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 $\pi/2 < \theta < \pi/2$, intersect the coastline at a single point. Detectors on the coastline record only the flashes' locations x_k (where $k = 1, 2, \ldots, N$) for N flashes received. The setup is illustrated in Figure 1.

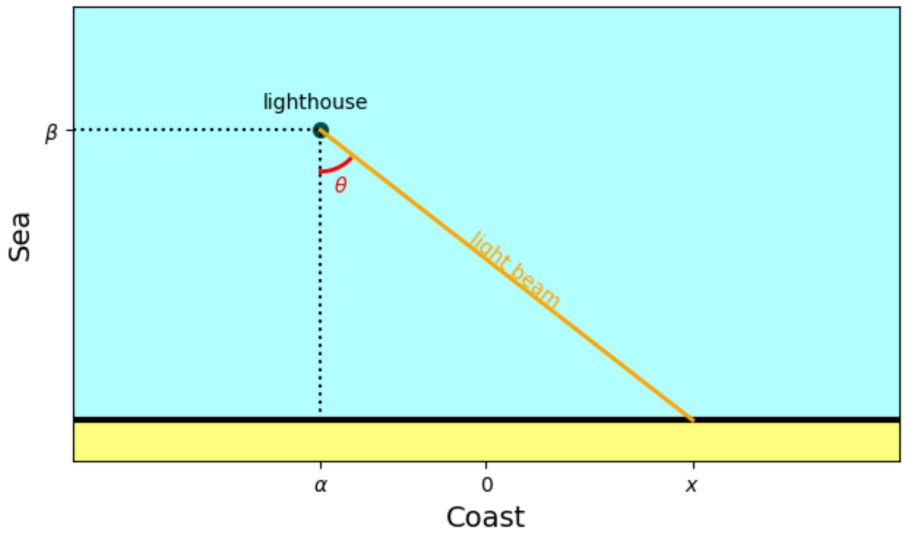

Figure 1: The lighthouse problem setup

The goal of this project is to find the location of the lighthouse from the observed detector data $\{x_k\}$, and $\{I_k\}$.

(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 lightouse location parameters α and β. Using trigonometry, the following relationship can be derived:

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

(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:

$$\theta \sim U(-\pi/2, \pi/2) \tag{2}$$

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

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:

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

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:

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

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

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

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

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

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

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

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

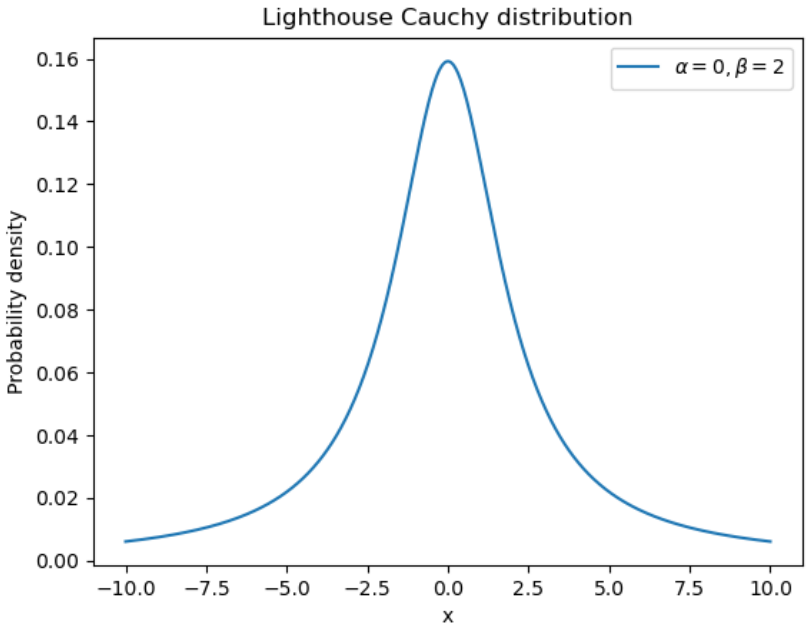

Figure 2: Cauchy distribution plot, for hypothetical values of $\alpha = 0$ and $\beta = 1$

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:

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

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

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 $\{x_k\}$ $(k = 1, 2, \ldots, N)$. 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.

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

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

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

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

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].

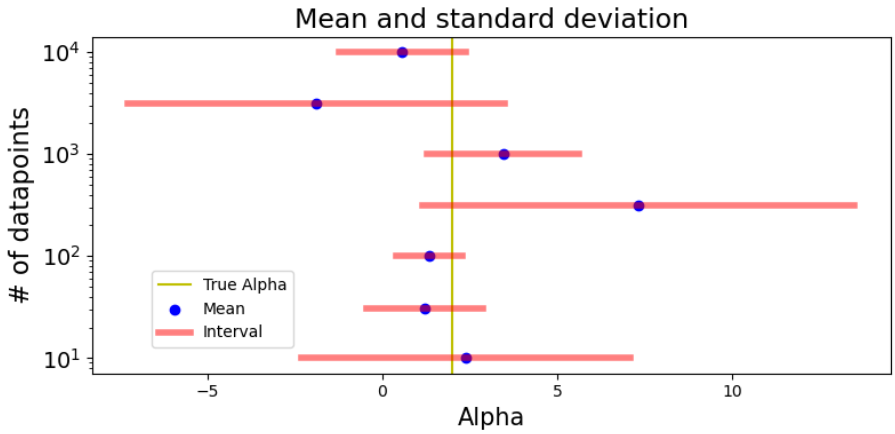

Figure 3: Plot of smaple mean and confidence intervals for the lighthouse problem, for varying sample sizes, with true parameters set to $\alpha = 2$ and $\beta = 2$

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:

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

(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 know 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:

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

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

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:

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

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.

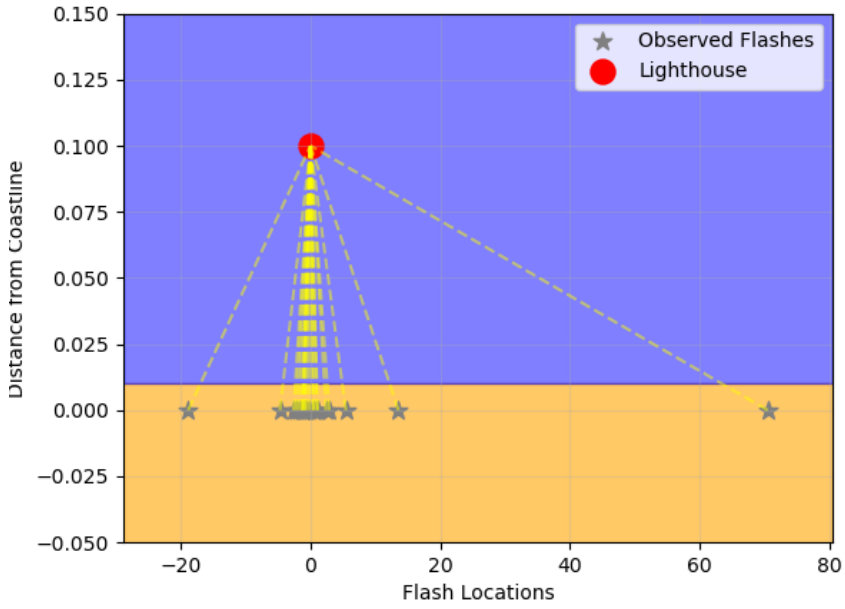

Figure 4: The observed flash locations on the coastline, with a hypothetical lighthouse location at $\alpha = 0$ and $\beta = 0.1$

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:

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

Giving the following expression:

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

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

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

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 Markov Chains Monte-Carlo (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 $i = 0, 1, \ldots, N$
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.

Second, 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]

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

The total acceptance fraction for the chain was 0.27555, which is reasonable for the MH algorithm. In order to reflect the a priori knowledge we had of the position of the lighthouse, the initial position of the chain was set randomly, following a uniform distribution in between the set limits. This, however, meant the chain would have to converge first to the region of high probability, which requires us to ignore the first few samples of the chain. To check how many samples it took for the chain to converge, a plot of the first 1000 samples was made.

Fig. 5 shows that the chain converges in just about 1000 steps. Thus, before estimating the parameters from the chain, this burn-in period of 1000 steps must be

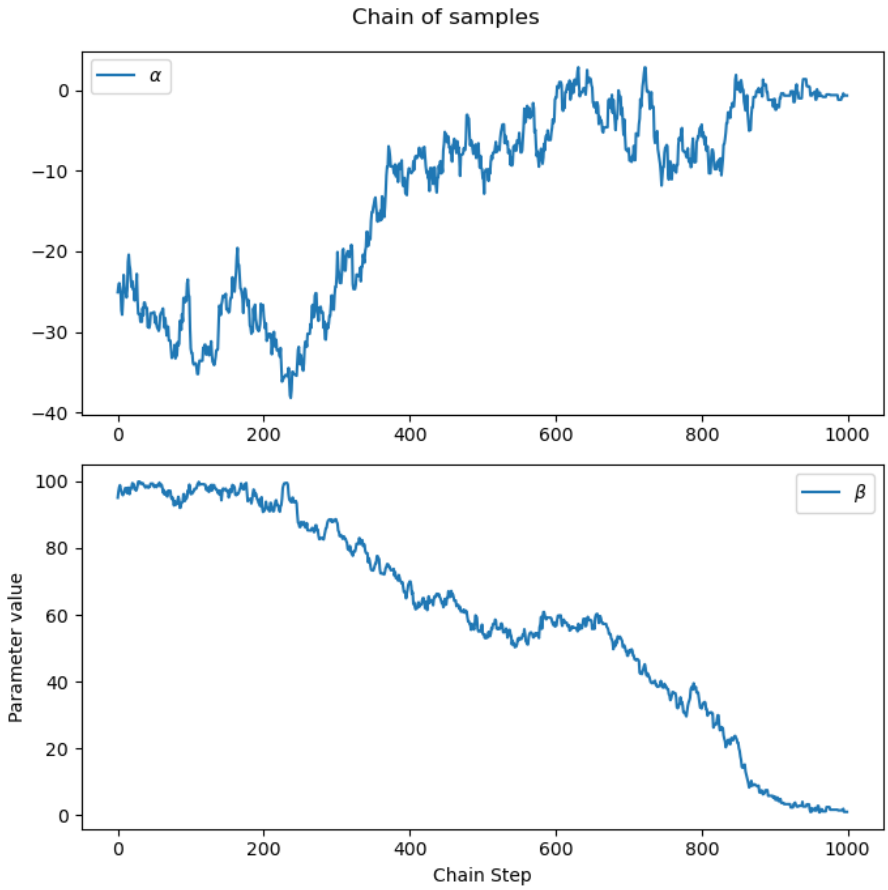

Figure 5: First 1000 samples of the MH chain for the lighthouse problem

ignored and dropped. Furthermore, once this is done, due to the nature of MCMC chains, subsequent samples are correlated. And so, to obtain independent samples, the chain must be thinned, by only taking every n-th sample. The value of n is determined by seeing how long it takes the chain to oscillate from one end of the distribution to the other. This is called the autocorrelation time. For MH, it was found to be 10.9 steps. Thus, the chain was thinned by a factor of 11. The following samples were kept:

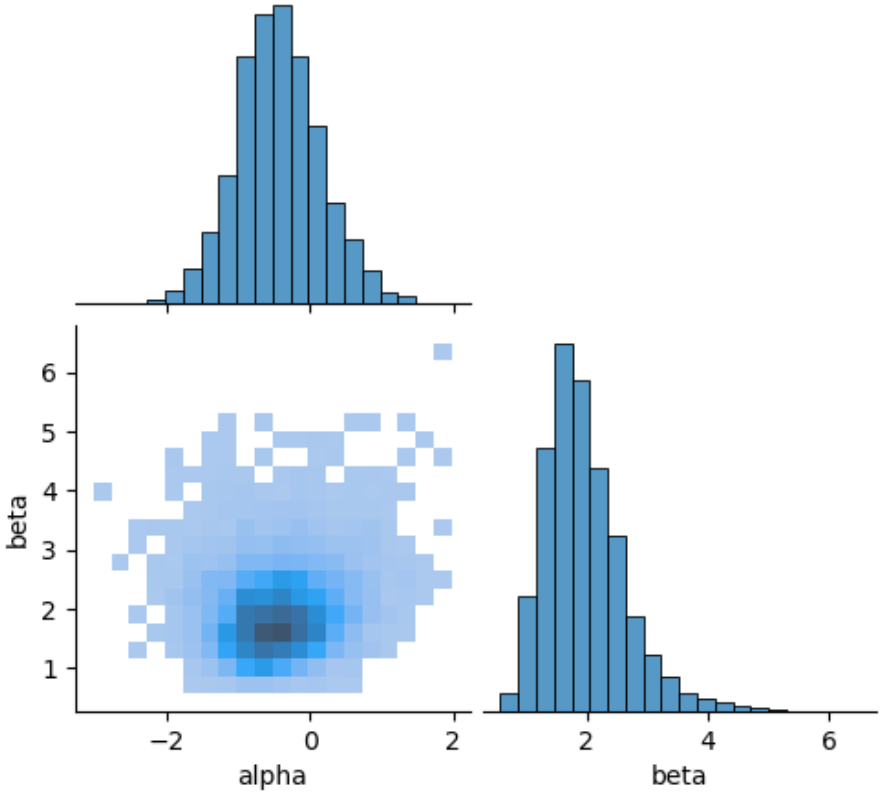

Figure 6: Metropolis-Hastings samples for the lighthouse problem,

Fig. 6 shows a corner plot of the samples generated using the MH algorithm. Taking the sample mean and standard deviation of the samples, the following estimates for α and β were obtained:

$$\hat{\alpha}_{MH} = -0.4428 \pm 0.5977 \text{ and } \hat{\beta}_{MH} = 1.9595 \pm 0.6669 \tag{22}$$

`zeus` Ensemble Slice Sampler

Using `zeus`, the acceptance rate is always 1, as mentioned earlier. Similarly, starting locations were randomly sampled from the prior distributions. 10,000 samples were generated, for 10 walkers. Looking at a plot of the chain in Fig. 7, it is clear that the chain converges much quicker than MH, and staying conservative, the burn-in period was set to be 100 steps. This can be attributed to the nature of slice sampling, which allows for much bigger steps to be taken as they will always be under the curve of the posterior distribution.

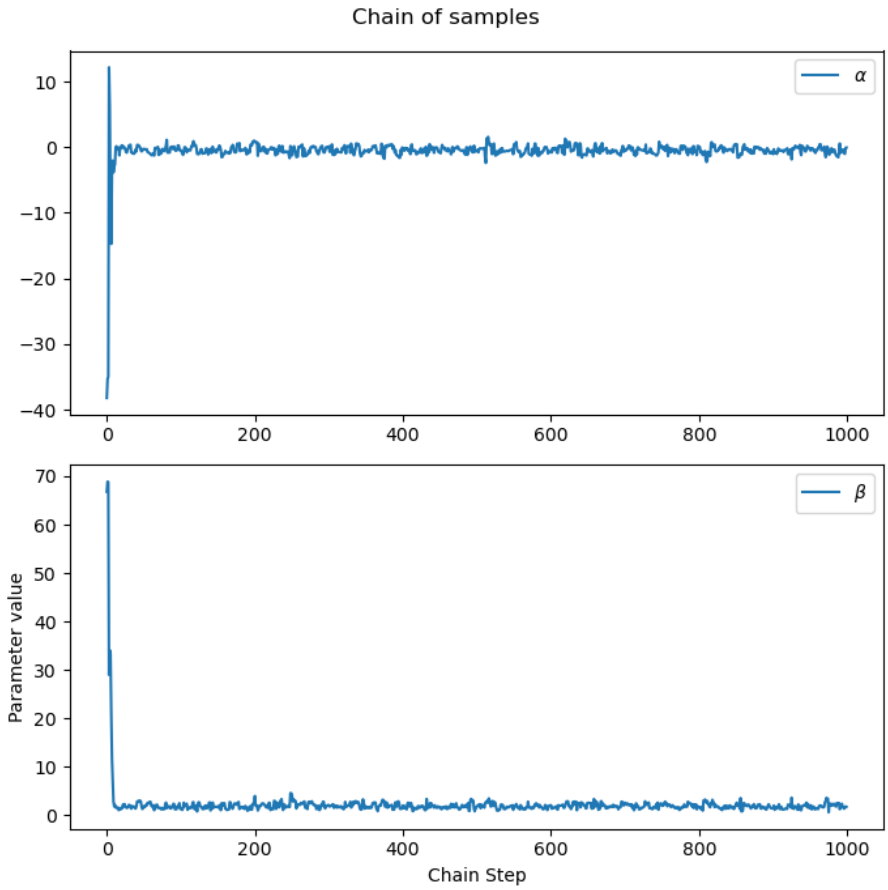

Figure 7: First 1000 samples of the SS chain for the lighthouse problem

This is also backed by the fact that the autocorrelation time for the chain was found to be 4.8 steps. Thus, the chain was thinned by a factor of 5. The following samples were kept:

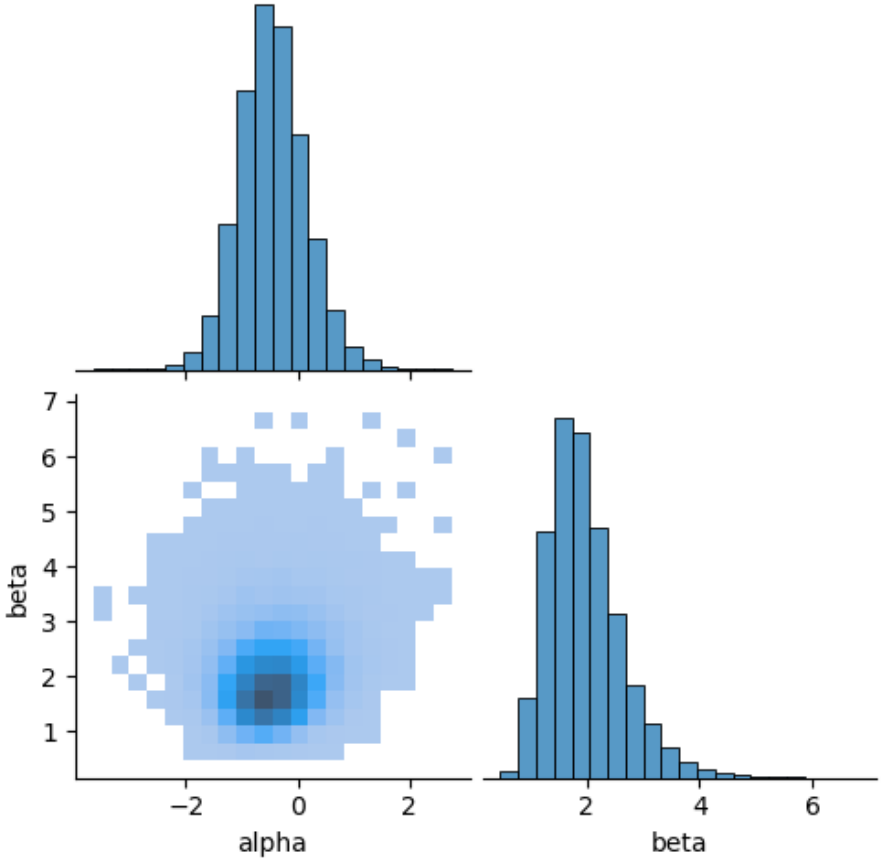

Figure 8: `zeus` Ensemble Slice Sampler samples for the lighthouse problem,

Giving the following estimates for α and β:

$$\hat{\alpha}_{SS} = -0.4558 \pm 0.6038 \text{ and } \hat{\beta}_{SS} = 1.9678 \pm 0.6757 \tag{23}$$

(vi) Prior for the Intensity I_0 Parameter

Again, the prior describing the knowledge we have about the intensity I_0. Because the intensity is a physical property of the lighthouse, it would be useful for it to be scale invariant. That is, the prior should be invariant under a change of units:

$$\pi(I_0)dI_0 = \pi(\gamma I_0)d(\gamma I_0) \Rightarrow \pi(I_0) \propto \frac{1}{I_0} \tag{24}$$

This is equivalent to having:

$$\log I_0 \sim U(I_{0_{max}}, I_{0_{max}}) \iff I_0 \sim \frac{1}{I_0 \log(\frac{I_{0max}}{I_{0_{min}}})} \tag{25}$$

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

Defining the Posterior

The likelihood for the measured intensity I is given by a log-normal distribution with uncertainty $\sigma = 1/$ The expectation μ given by the inverse square law, and therefore depending on x, α and β, such that $\mu = \log(I_0/((x-\alpha)^2 + \beta^2))$. The log-likelihood for the measured intensity I is then given by:

$$I \sim \mathcal{L}_I(\log I|\alpha, \beta, I_0) = \frac{exp(\frac{-(\log I - \mu)^2}{2\sigma^2})}{\sqrt{2\pi\sigma^2}} \tag{26}$$

Since the location and intensity measurements are independent, the joint likelihood for the location and intensity measurements is the product of the individual likelihoods. The log-likelihood for the location and intensity measurements is then given by:

$$\ln(\mathcal{L}_{x,I}(\{x_k\}, \{\log I_k\}|\alpha, \beta, I_0)) = \ln(\mathcal{L}_x(\{x_k\}|\alpha, \beta)) + \ln(\mathcal{L}_I(\{\log I_k\}|\alpha, \beta, I_0)) \tag{27}$$

For which the likelihood of the location is given in Eq.(8) and (11). Finally, with the joint prior also being the product of the individual priors, and the prior for intensity being a log-uniform distribution, the log prior is given by:

$$\ln(\pi(\alpha, \beta, I_0)) = \ln(\pi(\alpha)) + \ln(\pi(\beta)) + \ln(\pi(I_0)) \tag{28}$$

Thus the log posterior is given by:

$$\ln(P(\alpha, \beta, I_0|\{x_k\}, \{\log I_k\})) = \ln(\mathcal{L}_{x,I}(\{x_k\}, \{\log I_k\}|\alpha, \beta, I_0)) + \ln(\pi(\alpha, \beta, I_0)) \tag{29}$$

Sampling Methods

The same three MCMC algorithms were used to sample from the posterior distribution of the parameters. The parameter limits set are the same for α and β, and the prior for I_0 is set to be between a near-zero value (0.0001) and 50. The Metropolis-Hastings and `zeus` Ensemble Slice Sampler algorithms were used as before, but with

a larger number of samples generated (500,000) for the MH algorithm. The initial positions for the location parameters were obtained in the same way, with the intensity parameter being randomly sampled from the prior distribution as well.

Results

Metropolis-Hastings

Due to the higher dimension, the total acceptance fraction for this chain was much lower than in part (v). With just 7.65% of the total samples being accepted, this is still a reasonable amount though much worse than before. In case the burn-in was also worse in this context, the first steps of the chain were once again plotted to check when the chain converged.

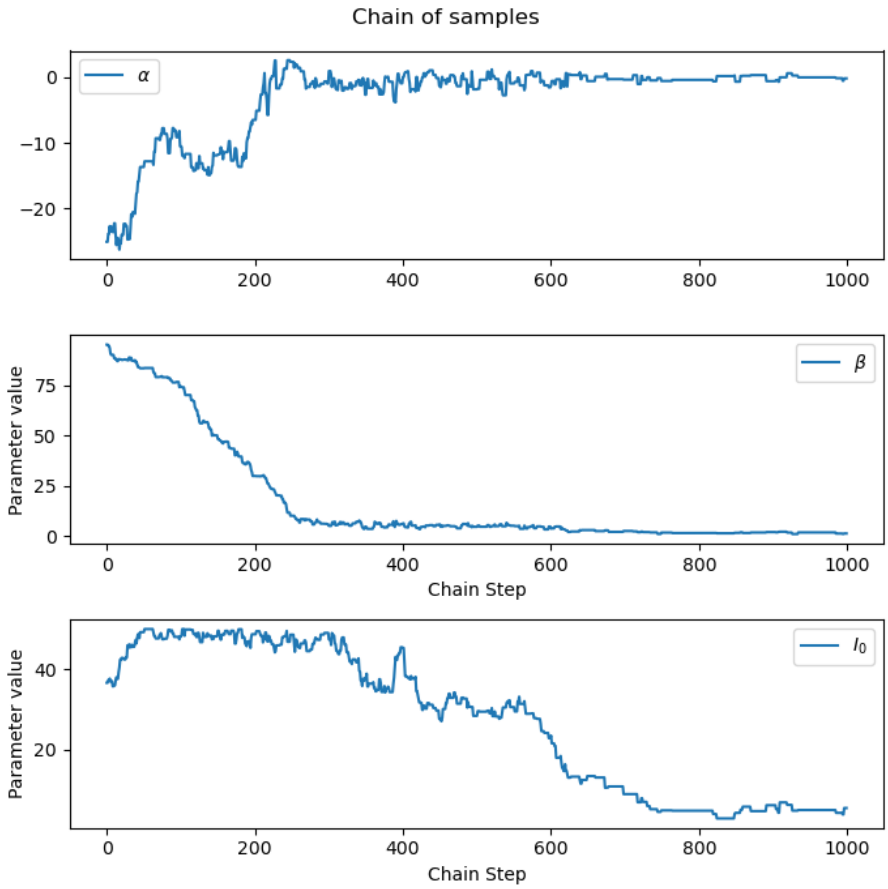

Figure 9: First 1000 samples of the MH chain for the lighthouse problem

Fig. 9 shows that the chhain again converges in about 1000 steps, maybe even

shorter. To stay conservative however, the burn-in period is set to 1000 steps and is dropped. However, the autocorrelation time is much larger for this case as it was found to be 94 steps. So the chain is thinned by a factor of 90+. This is why a larger number of samples were generated, in order to have a larger amount of independent samples to estimate the parameters with. With the samples left, the corner plot of the 2-D and 1-D marginal distributions of the posterior was obtained in Fig. 10.

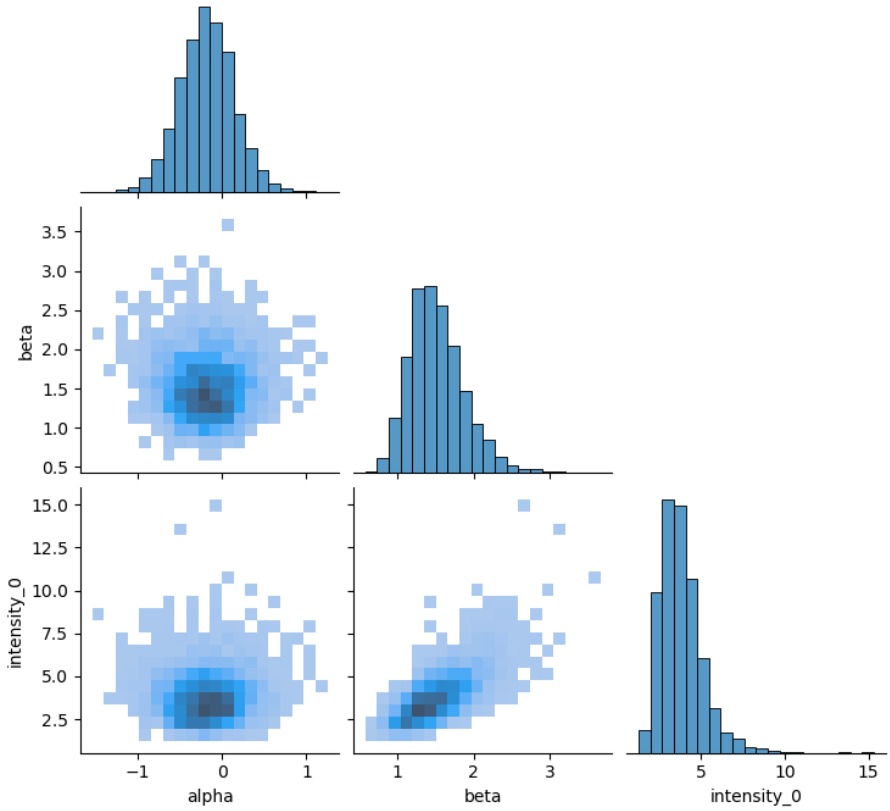

Figure 10: Metropolis-Hastings samples for the lighthouse problem,

Taking the sample mean and standard deviation of the samples, estimates of the lighthouse parameters were obtained:

$$\hat{\alpha}_{MH} = -0.1821 \pm 0.3311 \text{ and } \hat{\beta}_{MH} = 1.5193 \pm 0.3608 \text{ and } \hat{I}_0 = 3.7809 \pm 1.2880 \quad (30)$$

`zeus` Ensemble Slice Sampler

Again, using `zeus`, the acceptance rate is always 1. Looking at a plot of the chain in Fig. 11, it once again converges much quicker than MH. So the burn-in is kept at 100 steps.

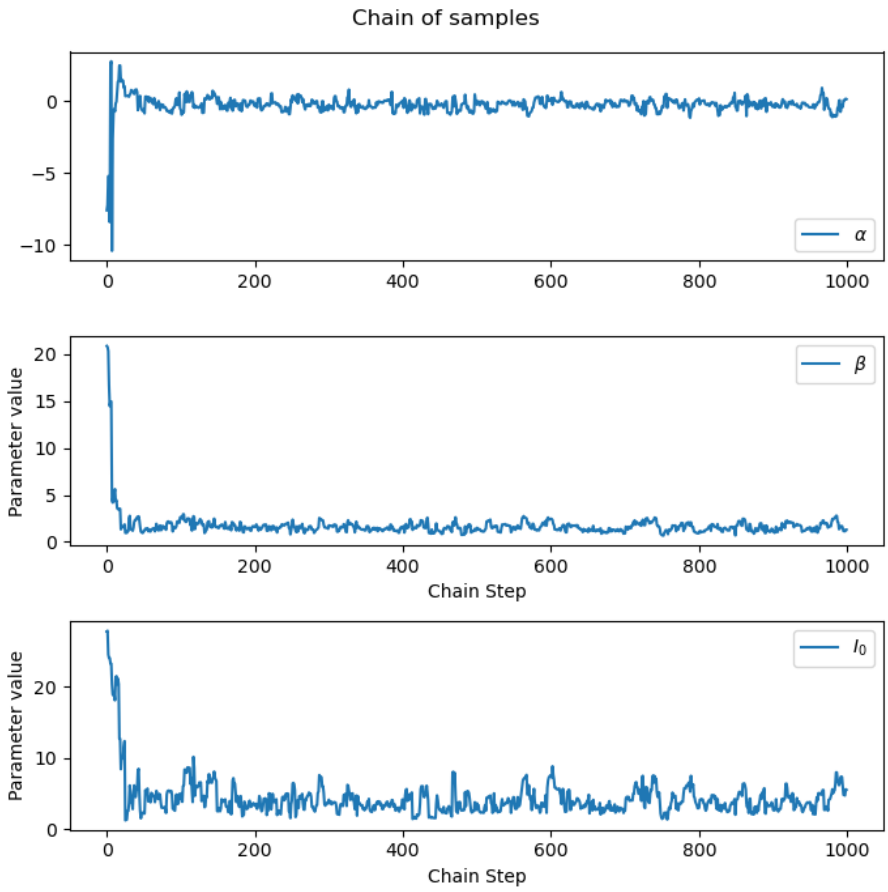

Figure 11: First 1000 samples of the SS chain for the lighthouse problem

The autocorrelation time also increases for this algorithm, with a value of 9 steps. Thus, the chain was thinned by a factor of 9. The following samples were kept:

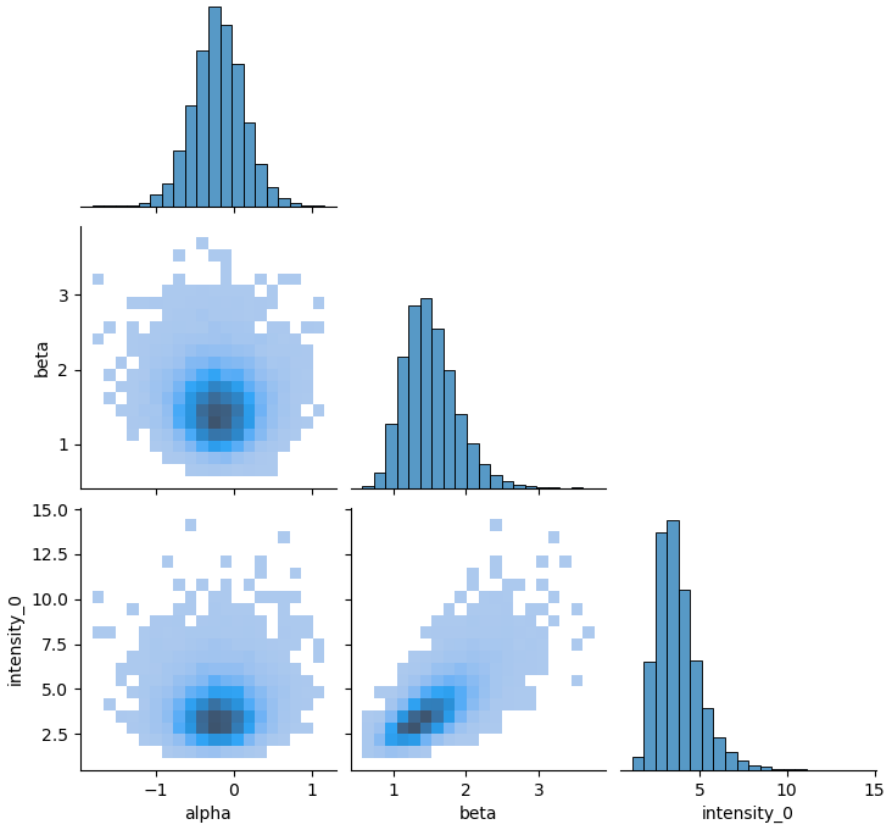

Figure 12: `zeus` Ensemble Slice Sampler samples for the lighthouse problem,

Giving the following estimates for α and β:

$$\hat{\alpha}_{SS} = -0.1945 \pm 0.3311 \text{ and } \hat{\beta}_{SS} = 1.5229 \pm 0.3639 \text{ and } \hat{I}_0 = 3.7814 \pm 1.2565 \quad (31)$$

(viii) Comparison of the results

First off, it is reassuring that the estimates of the 2 algorithms are very close. Adding the intensity measurement and parameter to the estimation does lead to a change in the values of α and β. More importantly the uncertainties do decrease, which does bring more confidence to the estimates.

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.