \times \frac{\beta}{\beta^2 + (x - \alpha)^2} \quad (8)$$

as required. Below is an example plot of the likelihood function for hypothetical values of $\alpha = 0$ and $\beta = 1$.

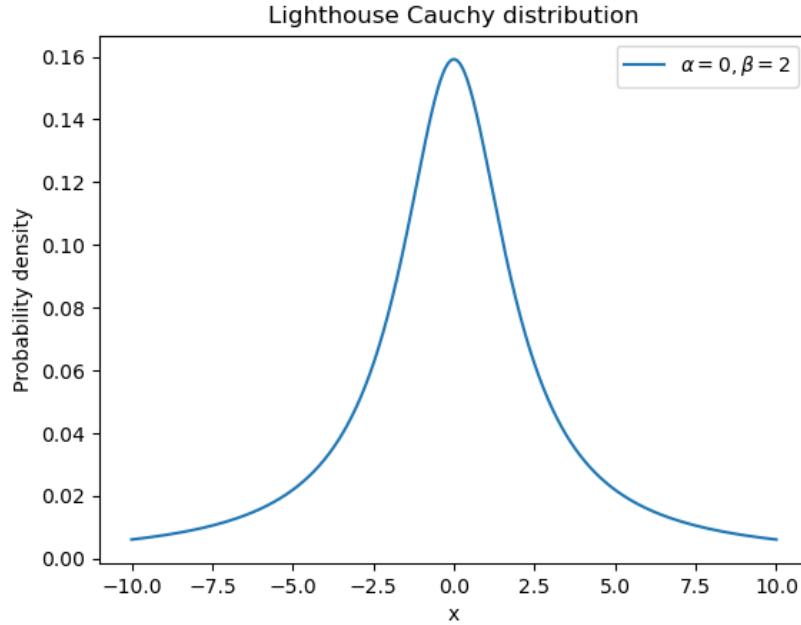


Figure 2: Cauchy distribution plot, for hypothetical values of $\alpha = 0$ and $\beta = 1$

Fig. 2 shows the Cauchy distribution, which is a heavy-tailed distribution. This means that the likelihood stays non-negligible for values of x very different from α . This is a result of the trigonometric relationship between θ and x in Eq. (1). When the angle θ is close to $\pm\pi/2$, small changes in θ can result in large changes in x . And since the angle θ is uniformly distributed, the likelihood of observing a flash at a location x far from α is non-negligible.

(iii) Frequentist Claim

A frequentist colleague claims the most likely location, \hat{x} , for any flash to be received is given by the parameter α , the location of the lighthouse along the coastline. They also suggest using the sample mean to estimate the α parameter. First, the $\hat{x} = \alpha$ claim. \hat{x} is the location for which the likelihood in Eq. (8) is maximised. This is found by taking the derivative of the likelihood with respect to x and setting it to zero. This gives:

$$\left. \frac{d\mathcal{L}_x(x|\alpha, \beta)}{dx} \right|_{x=\hat{x}} = 0 \iff \left. \frac{d}{dx} \left(\frac{\beta}{\beta^2 + (x - \alpha)^2} \right) \right|_{x=\hat{x}} = 0 \quad (9)$$

$$-\frac{1}{\pi}\beta(2\hat{x} - 2\alpha) \frac{1}{\beta^4 + 2\beta^2(\hat{x} - \alpha)^2 + (\hat{x} - \alpha)^4} = 0 \quad (10)$$

The $-\frac{1}{\pi}\beta$ terms can be dropped, leaving only the numerator $(2\hat{x} - 2\alpha)$ term to satisfy the equation. This is satisfied when $\hat{x} = \alpha$, as claimed. Thus, the frequentist's claim is accurate for the value of the most likely flash location. However, estimating the value of α with the sample mean is not robust. To show this, the Maximum Likelihood Estimate (MLE) method is used to compare the two estimators. The aim is to find the value of α that maximises the likelihood derived in Eq. (8), based on the observed sample $\{x_k\}$ ($k = 1, 2, \dots, N$). This means now having the likelihood being the product of individual likelihoods for all values of $\{x_k\}$, as the flashes are emitted at independent angles. Finding $\hat{\alpha}$ is done in a similar way as above, by taking the derivative of the likelihood with respect to α and setting it to zero. However, for simplicity in derivation and computation, the log-likelihood is used instead.

$$\hat{\alpha}_{MLE} = \arg \max_{\alpha} \ln \left(\prod_{k=1}^N \mathcal{L}_x(\{x_k\}|\alpha, \beta) \right) = \arg \max_{\alpha} \sum_{k=1}^N \ln \left(\frac{1}{\pi} \times \frac{\beta}{\beta^2 + (x_k - \alpha)^2} \right) \quad (11)$$

$$\hat{\alpha}_{MLE} = \arg \max_{\alpha} \sum_{k=1}^N \ln \left(\frac{\beta}{\pi} \right) + \ln \left(\frac{1}{(\beta^2 + (x_k - \alpha)^2)} \right) \quad (12)$$

$$\left. \frac{d}{d\alpha} \left(\sum_{k=1}^N \ln \left(\frac{\beta}{\pi} \right) - \ln(\beta^2 + (x_k - \alpha)^2) \right) \right|_{\alpha=\hat{\alpha}_{MLE}} = 0 \quad (13)$$

$$\sum_{k=1}^N \frac{(2\hat{\alpha}_{MLE} - 2x_k)}{\beta^2 + (x_k - \hat{\alpha}_{MLE})^2} = 0 \quad (14)$$

To get an estimate of $\hat{\alpha}_{MLE}$, the above equation needs to be rearranged to isolate α .

However, this is difficult to do analytically. The sample mean estimator can be tested numerically. Using Pedro Pessoa’s example code for the lighthouse problem [3], flash location sample data is generated for $\alpha = 2$ and $\beta = 2$. The sample mean and confidence intervals are then calculated and compared to the true mean. The results are shown in Fig. 3 and are similar to the example shown in Fig 2.9 of Section 2.4 in Silvia’s book [4].

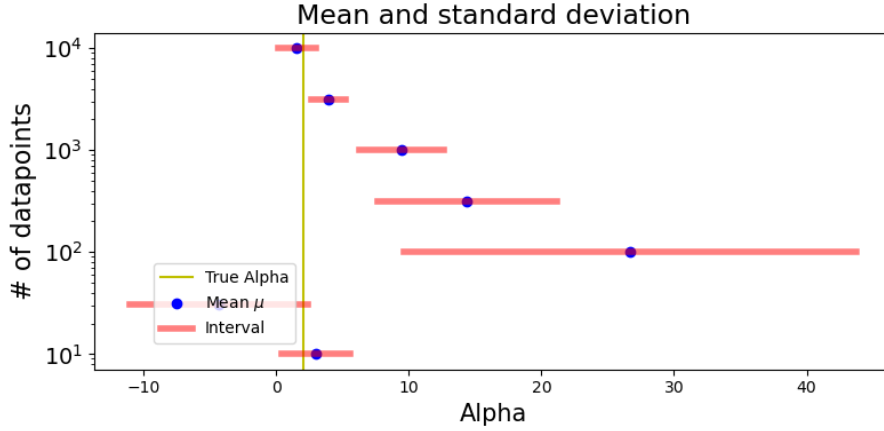


Figure 3: Plot of sample mean and confidence intervals for the lighthouse problem, for varying sample sizes, with true parameters set to $\alpha = 2$ and $\beta = 2$

Fig. 3 shows that the sample mean estimator is not robust as it does not converge with the sample size increasing, sometimes even getting worse. This is due to the heavy-tailed nature of the Cauchy distribution. And in the context of the lighthouse problem, this was discussed shortly at the end of section (ii). The overarching issue is that the Cauchy distribution does not follow the Central Limit Theorem (CLT) [4, p34]. This is because the mean, or expected value of the Cauchy distribution is undefined [1, p84]. This is shown in James’ book [1, p34] where the characteristic function of the Cauchy distribution has no derivatives at $t=0$, which from the definition of the central moment, means it has no defined mean:

$$\phi(t) = \int_{-\infty}^{\infty} e^{itx} \frac{1}{\pi} \frac{\beta}{\beta^2 + (X - \alpha)^2} dX \quad (15)$$

(iv) Priors for α and β

Since priors are primarily to express how much we know about the parameters before observing the data. For the lighthouse problem, it is simply known that the lighthouse is at some location α along the coastline and at some distance β from the coastline,

and no more. Thus, the priors for α and β should be ignorant ones, such as the uniform distribution. These can be written as:

$$\alpha \sim U(\alpha_{min}, \alpha_{max}) \text{ and } \beta \sim U(0, \beta_{max}) \quad (16)$$

$$P(\alpha) = \mathbb{1}_{(\alpha_{min}, \alpha_{max})}(\alpha) \times \frac{1}{\alpha_{max} - \alpha_{min}} \text{ and } P(\beta) = \mathbb{1}_{(0, \beta_{max})}(\beta) \times \frac{1}{\beta_{max}} \quad (17)$$

where $\mathbb{1}$ is the indicator function. β is non-negative, hence the lower bound of the uniform distribution for β is 0.

(v) Stochastic Sampling

Defining the Posterior

With the likelihood and priors defined, and the general lighthouse problem understood, estimating the location parameters α and β can be attempted. Here we use stochastic sampling approaches, directly drawing samples from the posterior distribution of the parameters. The posterior distribution is given by Bayes' theorem as:

$$P(\alpha, \beta|x) = \frac{\mathcal{L}_x(x|\alpha, \beta) \times \pi(\alpha, \beta)}{Z} \quad (18)$$

where $\pi(\alpha, \beta)$ is the joint prior distribution of α & β , and Z is the evidence. The evidence is the normalising constant, which is the integral of the likelihood function over the entire parameter space. Since it is a constant, evaluating it is not necessary for the purposes of sampling values of α and β from the posterior. Furthermore, the joint prior distribution $\pi(\alpha, \beta)$ is simply the product of the individual priors for α and β , as they are independent. Finally, the estimation of the location parameters α and β is based on $N = 20$ observed flash locations. These observations are plotted in Fig. 4's diagram.

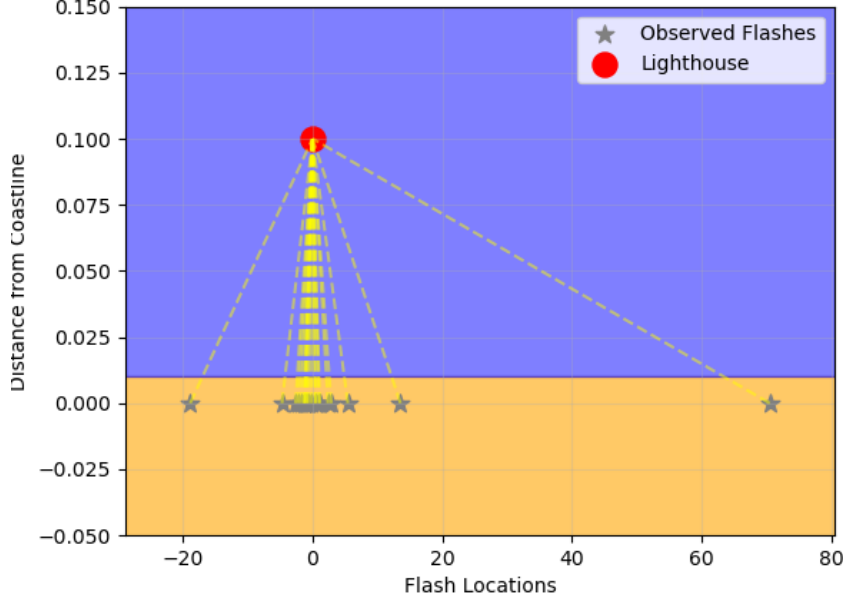


Figure 4: The observed flash locations on the coastline, with a hypothetical lighthouse location at $\alpha = 0$ and $\beta = 0.1$

With these data points, the likelihood in Eq. (18) is the product of the likelihood for each observation as in Eq. (11). This means we can rewrite the posterior distribution, P as:

$$P(\alpha, \beta | \{x_k\}) \propto \mathcal{L}_x(\{x_k\} | \alpha, \beta) \times \pi(\alpha) \times \pi(\beta) \quad (19)$$

Giving the following expression:

$$P(\alpha, \beta | \{x_k\}) \propto \left(\frac{\beta}{\pi}\right)^N \times \prod_{k=1}^N \frac{1}{(\beta^2 + (x_k - \alpha)^2)} \times \frac{1}{\beta_{max}(\alpha_{max} - \alpha_{min})} \quad (20)$$

Finally, for numerical precision reasons, to avoid product of small numbers, the log posterior is used instead:

$$\ln(P(\alpha, \beta | \{x_k\})) \propto N \ln(\beta) - N \ln(\pi) - \sum_{k=1}^N \ln(\beta^2 + (x_k - \alpha)^2) - \ln(\beta_{max}) - \ln(\alpha_{max} - \alpha_{min}) \quad (21)$$

Since the natural logarithm is a one-to-one mapping and is invertible, this will not affect the final result of the sampling, i.e. estimates of α and β .

Sampling Methods

Here, multiple Monte-Carlo Markov Chains (MCMC) sampling algorithms were used. The first is the Metropolis-Hastings (MH) algorithm described below. One of the good things of the MH algorithm is that it does not require us to know the normalised target distribution.

Metropolis-Hastings Algorithm

```

1:  $l_0 \sim (\alpha, \beta)$  ▷ Initialisation
2:  $i \leftarrow 0$ 
3: while  $i \geq 0$  do ▷ Iterate  $i = 0, 1, \dots, N$ 
4:    $y \sim \mathcal{N}(\alpha, \beta | l_i, cov)$  ▷ Proposal distribution, here, a multivariate normal
5:    $a \leftarrow (P(y | \{x_k\}) \mathcal{N}(l_i | y, cov)) / (P(l_i | \{x_k\}) \mathcal{N}(y | l_i, cov))$  ▷ MH acceptance probability
6:    $u \sim U(0, 1)$  ▷ Uniform random number
7:   if  $u < a$  then
8:      $l_{i+1} \leftarrow y$  ▷ Accept the proposal
9:   else
10:     $l_{i+1} \leftarrow x_i$  ▷ Reject the proposal
11:   end if
12:    $i \leftarrow i + 1$ 
13: end while

```

Where cov is the covariance matrix of the multivariate normal distribution, $P(l_i | \{x - k\})$ is the posterior distribution, and $l, y = (\alpha, \beta)$ are proposed points in the parameter space.

The key step of MH, line 7 of the algorithm, is the check of the acceptance probability, a . This is the ratio of the posterior distribution at the proposed point, y , to the posterior distribution at the current point, l_i , multiplied by the ratio of the proposal distribution at the current point, l_i , to the proposal distribution at the proposed point, y . The proposal distribution is a multivariate normal distribution, with mean l_i and covariance matrix cov , which is to be chosen. The acceptance probability is then compared to a uniformly drawn number between 0 and 1. If the acceptance probability is greater than u , the proposed point is accepted, and the chain moves to the proposed point. Otherwise, the proposed point is rejected, and the chain stays at the current point. The chain then iterates, moving through the parameter space, until a sufficient number of samples are obtained. One downside to MH is that it can be slow to converge, and line 10 of the algorithm makes it so MH

can stay at the same point for multiple iterations, resulting in repeated samples.

This is why the second algorithm used is the Hamiltonian Monte Carlo (HMC) algorithm. HMC is a more sophisticated algorithm that uses the gradient of the log posterior to guide the sampling. This means it can take larger steps through the parameter space without the "blindness" of MH as it will take these large steps in the direction of the steepest gradient. This is done by introducing a momentum variable, which is then used to guide the sampling. The HMC algorithm is described below.

Hamiltonian Monte Carlo Algorithm

```

1:  $l_0 \sim (\alpha, \beta)$  ▷ Initialisation
2:  $i \leftarrow 0$ 
3: while  $i \geq 0$  do ▷ Iterate  $i = 0, 1, \dots, N$ 
4:    $p \sim \mathcal{N}(0, 1)$  ▷ Sample momentum
5:    $q \leftarrow l_i$  ▷ Initial position
6:    $p \leftarrow p - \epsilon \nabla \ln(P(q|\{x_k\}))$  ▷ Leapfrog integration
7:    $q \leftarrow q + \epsilon p$  ▷ Leapfrog integration
8:    $p \leftarrow p - \epsilon \nabla \ln(P(q|\{x_k\}))$  ▷ Leapfrog integration
9:    $p \leftarrow -p$  ▷ Reverse momentum
10:   $H \leftarrow \ln(P(q|\{x_k\})) + \frac{p^2}{2}$  ▷ Hamiltonian
11:   $q_{new} \sim \mathcal{N}(q, cov)$  ▷ Proposal distribution
12:   $p_{new} \leftarrow p_{new} - \epsilon \nabla \ln(P(q_{new}|\{x_k\}))$  ▷ Leapfrog integration
13:   $p_{new} \leftarrow -p_{new}$  ▷ Reverse momentum
14:   $H_{new} \leftarrow \ln(P(q_{new}|\{x_k\})) + \frac{p_{new}^2}{2}$  ▷ Hamiltonian
15:   $a \leftarrow \min(1, e^{H-H_{new}})$  ▷ HMC acceptance probability
16:   $u \sim U(0, 1)$  ▷ Uniform random number
17:  if  $u < a$  then
18:     $l_{i+1} \leftarrow q_{new}$  ▷ Accept the proposal
19:  else
20:     $l_{i+1} \leftarrow x_i$  ▷ Reject the proposal
21:  end if
22:   $i \leftarrow i + 1$ 
23: end while

```

Where cov is the covariance matrix of the multivariate normal distribution, $P(l_i|\{x - k\})$ is the posterior distribution, and $l, y = (\alpha, \beta)$ are proposed points in the parameter space.

Third, the slice sampling algorithm is used. This algorithm is also a MCMC algorithm and follows a relatively simple procedure. The main idea is to sample

uniformly from the region under the curve of the posterior distribution. One of the main advantages of slice sampling is that it's acceptance rate is always 1, meaning that it is very efficient. Again, the slice sampling algorithm is described below.

Slice Sampling Algorithm [2]

```

1: Initialize:
2:    $l_0 \leftarrow$  Initial state
3:    $i \leftarrow 0$ 
4: while  $i \geq 0$  do                                      $\triangleright$  Iterate until desired number of samples
5:    $y_0 \sim U(0, P(l_i|\{x_k\}))$ 
6:   Stepping Out Procedure: To find  $I = (L, R)$  that contains all or at least part of  $S$ , expand  $I$  by a
     step of size  $\mu$  until both ends of  $I$ :  $L$  and  $R$  are outside of  $S$ 
7:    $l_{i+1} \sim U(I \cap S)$                                 $\triangleright$  Shrinking Procedure: Sample  $l_{i+1}$  from  $I$  until one of them is inside  $S$ 
8:    $i \leftarrow i + 1$ 
9: end while

```

Where S is the slice for which $\{l : y_0 < (l_i|\{x_k\})\}$, and $P(l_i|\{x_k\})$ is the posterior distribution.

The issue that arises with simple slice sampling here is that the stepping out procedure has no guarantee of finding the entire slice S , especially in case of multi-modal distributions. This is where the software package used, Zeus, comes in. Zeus is a Python package that implements a more sophisticated version of slice sampling, called Ensemble Slice Sampler. As its name suggests, it uses an ensemble of walkers to sample from the posterior distribution, each with a different starting point. This minimises the chances of the total chain sampled missing an entire mode of the posterior distribution [2].

Results

Metropolis-Hastings

(vi) Prior for the Intensity I_0 Parameter

Again, the prior describing the knowledge we have about the intensity I_0 . Because the intensity follows a log-normal distribution due to the inverse square law, the prior for I_0 should be a log-uniform distribution. MAYBE UNIFORM? Cuz it's just a constant set for the lighthouse

$$I_0 \sim U(I_{0_{min}}, I_{0_{max}}) \quad (22)$$

(vii) Stochastic Sampling, with the added Intensity I_0 Parameter

(viii) Comparison of the results

Bibliography

- [1] Frederick James. Statistical methods in experimental physics: 2nd edition. 11 2006.
- [2] Minas Karamanis, Florian Beutler, and John A. Peacock. zeus: A python implementation of ensemble slice sampling for efficient bayesian parameter inference. *Monthly Notices of the Royal Astronomical Society*, 2021.
- [3] PessoaP. Lighthouse problem.
- [4] Devinderjit Sivia and John Skilling. *Data Analysis: A Bayesian Tutorial*. Oxford University Press, Incorporated, 2006.