S2: Statistics for Data Science - Coursework

William Knottenbelt, wdk24

Word Count: 2904

April 3, 2024

Part i)

A flash with the angle, $\theta \in (-\frac{\pi}{2}, \frac{\pi}{2})$, and location along the shore, x, are related to the perpendicular distance of the lighthouse to the shore, β and the location of the lighthouse along the shore, α, by the equation

$$\tan(\theta) = \frac{x - \alpha}{\beta},$$

This gives the monotonically increasing functions $\theta : \mathbb{R} \to (-\frac{\pi}{2}, \frac{\pi}{2})$:

$$\theta = \arctan \frac{x - \alpha}{\beta}, \tag{1}$$

and $x : (-\frac{\pi}{2}, \frac{\pi}{2}) \to \mathbb{R}$

$$x = \alpha + \beta \tan \theta. \tag{2}$$

Part ii)

Let $g(x)$ and $u(\theta)$ be the PDFs of x and θ, respectively. θ is uniformly distributed over $(-\frac{\pi}{2}, \frac{\pi}{2})$, so we have

$$u(\theta) = \begin{cases} \frac{1}{\pi} & \text{for } \theta \in (-\frac{\pi}{2}, \frac{\pi}{2}), \\ 0 & \text{otherwise.} \end{cases}$$

Equation (1) is monotonically increasing, thus the probability for interval $[\theta, \theta + d\theta)$ is $u(\theta)d\theta = g(x)dx$. Hence, for $\theta \in (-\frac{\pi}{2}, \frac{\pi}{2})$, we have $g(x) = u(\theta)|\frac{d\theta}{dx}| = \frac{1}{\pi}|\frac{d\theta}{dx}|$ (where the absolute value is taken to enforce positive probability density). The likelihood for a single flash $\mathcal{L}_x(x|\alpha, \beta)$ is just given by the probability density function evaluated at x,

$$\mathcal{L}_x(x|\alpha, \beta) = g(x) = \frac{1}{\pi} \left| \frac{d\theta}{dx} \right|.$$

Using equation (1) and the identity $\frac{d}{dz} \arctan(z) = \frac{1}{1+z^2}$, we get

$$\frac{d\theta}{dx} = \frac{1}{1 + (\frac{x-\alpha}{\beta})^2} \frac{1}{\beta} = \frac{\beta}{\beta^2 + (x - \alpha)^2}.$$

This expression is always positive, thus:

$$\mathcal{L}_x(x|\alpha, \beta) = g(x) = \frac{\beta}{\pi(\beta^2 + (x - \alpha)^2)}. \tag{3}$$

Part iii)

The colleague is indeed correct that the mode of x is α. This can be seen by setting the first derivative of the PDF, $g(x)$, to zero:

$$0 = \frac{dg}{dx} = -\frac{2\beta(x-\alpha)}{\pi(\beta^2 + (x-\alpha)^2)^2},$$

which is only solved by $x = \alpha$. We can check that this is a maximum by checking $\frac{d^2 f}{dx^2}|_{x=\alpha} < 0$.

However, the colleagues' suggestion to use sample mean to estimate α is flawed since it does not converge, hence it is not a consistent estimator. The reason for this is that equation (3) is the Cauchy distribution, which has extremely heavy tails. Both the mean and the standard deviation of the Cauchy distribution are undefined [1], since their integrals do not converge. The sample mean of independent Cauchy random variables is itself a Cauchy random variable with the same parameters (α, β) [2]. Hence, the sample mean also has an undefined mean and standard deviation, and will not converge.

To illustrate this point, we ran a simulation experiment to compare the sample mean to the maximum likelihood estimate (MLE) of α. For distributions with finite moments, maximum likelihood estimates are consistent, efficient and unbiased. However, in the case of the Cauchy distribution the moments are not defined so consistency must be shown via simulation. For each sample size, N, in a series of sample sizes, we generated $M = 50$ samples from the Cauchy distribution by sampling $\theta \in (-\frac{\pi}{2}, \frac{\pi}{2})$ uniformly and transforming it using equation (2). For each sample, $\{x_k \sim \text{Cauchy}(\alpha, \beta)\}_{k=1}^{N}$, we computed the sample mean, $\bar{x} = \frac{1}{N}\sum_k^N x_k$. The likelihood function is:

$$\mathcal{L}_x(x|\alpha, \beta) = \Pi_{k=1}^{N} \frac{\beta}{\pi(\beta^2 + (x_k - \alpha)^2)} \tag{4}$$

There is no closed form expression for the maximum likelihood estimate of the location parameter, α, of the Cauchy distribution [3], thus it is computed using the Newton-Raphson method. We pick an initial guess, α_0, then iterate the following expression until convergence:

$$\alpha_{t+1} = \alpha_t - \frac{l'}{l''}, \tag{5}$$

where l', l'' are the first and second derivatives of the log-likelihood $l = \ln \mathcal{L}_x$ with respect to α, respectively. Expressions for l, l' and l'' can be found in appendix A. We then calculated the mean squared error (MSE) for each estimator as

$$MSE = \frac{1}{M}\sum_{i=1}^{M}(\hat{\alpha}_i - \alpha)^2,$$

where $\hat{\alpha}_i$ is the estimate from the i^{th} sample. This process was repeated for a series of sample sizes.

We performed this experiment using two sets of parameters: $(\alpha = 0, \beta = 1)$ and $(\alpha = 1, \beta = 2)$, to verify that the behaviour is consistent for different parameters. The results are plotted in Fig. 1. We see that for both sets of parameters, the MSE of the maximum likelihood estimates converges to zero as the sample size increases, which supports our hypothesis that MLE is a consistent estimator for α.

As expected, for the sample means there is no sign of convergence for either sets of parameters, and the MSE is several orders of magnitude larger than for the MLE. This is consistent with our earlier discussion that the sample mean does not converge and therefore cannot be used to estimate α.

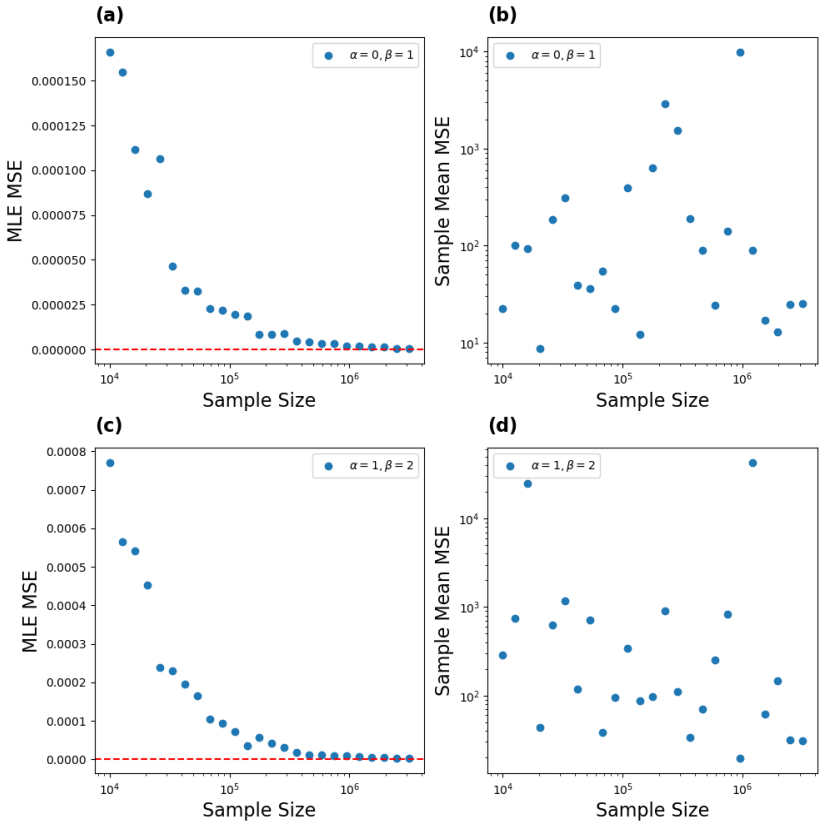

Figure 1: Plot showing the mean squared error (MSE) of estimators of the location parameter, α, of the Cauchy distribution, for a range of sample sizes. Each calculation of MSE is based on 50 Cauchy samples of a given size. Using samples generated from the standard Cauchy distribution, we plot MSE of the (a) maximum likelihood estimate and (b) sample mean. Similarly, for Cauchy with $\alpha = 1, \beta = 2$, we plot MSE of the (c) MLE and (d) sample mean. The MSE of MLE are given on a linear scale, with an axis line at 0, while for the sample mean they are on a logarithmic scale. The sample sizes are given on a logarithmic scale.

Part iv)

For α we chose to use an improper flat prior over all the reals. This was selected since we have no information about α so we want an uninformative uniform prior. We do not set upper/lower bounds so as to not accidentally exclude an important region of the parameter space from the posterior. This lack of normalisation should not negatively impact the Bayesian inference, since we do not require that the posterior is normalised in order to sample from it. We also have no information about β except that $\beta > 0$, hence we chose an improper flat prior with a lower bound at 0. There is no reason to believe the parameters are dependant on each other, thus the joint prior is just the product of each.

Part v)

The posterior, $P(\alpha, \beta|\{x_k\})$, is proportional to the likelihood in equation (4) for all $\alpha \in \mathbb{R}, \beta > 0$, and zero elsewhere. To sample from this posterior, we chose to use the No-U-Turn Sampler (NUTS) [4], which is an extension of Hamiltonian Monte Carlo (HMC) sampling [5] that attempts to automate the user inputs. This was chosen for its ability to explore the sample space efficiently, leading to a low autocorrelation time, without the risk of user error.

At a given point in the Markov chain $\mathbf{x}_i \in \mathbb{R}^d$, HMC sampling works by generating a momentum variable $\mathbf{p}_i \in \mathbb{R}^d$ from a multivariate Gaussian $Q = \mathcal{N}(0, M)$, with mean 0 and positive definite covariance M (called the 'mass' matrix). We define the Hamiltonian by $H(\mathbf{x}, \mathbf{p}) \propto -\ln[P(\mathbf{x})] - \ln[Q(\mathbf{p})]$, where P is the target distribution, and evolve the system according to Hamiltonian dynamics:

$$\frac{dx^{(j)}}{dt} = \frac{\partial H}{\partial p^{(j)}} \qquad , \qquad \frac{dp^{(j)}}{dt} = -\frac{\partial H}{\partial x^{(j)}}, \tag{6}$$

where superscript (j) denotes components of position/momentum. Integrating exactly is intractable, so we evolve the system via L discrete Leapfrog steps of size Δt (algorithm 1), to a new point in phase space $(\mathbf{x}', \mathbf{p}')$. Leapfrog does not guarantee that the Hamiltonian is conserved (as it is for exact Hamiltonian dynamics), so we accept the transition with a probability of

$$a = e^{H(\mathbf{x}_i, \mathbf{p}_i) - H(\mathbf{x}', \mathbf{p}')}. \tag{7}$$

Algorithm 1 Leapfrog Step

1: **Input:** Current position $\mathbf{x}$, current momentum $\mathbf{p}$, target distribution $P(\mathbf{x})$, time-step size Δt.
2: $\tilde{\mathbf{p}} \leftarrow \mathbf{p} - \frac{\Delta t}{2} \nabla_{\mathbf{x}} \ln P(\mathbf{x})$
3: $\tilde{\mathbf{x}} \leftarrow \mathbf{x} + \Delta t \tilde{\mathbf{p}}$
4: $\tilde{\mathbf{p}} \leftarrow \tilde{\mathbf{p}} - \frac{\Delta t}{2} \nabla_{\mathbf{x}} \ln P(\mathbf{x})$
5: **return** $\tilde{\mathbf{x}}$, $\tilde{\mathbf{p}}$

If accepted, the next position in the chain becomes $\mathbf{x}_{i+1} = \mathbf{x}'$, otherwise $\mathbf{x}_{i+1} = \mathbf{x}_i$. This results in a time-homogeneous Markov chain which satisfies the detailed balance condition, ensuring that the target distribution is the stationary distribution of the chain [6]. This algorithm explores the sample space very efficiently resulting with a low autocorrelation time, however it is very sensitive to the choice of $L, \Delta t$ and M. If $L\Delta t$ is too large, we get wasteful oscillations around the sample space, resulting in unnecessary computation time. Conversely, if $L\Delta t$ is too small, we get random walk behaviour,

and the autocorrelation time is high. Also, if Δt is high then the Leapfrog integration is inaccurate, resulting in a low acceptance rate, and increasing the autocorrelation time.

The No-U-Turn Sampler (NUTS) [4] automates L by evolving the trajectory in phase space forwards and backwards in time until a U-Turn occurs on either path, and then chooses the next point in the chain from the trajectory with probabilities weighted according to the target distribution density at each point. The acceptance criterion in equation (7) is then used. This algorithm allows the particle to travel as far as possible without wasteful oscillations, and still satisfies detailed balance. NUTS also uses an initial tuning phase to tune the step size parameter, Δt, with a method based on primal-dual averaging [7], to strike an optimal balance between efficiency and autocorrelation time. We used the PyMC [8] implementation of NUTS, which adapts the (diagonal) mass matrix during this tuning phase, starting from the identity matrix, to match the variance of the samples so far. This completely eliminates the need for the user to pick any parameters.

We ran $M = 4$ chains, with a tuning phase of 500 steps, an additional 50 steps discarded as a burn-in, and then $N = 10000$ steps (draws). The burn-in is necessary since $L, \Delta t$ and M change during tuning, meaning that the algorithm does not satisfy detailed balance in this period, thus we cannot make the assumption that the chain has converged to the posterior by the end of the tuning phase. To assess convergence, we use the potential scale reduction factor [9], which is defined:

$$\hat{R} = \sqrt{\frac{\hat{V}}{W}}, \tag{8}$$

where

$$\hat{V} = \frac{N-1}{N}W + \frac{1}{N}B,$$

where W is within-chain variance (the mean of the variance of samples within each chain) and B is the between-chain variance. If all chains have converged to the target distribution, these will be the same and $\hat{R} = 1$, otherwise $\hat{R} > 1$. Gelman and Rubin (1992) [9] recommended a threshold of $\hat{R} < 1.1$ to statistically justify convergence. We calculate the "split" $\hat{R}$ statistic for each parameter (ArviZ implementation [10]), which involves splitting each chain into two halves before calculating $\hat{R}$ [11]. This provides an additional way to detect non-equilibrium behaviour in the individual chains. We obtain $\hat{R} < 1.0001$ for both α and β, which is very strong evidence that the burn-in was sufficient to allow the chains to converge. We visualise the first 500 draws of α and β (after the burn-in) for one chain in Fig. 2. The chains appear to be mixing well, with no overarching trends or patterns, which reinforces our hypothesis that the chains had reached their stationary state by the end of the burn-in.

For a single un-thinned Markov chain $\{f_i\}_{i=1}^N$ (burn-in removed), the empirical autocorrelation at lag δ is given [12]:

$$\hat{\rho}_f(\delta) = \frac{\sum_{i=1}^{N-\delta}(f_i - \bar{f})(f_{i+\delta} - \bar{f})}{\sum_{i=1}^{N}(f_i - \bar{f})^2}, \tag{9}$$

where $\bar{f} = \frac{1}{N}\sum_{i=1}^{N} f_i$. We estimate the integrated autocorrelation time (IAT) as [13]:

$$\hat{\tau}_f = 1 + 2\sum_{\delta=1}^{T}\hat{\rho}_f(\delta), \tag{10}$$

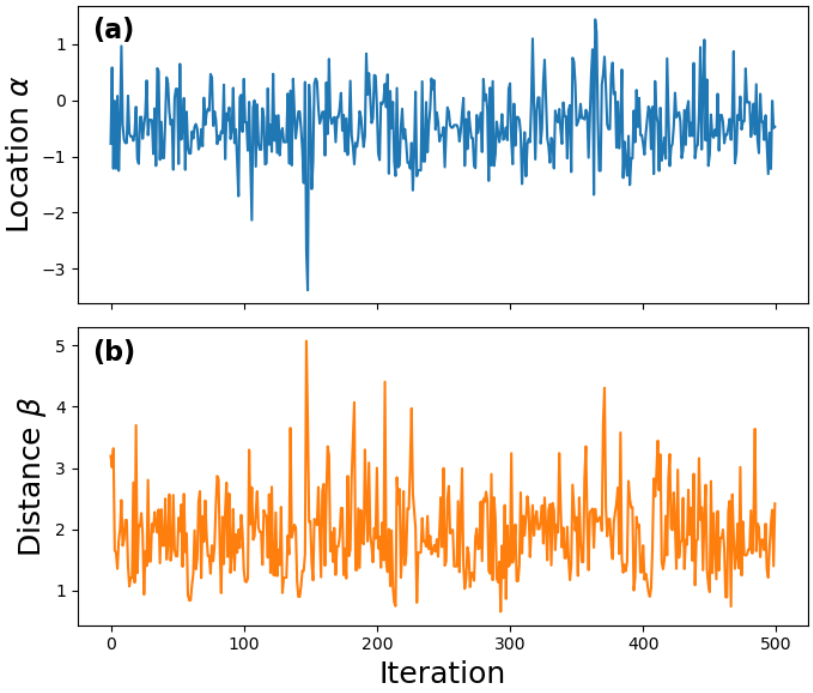

Figure 2: Trace plot depicting the first 500 draws of one Markov chain (after tuning and burn-in) for the (a) location, α, and (b) distance, β, of the lighthouse. The stationary distribution of the chain is the joint posterior $P(\alpha, \beta|\{x_k\})$ (based on only the flash position measurements). The chain was generated using the NUTS sampler.

where T is the 'window size'. It is important that $T \ll N$ to ensure convergence of the autocorrelation. We calculated the autocorrelation times, τ_α and τ_β, for each chain using a method from the emcee package [14], which uses an iterative procedure to find a good window size, as described in page 16 of Sokal's notes [15]. We then took the mean of the IATs across the chains, to produce estimates $\hat{\tau}_\alpha = 1.31$ and $\hat{\tau}_\beta = 1.30$. This method of estimating the IAT from multiple chains is recommended in the emcee documentation [13]. In Fig. 3, we plot the empirical autocorrelation functions, $\hat{\rho}_\alpha, \hat{\rho}_\beta$ (equation (9)) of all chains, for lags $1 \leq \delta \leq 50$. We see that the autocorrelation drops sharply for all chains, converging to roughly 0 after the first couple of lags. This demonstrates that the chain is mixing efficiently, allowing for a large effective sample size and more precise posterior measurements.

We choose not to thin the chains since it reduces the precision of measurements, as discussed by Link and Eaton (2012) [16], who recommend against thinning except for computer memory reasons. Furthermore, both estimates of the IAT are below 2, hence thinning would cause the loss of valuable statistics. We estimate the mean, $\hat{\mu}_f$, and standard deviation, $\hat{\sigma}_f$, for $f \in \{\alpha, \beta\}$, using the sample mean and standard deviation of the combined draws across all the chains [11]. Then, based on the Markov chain central limit theorem [17], we can estimate the Markov chain standard error on the mean as $MCSE = \frac{\hat{\sigma}_f}{\sqrt{\hat{N}_{eff}}}$ [11], where $\hat{N}_{eff}$ is the estimated effective sample size, given by [18] (BDA formula 11.8):

$$\hat{N}_{eff} = \frac{M \cdot N}{\hat{\tau}_f},$$

where M is the number of (stationary) chains, and N is the length of each chain. Quoting estimates

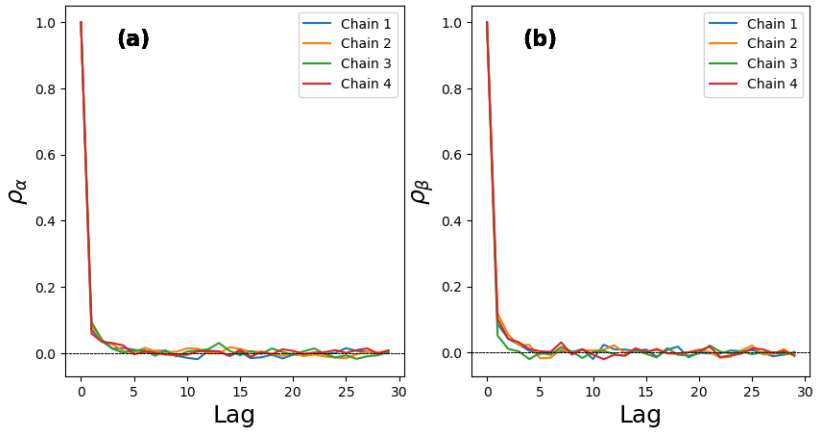

Figure 3: Plot of the empirical autocorrelation function of the (a) location, α, and (b) distance, β, of the lighthouse, for 4 Markov chains generated using the NUTS sampler. The joint posterior $P(\alpha, \beta|\{x_k\})$ (based on flash positions) is the stationary distribution.

as "mean $\pm$ sd", we obtain

$$\alpha = -0.447 \pm 0.604 \quad \text{(MCSE 0.003)}$$
$$\beta = 1.967 \pm 0.669 \quad \text{(MCSE 0.004)}.$$

The low MCSE indicates that these are precise estimates of the posterior means.

Using our combined draws across the chains, we constructed a 2D density histogram of the joint posterior, $P(\alpha, \beta|\{x_k\})$, depicted in Fig. 4. Bins were only plotted above a certain density threshold and elsewhere we scattered the draws to indicate outliers. We see that the joint posterior exhibits a single well-defined peak and significant skew in the positive β direction. This implies that the flash positions provide a good amount of information about the location and distance of the lighthouse, allowing for a well-defined central tendency to emerge when using uninformative flat priors. There appears to be no significant correlation between α and β, except for the possibility of a very slight positive correlation. This implies that the position of the lighthouse along the shore does not significantly influence our estimates of how far away it is from the shore.

We also plot histograms of the 1D marginal distributions, $P(\alpha|\{x_k\})$ and $P(\beta|\{x_k\})$, in Fig. 5, with the mean indicated by a vertical line. We see that $P(\alpha|\{x_k\})$ resembles a Gaussian bell-curve, which suggests that there are enough measured flash positions spread symmetrically around the true position for a well-defined peak in $P(\alpha|\{x_k\})$, despite the Cauchy distribution having no defined mean. The mean approximately lines up with the mode of the distribution, which makes the mean an effective point estimate for α. Furthermore, since the distribution is symmetric about the mean, the standard deviation is reflective of the true uncertainty in the lighthouse location.

$P(\beta|\{x_k\})$ also has a single peak, but with a long positive tail. This skew indicates more uncertainty in β, conveying that there is a non-negligible probability of the lighthouse being much further away from the shore than the mode of distribution suggests. We see that, as a result of the skew, the mean is higher than the MAP, suggesting that it is a potentially misleading point estimate. In addition, the standard deviation is a potentially misleading estimate of spread, since the distribution is not symmetric about the mean. To improve this analysis, it may be more appropriate to quote an estimate of the Maximum a

posteriori (MAP - mode of the joint posterior) as the point estimates, and use highest density intervals to indicate uncertainty.

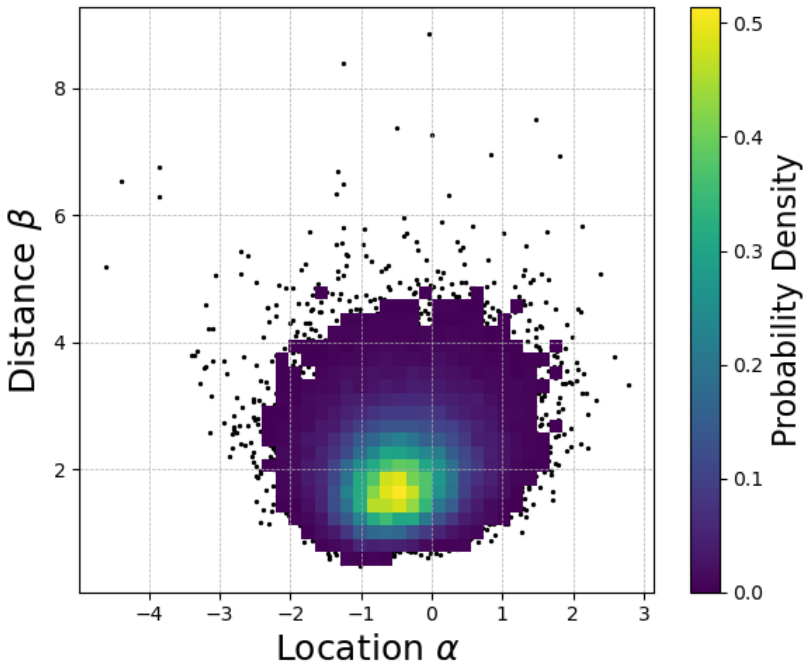

Figure 4: 2-dimensional density histogram depicting the joint posterior, $P(\alpha, \beta|\{x_k\})$, on the location, α, and distance, β, of the lighthouse, based on flash positions. Bins are only plotted for density $>$ 0.002, and elsewhere the samples are scattered, to indicate outliers. The colour of the bins indicates the probability density.

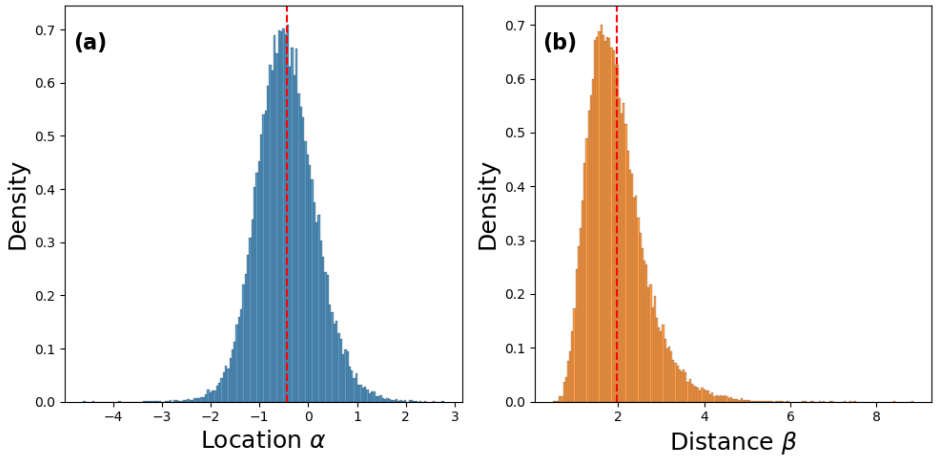

Figure 5: Density histograms depicting the marginal posterior distributions for the (a) location, α, and (b) distance, β, of the lighthouse, based on solely the position measurements of the flashes. Vertical red lines indicate the sample mean.

Part vi)

I_0 is the source (initial) intensity of the lighthouse. We chose to use Jeffrey's prior, $\pi(I_0) \propto \frac{1}{I_0}$, over the positive reals (improper). This was chosen since we have no information about I_0 except that it is positive (including no knowledge about the units it is measured in), hence we want a non-informative scale-invariant prior. We also do not want to set upper/lower bounds since we do not know the scale, thus we leave the prior un-normalised (improper). Practically, this was implemented using an improper flat prior for $\ln(I_0)$, then transforming to I_0.

Part vii)

Using the priors in part (iv) and (vi), we get

$$P(\alpha, \beta, I_0 | \{x_k\}, \{I_k\}) \propto \prod_{k=1} \frac{\beta}{I_0 \pi(\beta^2 + (x_k - \alpha)^2)} \exp\left[-\frac{1}{2\sigma^2}\left(\ln(I_k) - \ln(\frac{I_0}{\beta^2 + (x_k - \alpha)^2})\right)\right] \quad (11)$$

for all $\alpha \in \mathbb{R}, \beta > 0, I_0 > 0$, and zero elsewhere.

We sample from this posterior using the same NUTS sampler as in part (v) for $M = 4$ chains, using 500 tuning steps, 50 burn-in steps, and 10000 draws in each chain. We calculate the split $\hat{R}$ statistics, yielding values below 1.001 for all parameters, which provides strong evidence of convergence. We visualise the first 500 draws of the first chain in Fig. 6, and observe that the chains appear to be mixing well, as is characteristic of equilibrium behaviour.

We estimated the autocorrelation times using the same method as part (v), yielding $\hat{\tau}_\alpha = 1.72, \hat{\tau}_\beta = 1.99, \hat{\tau}_{I_0} = 1.98$. These short autocorrelation times show that NUTS is an effective sampler for this posterior, yielding a high effective sample size for precise posterior measurements. For the same reasons in part (v), we did not thin the chains. Using the same methods as in part (v), we estimate the posterior mean and standard deviation, as well as the Markov chain standard error (MCSE) on the mean for each parameter. We obtain:

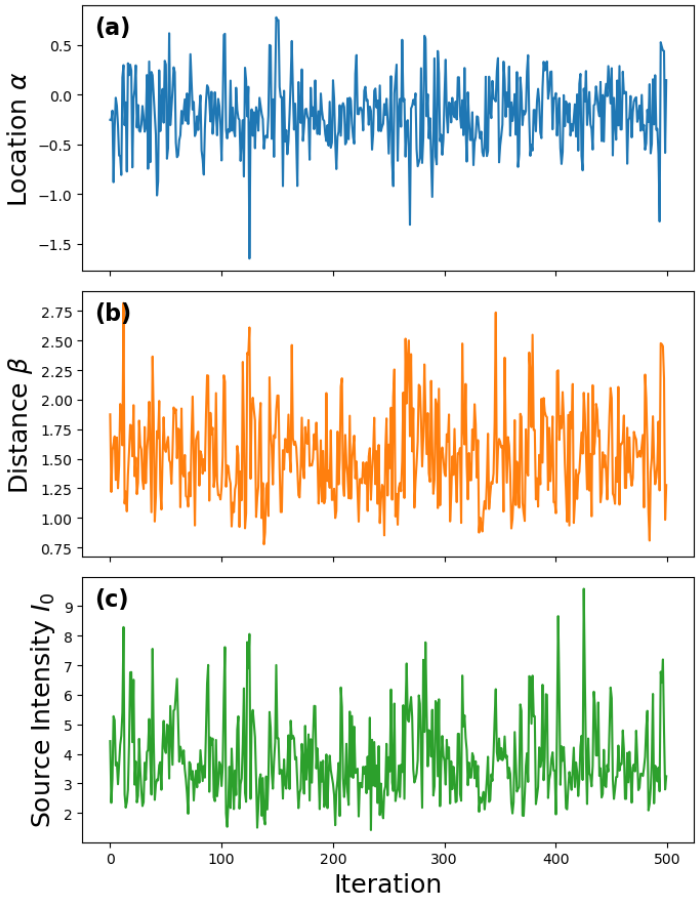

Figure 6: Trace plot depicting the first 500 draws of one Markov chain (after tuning and burn-in) for the (a) location, α, (b) distance, β, and (c) source intensity, I_0, of the lighthouse. The stationary distribution of the chain is the posterior, $P(\alpha, \beta, I_0|\{x_k\}, \{I_k\})$ (based on the flash position and intensity measurements). The chain was generated using the NUTS sampler.

$$\alpha = -0.197 \pm 0.326 \quad \text{(MCSE 0.002)},$$
$$\beta = 1.515 \pm 0.361 \quad \text{(MCSE 0.003)},$$
$$I_0 = 3.765 \pm 1.247 \quad \text{(MCSE 0.009)}.$$

In Fig. 7 we depict a corner plot of the 2D marginal distributions. Each plot was constructed as a scatter plot overlaid by a 2D density histogram, using the draws of all chains collectively. We see that all plots exhibit a single well defined peak, which suggests that the data is sufficient to provide a high degree of confidence about a single set of parameter values. The marginal distribution $P(\alpha, \beta|\{x_k\}, \{I_k\})$ in Fig. 7(a) has a similar shape to the joint posterior from part (v), except with a narrower peak, which reflects the improved state of knowledge from including the intensity measurements in the inference. The marginal distribution $P(\alpha, I_0|\{x_k\}, \{I_k\})$ has a skew in the positive I_0 direction, and no clear correlation between I_0 and α. This suggests that the location of the lighthouse along the shore is not informative about its source intensity, and vice versa. In Fig. 7(c), we see that there is a strong positive correlation between β and I_0. This is natural since the mean intensity decays with distance

by the inverse square law, and a larger distance between the lighthouse and the shore means that the light has to travel further, hence the source intensity would need to be higher to achieve the same measured intensity.

In Fig. 8 we plot histograms of the 1D marginal distributions, and indicate the means using vertical lines. Both α and β have similarly-shaped marginal distributions as in part (v), except both distributions have narrower peaks. This demonstrates that the inclusion of intensity data has allowed for more precise estimates of the location and distance of the lighthouse. Additionally, $P(\beta|\{x_k\},\{I_k\})$ (Fig. 7(b)) has a much shorter tail than the same marginal from part (v) (Fig. 4(b)), which suggests that considering the intensity measurements has substantially constrained the upper limits of how far from the shore the lighthouse can plausibly be. This is because measuring intensity naturally informs the range of distances over which the light may have travelled. Despite the tail being shorter, we see that the mean is still slightly higher than the peak of the distribution, which suggests that the mean estimate is potentially misleading. In Fig. 8(c) we see that the distribution for I_0 has a single peak and a significant positive skew. This indicates that there is a non-negligible probability that the actual source intensity could be significantly higher than our mean and standard deviation estimates suggest. This uncertainty is not properly reflected by the standard deviation estimate, as the distribution is asymmetric about the mean. Furthermore, this tail skews the mean to be slightly higher than the mode of the distribution, making it a potentially misleading point estimate for I_0.

Part viii)

In part (v), we estimated $\alpha = -0.447 \pm 0.604$, and in part (vii) we estimated $\alpha = -0.196 \pm 0.325$. Both of the marginal posteriors on α are symmetric about the mean, making these standard deviations reflective of the true uncertainty in α given by the posterior. The inference in part (v) is based only on the flash positions, whereas in part (vii) we included both the flash position and intensity measurements. Hence, the inclusion of the intensity data has yielded a marginal posterior on α with almost half the standard deviation as that with only the flash positions, demonstrating an improved state of knowledge about the location of the lighthouse.

This makes intuitive sense as the mean intensity decays with distance travelled according to the inverse square law, hence we would expect the flash positions with higher intensity to be closer to the lighthouse. This means that including the intensity data would naturally allow us to make a more precise estimate of α, based on where the higher intensity flashes are located compared to the low intensity flashes. Furthermore, the Cauchy distribution has heavy tails, hence it is natural that the Cauchy-distributed flash positions give an uncertain estimate of α, which can be substantially improved by the intensity measurements that follow a log normal distribution.

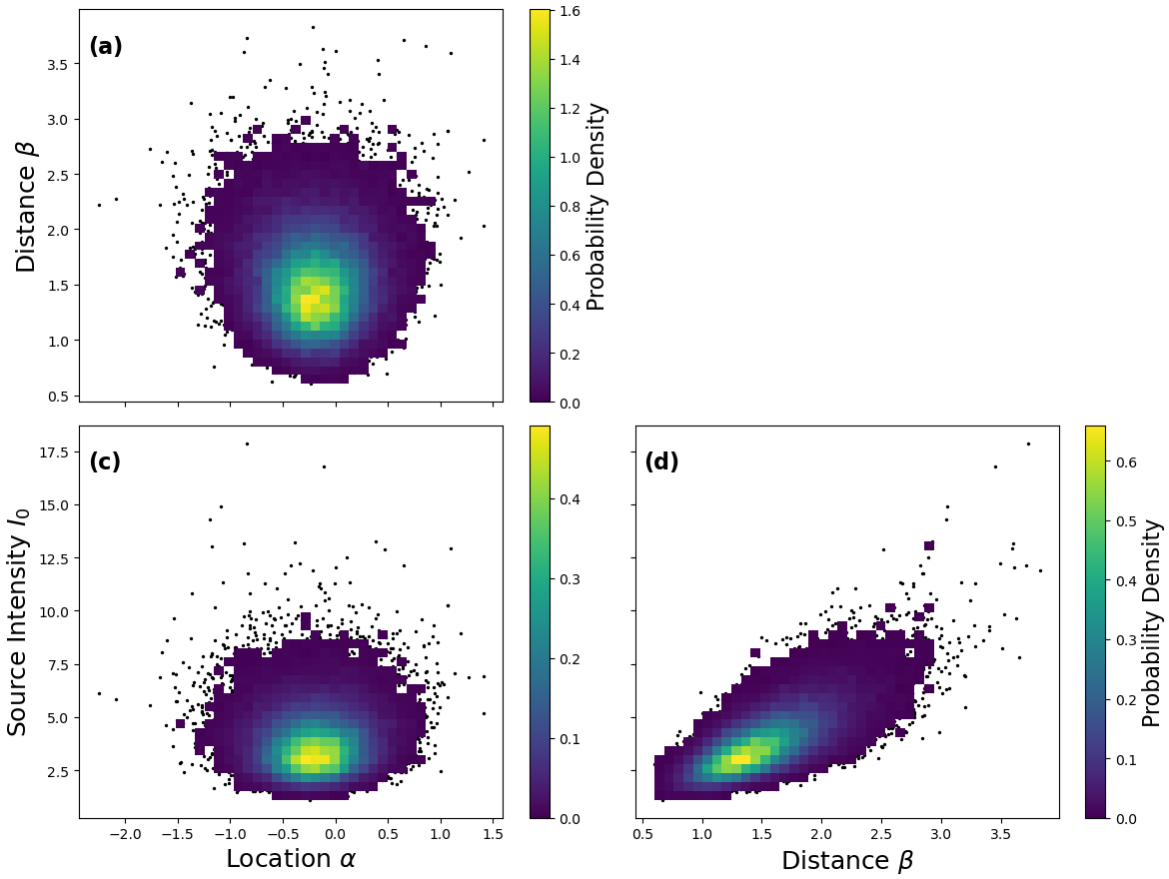

Figure 7: Corner plot of the 2D marginal distributions, (a) $P(\alpha, \beta|\{x_k\}, \{I_k\})$, (b) $P(\alpha, I_0|\{x_k\}, \{I_k\})$ and (c) $P(\beta, I_0|\{x_k\}, \{I_k\})$, based on the flash positions and intensities. The distributions are plotted as scatter plots overlaid by density histograms, using the combined draws across four Markov chains generated with NUTS. The colour of the bins indicates the probability density.

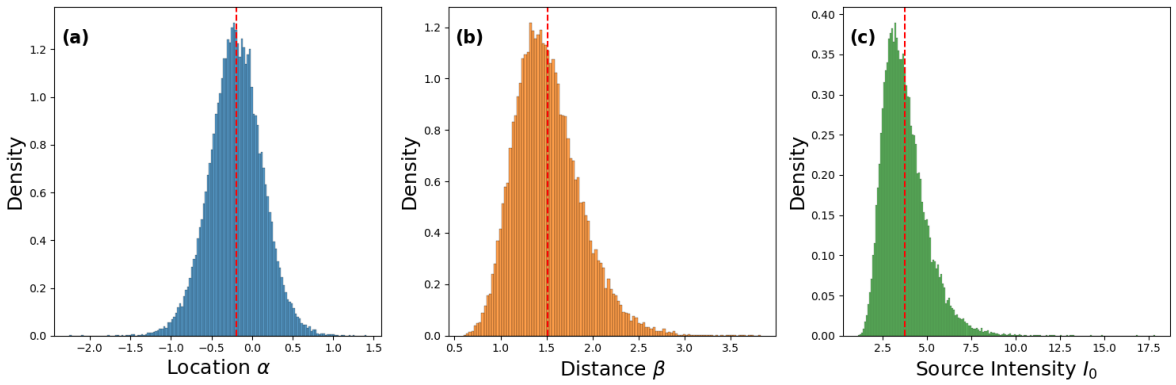

Figure 8: Density histograms depicting the 1D marginal posterior distributions for the (a) location, α, (b) distance, β, and (c) source intensity, I_0, of the lighthouse, based on position and intensity measurements of the flashes. Vertical red lines indicate the sample mean.

References

[1] National Institute of Standards and Technology. (2012). *e-Handbook of Statistical Methods*, Section 1.3.6.6.3. Cauchy Distribution. Available at: https://www.itl.nist.gov/div898/handbook/eda/section3/eda3663.htm

[2] Michael P. Cohen (2012) *Sample Means of Independent Standard Cauchy Random Variables Are Standard Cauchy: A New Approach*, The American Mathematical Monthly, 119:3, 240-244. DOI: 10.4169/amer.math.monthly.119.03.240

[3] Gerald Haas, Lee Bain, and Charles Antle (1970). *Inferences for the Cauchy Distribution Based on Maximum Likelihood Estimators*, Biometrika, 57(2), 403–408.

[4] Matthew D. Hoffman & Andrew Gelman (2014). *The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo*, Journal of Machine Learning Research, 15, 1593-1623.

[5] Radford M. Neal (2011). *MCMC using Hamiltonian Dynamics*, Chapter 5 in the Handbook of Markov Chain Monte Carlo, edited by Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng.

[6] Jizhou Kang (2021). *Chapter 10 Limiting Distribution of Markov Chain*, In STAT 243: Stochastic Process. Available at: https://bookdown.org/jkang37/stochastic-process-lecture-notes/lecture10.html.

[7] Y. Nesterov (2009). *Primal-dual Subgradient Methods for Convex Problems*, Mathematical Programming, 120(1), 221–259.

[8] Oriol Abril-Pla, Virgile Andreani, Colin Carroll, Larry Dong, Christopher J. Fonnesbeck, Maxim Kochurov, Ravin Kumar, Jupeng Lao, Christian C. Luhmann, Osvaldo A. Martin, Michael Osthege, Ricardo Vieira, Thomas Wiecki, Robert Zinkov (2023). *PyMC: A Modern and Comprehensive Probabilistic Programming Framework in Python*, PeerJ Computer Science, 9, e1516. DOI: 10.7717/peerj-cs.1516

[9] Andrew Gelman and Donald B. Rubin (1992). *Inference from Iterative Simulation Using Multiple Sequences*, Statistical Science, 7(4), 457-472.

[10] Ravin Kumar, Colin Carroll, Ari Hartikainen, Martin Osvaldo (2023). *ArviZ a unified library for exploratory analysis of Bayesian models in Python*, Journal of Open Source Software.

[11] Stan Development Team. (2024). "Chapter 16: Posterior Analysis". in: Stan Reference Manual, version 2.34. https://mc-stan.or

[12] Charles J. Geyer (2011). *Section 1.8.2 The Autocovariance Function. "Introduction to MCMC."* In *Handbook of Markov Chain Monte Carlo*, edited by S. P. Brooks, A. E. Gelman, G. L. Jones, and X. L. Meng. Chapman & Hall/CRC, Boca Raton, pp. 3-48.

[13] D. Foreman-Mackey, D. W. Hogg, D. Lang, J. Goodman (2013). *Autocorrelation analysis & convergence*. Emcee documentation. Available at: https://emcee.readthedocs.io/en/stable/tutorials/autocorr/

[14] D. Foreman-Mackey, D. W. Hogg, D. Lang, J. Goodman (2013). *emcee: The MCMC Hammer*, PASP, 125, 306-312. DOI: 10.1086/670067

[15] Alan D. Sokal (1996). *Monte Carlo Methods in Statistical Mechanics: Foundations and New Algorithms Note to the Reader*. Available at: https://api.semanticscholar.org/CorpusID:14817657

[16] William Link and Mitchell Eaton (2011). *On thinning of chains in MCMC*, Methods in Ecology and Evolution, 3, 112-115. DOI: 10.1111/j.2041-210X.2011.00131.x

[17] M. I. Gordin and B. A. Lifšic (1978). *Central limit theorem for stationary Markov processes*, Soviet Mathematics, Doklady, 19, 392-394. (English translation of Russian original).

[18] Gelman et al. (2014). *Bayesian Data Analysis*. Available at: http://www.stat.columbia.edu/~gelman/book/BDA3.pdf

Appendix

A Likelihood of Cauchy Distribution

The log-likelihood function for a set of independent Cauchy random variables is

$$l = \ln \mathcal{L}_x(x|\alpha, \beta) = N \ln \frac{\beta}{\pi} - \sum_{k=1}^{N} \ln \left(\beta^2 + (x_k - \alpha)^2\right)$$

The first derivative with respect to α is

$$l' = \frac{\partial l}{\partial \alpha} = \sum_{k=1}^{N} \frac{2(x_k - \alpha)}{\beta^2 + (x_k - \alpha)^2}$$

The second derivative is then

$$l'' = \frac{\partial^2 l}{\partial \alpha^2} = -2 \sum_{k=1}^{N} \frac{\beta^2 - (x_k - \alpha)^2}{(\beta^2 + (x_k - \alpha)^2)^2}$$

B Auto-Generation Tools

Github copilot was used to assist with basic programming tasks like plotting figures and writing print statements. It was also occasionally used to assist with the generation of doc-strings of functions and files.