Ivo Vladislavov Petrov

Statistics for Data Science

Coursework Report

Data-Intensive Science

March 28, 2024

Word count: 2977 (using Overleaf)

Contents

List of Figures

Chapter 1

Introduction

Bayesian inference offers a rigorous framework for tackling parameter inference problems, allowing for the incorporation of prior knowledge and the measurement of uncertainty in parameter estimates. This problem, originally posed on a Cambridge problem sheet by S. Gull and discussed in D.S. Sivia's "Data Analysis: A Bayesian Tutorial," involves a lighthouse situated at an unknown location along a straight coastline and a certain distance out to sea. The lighthouse emits light flashes at random angles, which are detected by an array along the coast, registering the locations of these flashes without direction. Its solution not only demonstrates the power of Bayesian methods in parameter estimation but also lays foundational concepts for applied statistics and data science.

This coursework explores the application of Bayesian inference to the Lighthouse Problem, employing Markov Chain Monte Carlo (MCMC) techniques to sample from the posterior distributions of the lighthouse's position and other parameters. We discuss not only the practical nature of the problem but also demonstrate rigorous approaches to verifying the obtained results are valid and robust.

Chapter 2

Parameter Estimation from Detection Location

2.1 Part (i)

By observing the schematic shown on Figure 2.1, we can derive the relationship:

$$\tan\theta = \frac{x-\alpha}{\beta} \iff \theta = \tan^{-1}\left(\frac{x-\alpha}{\beta}\right) \tag{2.1}$$

This can be further rewritten as an expression for x as:

$$x = \alpha + \beta\tan\theta \tag{2.2}$$

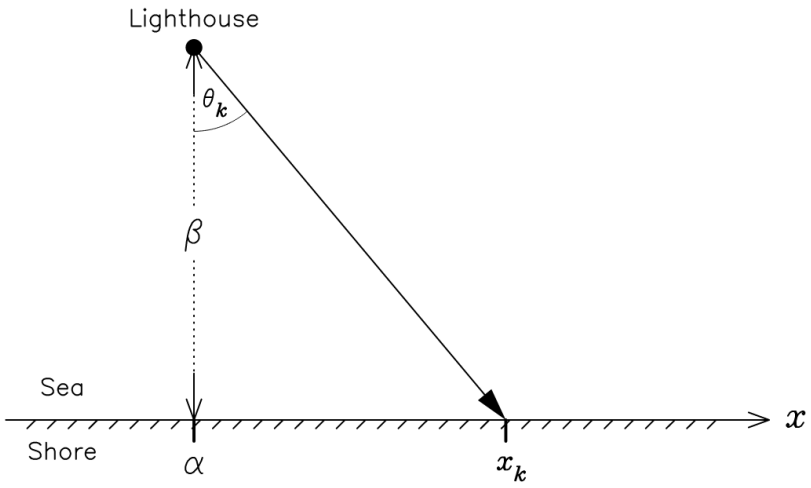


Figure 2.1: *A geometric illustration of the lighthouse problem. The figure was taken from the original textbook by D. S. Sivia* [1].

Definition 2.1.1 Below we explain each of the aforementioned terms in the context of the lighthouse problem:

- α – the location of the lighthouse along the coast.
- β – the height of the lighthouse's source.

- θ – the azimuthal angle at which the light leaves the source.
- x – the position along the coast, at which the light has been detected.

2.2 Part (ii)

We are given that the distribution for the angle θ as uniform, or:

$$\mathcal{L}_\theta(\theta|\alpha,\beta) = \mathbb{1}_{(-\frac{\pi}{2},\frac{\pi}{2})}\frac{1}{\pi} \tag{2.3}$$

To determine our likelihood using x as our data, we need to apply the change of variables formula in Definition 2.2.1 by calculating the Jacobian to the transformation given in Equation 2.1.

Definition 2.2.1 For 2 random variables X, Y, such that $Y = g(x)$, where g is invertible, the relationships between the respective probability density functions p_X, p_Y can be expressed by:

$$p_Y(y) = p_X(g^{-1}(y))\left|\frac{dg^{-1}(y)}{dy}\right| \tag{2.4}$$

Substituting for $X = \theta$, $Y = x$, we find that:

$$\mathcal{L}_x(x|\alpha,\beta) = \mathcal{L}_\theta(\theta|\alpha,\beta)\left|\frac{d\tan^{-1}(\frac{x-\alpha}{\beta})}{dx}\right| = \mathcal{L}_\theta(\theta|\alpha,\beta)\left|\frac{1}{\beta}\frac{1}{1+\frac{(x-\alpha)^2}{\beta^2}}\right| =$$
$$\mathcal{L}_\theta(\theta|\alpha,\beta)\left|\frac{\beta}{\beta^2+(x-\alpha)^2}\right| = \mathbb{1}_{(-\frac{\pi}{2},\frac{\pi}{2})}\frac{\beta}{\pi(\beta^2+(x-\alpha)^2)}$$

The bound $\theta \in (-\frac{\pi}{2},\frac{\pi}{2})$ is equivalent to $\tan\theta \in (-\infty,\infty) \iff x \in (-\infty,\infty)$ (from Eq. 2.2). Therefore, we can conclude that $\forall x.\mathbb{1}_{(-\frac{\pi}{2},\frac{\pi}{2})} = 1$, or finally:

$$\mathcal{L}_x(x|\alpha,\beta) = \frac{\beta}{\pi(\beta^2+(x-\alpha)^2)} \tag{2.5}$$

2.3 Part (iii)

To present an argument to the colleague, we need to consider the two parts of the question - showing that the Maximum Likelihood Estimator for α is given by $\hat{\alpha} = \text{argmax}_x\,\mathcal{L}_x(x|\alpha,\beta)$, as well as showing that the sample mean $\frac{1}{N}\sum_k x_k$ is not a good estimator for α.

2.3.1 Maximum Likelihood Location

We can infer from the derived likelihood in Equation 2.5 that because $(x-\alpha)^2 \geq 0$, it follows that:

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

The resulting upper bound is a constant, for which equality holds if and only if $(x-\alpha)^2 = 0$ or when $x = \alpha$. Therefore, the most likely location is indeed $\hat{x} = \alpha$.

We now proceed to evaluate the Maximum Likelihood Estimator for α. The log-likelihood for multiple samples can be given as:

$$\log(\mathcal{L}_x(\{x_k\}|\alpha,\beta)) = \log(\prod_{i=0} \frac{\beta}{\pi(\beta^2+(x_i-\alpha)^2)}) = \sum_{i=0} \log(\beta^2+(x_i-\alpha)^2) + C$$

Where C is a constant that does not rely on *alpha*

$$\frac{d}{d\alpha}\log(\mathcal{L}_x(\{x_k\}|\alpha,\beta)) = \sum_{i=0} \frac{2(x_i-\alpha)}{\beta^2+(x_i-\alpha)^2} = 0$$

This equation is infeasible to solve analytically for an arbitrary number of samples, but the numerical result will likely give us a good estimate. It is, however, simple enough for a single sample x_1. The solution we obtain is again $\hat{\alpha} = x_1$

2.3.2 Average Location Estimator

As shown in Part (ii), the likelihood distribution for x is a *Cauchy* distribution. We will now show that the Cauchy distribution does not have a defined mean (and hence has an undefined variance), meaning that the estimates we receive will not be valid.

Statement 2.3.1 The Cauchy distribution in the form $P(x|\alpha,\beta) = \frac{\beta}{\pi(\beta^2+(x-\alpha)^2)}$ does not have a defined mean.

To prove this statement, we begin by proving a property of the Cauchy distribution:

Lemma 2.3.2 The characteristic function of the Cauchy distribution can be given by $\varphi_X(t) = \mathbb{E}[e^{itX}] = e^{\alpha it-\beta|t|}$

Proof 2.3.3 We can check this is correct by inversing the characteristic function to obtain the probability density function: $P_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \varphi_X(t)e^{-itx}dt$.

$$P_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{\alpha it-\beta|t|}e^{-itx}dt = \frac{1}{2\pi}(\int_{-\infty}^{0} e^{\alpha it+\beta t-itx}dt + \int_{0}^{\infty} e^{\alpha it-\beta t-itx}dt) =$$

$$= \frac{1}{2\pi}(\int_{-\infty}^{0} e^{(i(\alpha-x)+\beta)t}dt + \int_{0}^{\infty} e^{(i(\alpha-x)-\beta)t}dt) = \frac{1}{2\pi(i(\alpha-x)+\beta)}e^{(i(\alpha-x)+\beta)t}\Big|_{t=-\infty}^{t=0} +$$

$$+\frac{1}{2\pi(i(\alpha-x)-\beta)}e^{(i(\alpha-x)-\beta)t}\Big|_{t=0}^{t=\infty} = \frac{1}{2\pi}(\frac{1}{i(\alpha-x)+\beta} - \frac{1}{i(\alpha-x)-\beta}) = \frac{-2\beta}{2\pi(-\beta^2-(x-\alpha)^2)}$$

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

This is exactly the PDF we obtain for the Cauchy distribution.

We now continue with the proof of Statement 2.3.1:

Proof 2.3.4 The mean of a distribution can usually be evaluated through integration. However, this causes unnecessary complications for the Cauchy distribution. Therefore, we will show the undefined nature of the mean using the characteristic function of the

Cauchy distribution. Using Lemma 2.3.2, we can compute the mean using the first derivative of the characteristic function [2]. Mathematically, this relationship is defined as:

$$\mathbb{E}[X] = -i\frac{d\varphi_X}{dt}(0) = ie^{i\alpha t-\beta|t|}\frac{d(i\alpha t-\beta|t|)}{dt}\Bigg|_{t=0}$$

However, this requires that we evaluate the derivative of the $|t|$ at $t = 0$, which is undefined. Hence, we can conclude that the mean of the Cauchy distribution is undefined.

This result shows us that unlike many other distributions, where the mean is at least a consistent estimator, the Cauchy distribution cannot rely on such a property. The undefined nature of the mean means we cannot apply the Central Limit Theorem [3], which would otherwise guarantee convergent behaviour for a sufficiently large sample set. We further show this result through simulations for varying sample sizes, as shown in Figure 2.2.

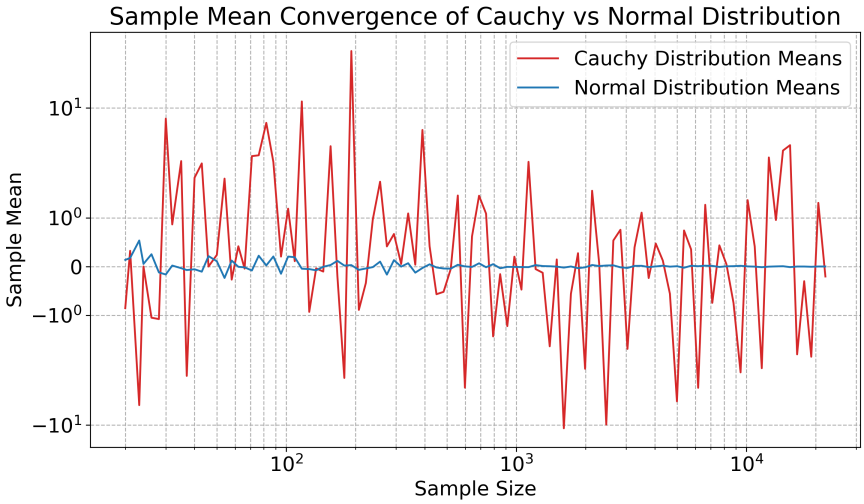


Figure 2.2: *The figure captures the (non-rolling) mean of samples of both the standard Cauchy and Normal distributions. The former's mean does not converge, regardless of the sample size, while the latter stabilizes at $\mu = 0$ after the initial stochasticity.*

2.4 Part (iv)

First of all, the prior we select should ideally be separable for simplicity – $\pi_{\alpha,\beta}(\alpha, \beta) = \pi_\alpha(\alpha)\pi_\beta(\beta)$.
The prior for α should be *translation-invariant*, as the problem depends on the horizontal *distance*, rather than the location. A suitable selection is a uniform prior:

$$\pi_\alpha = \frac{\mathbb{1}_{(\alpha_{min},\alpha_{max})}}{\alpha_{max} - \alpha_{min}}$$

The prior for β does not have any obvious properties we would like to retain, so we choose a relatively uninformed prior. We again choose the uniform distribution:

$$\pi_\beta = \frac{\mathbb{1}_{(\beta_{min},\beta_{max})}}{\beta_{max} - \beta_{min}}$$

This yields the final prior:

$$\pi(\alpha,\beta) = \frac{\mathbb{1}_{\alpha\in(\alpha_{min},\alpha_{max})}\mathbb{1}_{\beta\in(\beta_{min},\beta_{max})}}{(\alpha_{max}-\alpha_{min})(\beta_{max}-\beta_{min})}$$

With this in mind, we select appropriate bounds, namely $\alpha_{min} = -20$, $\alpha_{max} = 20$, $\beta_{min} = 0$ (the lighthouse must be above ground), $\beta_{max} = 50$.

2.5 Part (v)

Throughout the course, many methods of sampling were explored, and as so we found it necessary to compare and contrast different approaches. We explored 3 different sampling methods, one of which was implemented by hand to demonstrate understanding of the key algorithms. The different methods were also used to compare and contrast the quality of the produced samples. Convergence was evaluated by examining the samples and using different metrics to establish independence.

2.5.1 Sampling

The distribution we would like to sample from is the posterior $P(\alpha,\beta|\{x_k\})$. It is given by:

$$P(\alpha,\beta|\{x_k\}) = \frac{\mathcal{L}_x(\{x_k\}|\alpha,\beta)\pi(\alpha,\beta)}{\int_0^\infty\int_{-\infty}^\infty \mathcal{L}_x(\{x_k\}|\alpha,\beta)\pi(\alpha,\beta)\ d\alpha\ d\beta} \tag{2.6}$$

The evidence is computationally expensive to compute, so instead we use the unnormalized log-posterior as our target distribution.

$$\log P(\alpha,\beta|\{x_k\}) = \log(\pi(\alpha,\beta)) + \sum_i \log(\mathcal{L}_x(x_i|\alpha,\beta)) + const \tag{2.7}$$

The three sampling algorithms we will explore are the Metropolis-Hastings algorithm [4], Ensemble Sampling (as implemented in the *emcee* package[5]) and Nested Sampling (as implemented in the *nessai* package[6, 7, 8]).

1. **Metropolis-Hastings (MH)** – the MH algorithm is a Markov chain Monte Carlo (MCMC) method for sampling a distribution. It utilises a proposal distribution Q that produces moves in the sample space, which can be accepted or rejected depending on the target distribution. For this part, we used a multivariate normal distribution $\mathcal{N}(\mu,\mathbf{\Sigma})$ with μ being the previous point in the chain. We elect $\mathbf{\Sigma} = \mathbf{I_2}$ for simplicity. The pseudocode for the algorithm can be found in the Appendix.

2. **Ensemble Sampling** - another MCMC technique that makes use of an *ensemble* of separate Markov chains running in parallel called *walkers*. The key idea is to update the state of each walker based on the current state of all others. The *emcee* implementation makes progress using the so-called *stretch move*, which is described in the original paper [9]. Informally, the idea is to move in a direction opposite to one of the other walkers, essentially "stretching" the distance between them.

3. **Nested Sampling** – a key part of this project was to utilise a method that is not dependent on MCMC mehtods. The Nested Sampling algorithm, as originally detailed by John Skilling [10], simplifies finding the evidence of the data alongside creating a sample. While the algorithm is quite simplistic, with the pseudocode found in the Appendix, it requires a way to sample from a bounded prior $\pi(x|\mathcal{L}_x > L^*)$. The approach *nessai* takes

is to model this distribution using a variational method called **normalizing flows** [11]. The model is trained using the already existing samples, with new ones produced by inverting a transformation on a smaller latent representation of the data.

Both MCMC methods are prone to having correlated points in the chain, which would violate the independence requirement for our sample, as well as starting from a place in the chain that is not representative of the sample. To account for that, we initially clean the chain by factoring in the **burn-in** time, measured as a number of samples D (which we usually do not observe but we still conservatively remove 1–5% of the samples). Furthermore, we measure the **integrated autocorrelation time**, as described in a set of lecture notes by Alan Sokal[12]. Combining the measured times from all parameters $\{\hat{\tau}_\theta\}$ is done as so:

$$\hat{\tau} = 2 * \max_{\theta \in \{\alpha, \beta\}} (\hat{\tau}_\theta) - 1$$

Finally, we take the samples $S = \{x_D, x_{D+\hat{\tau}}, x_{D+2\hat{\tau}}, ...\}$ as our final set.

2.5.2 Results

After pre-processing the data, we summarize the results of each model visually and numerically. No special parameter finetuning was required to achieve satisfactory results. In Figure 2.3 we can observe the joint distribution and the marginals are both well-behaved. The smooth nature of the distributions means that it is also more likely that our samples have explored the entire sample space. The results we quote for each method are detailed in Table 2.1.

> **Remark:** Please note that the results related to the Nessai Nested sampler might **not** be completely reproducible using the accompanying code. The reason is that the algorithm only provides control of randomness for the initialization of the live points, but not for the later sampling. However, the results should nonetheless be qualitatively similar, meaning any conclusions we make should be valid upon repeated runs.

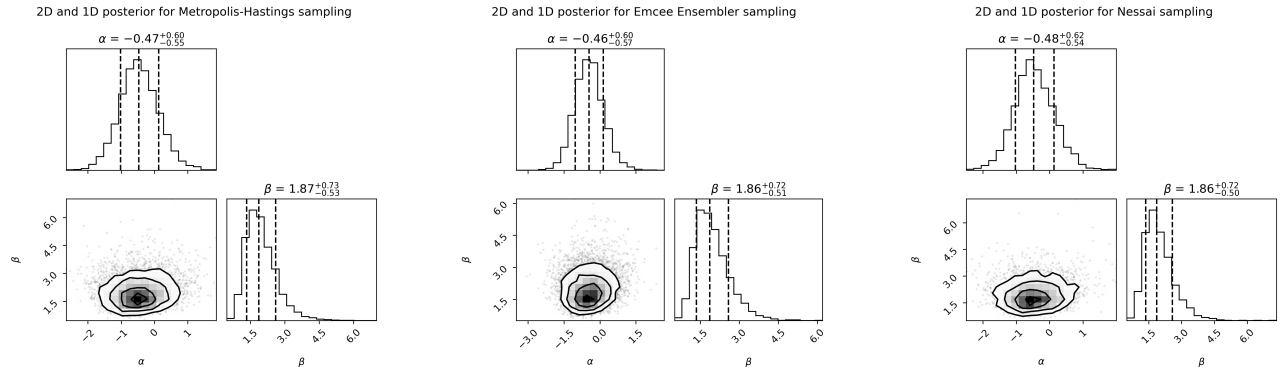


Figure 2.3: *Corner plots for each of the 3 methods. The distributions, while scaled differently, are quite similar, and the histograms are nearly identical, showing a good agreement. The parameters are given as the median with the distances to the 16- and 84-percent quantiles.*

As we can see, both from the visual representation, as well as the numerical results, the three methods show reasonable agreement, meaning any result we report from the selected ones will be valid.

Method	α	β
Metropolis-Hastings	-0.46 ± 0.59	1.98 ± 0.67
Ensemble Sampling	-0.45 ± 0.61	1.97 ± 0.66
Nested Sampling	-0.45 ± 0.61	1.98 ± 0.68

Table 2.1: *The results from each of the 3 methods, in the form mean ± std.*

2.5.3 Convergence and stability

To determine the state of convergence, we use both visual inspection methods, as well as different metrics to determine the viability of the sample. We can examine the trace plots in Figure 2.4 of the parameters to determine whether the parameters are stuck in any region unexpectedly. Furthermore, we observe the autocorrelation plots in Figure 2.5 of the chain so we can confirm whether we have correctly isolated the samples. Through both sets of plots, we can conclude that our cleaning procedure was valid and that a majority of the space of interest was explored. As our diagnostic measures, we measure the effective sample size, i.e.

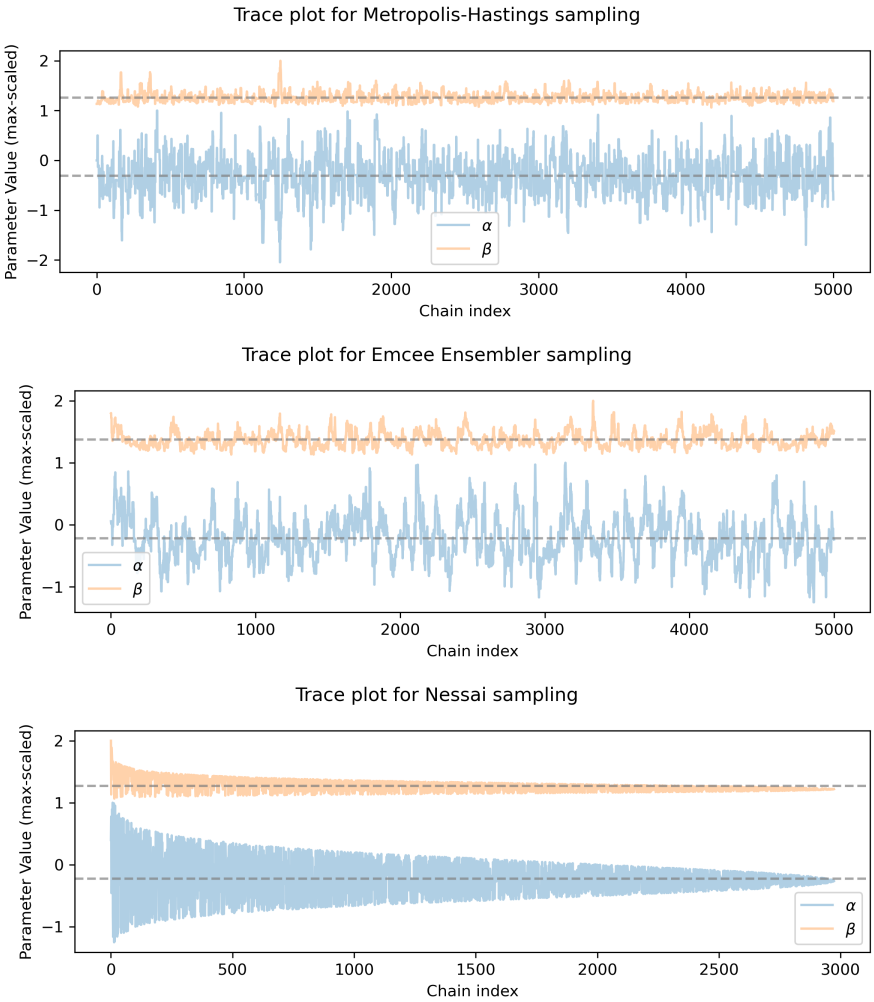


Figure 2.4: *The trace plots were normalized by the maximum traversed value and separated for visualization purposes. The dashed lines represent the mean of each feature. The results were limited to 5000 entries to ensure visibility of the traversal, showing that the chain is not stuck in any region. The Nested sampling trace plot shows how the likelihood condition squishes the distributions.*

how many of the generated points we finally accept, as well as the Gelman-Rubin statistic [13]

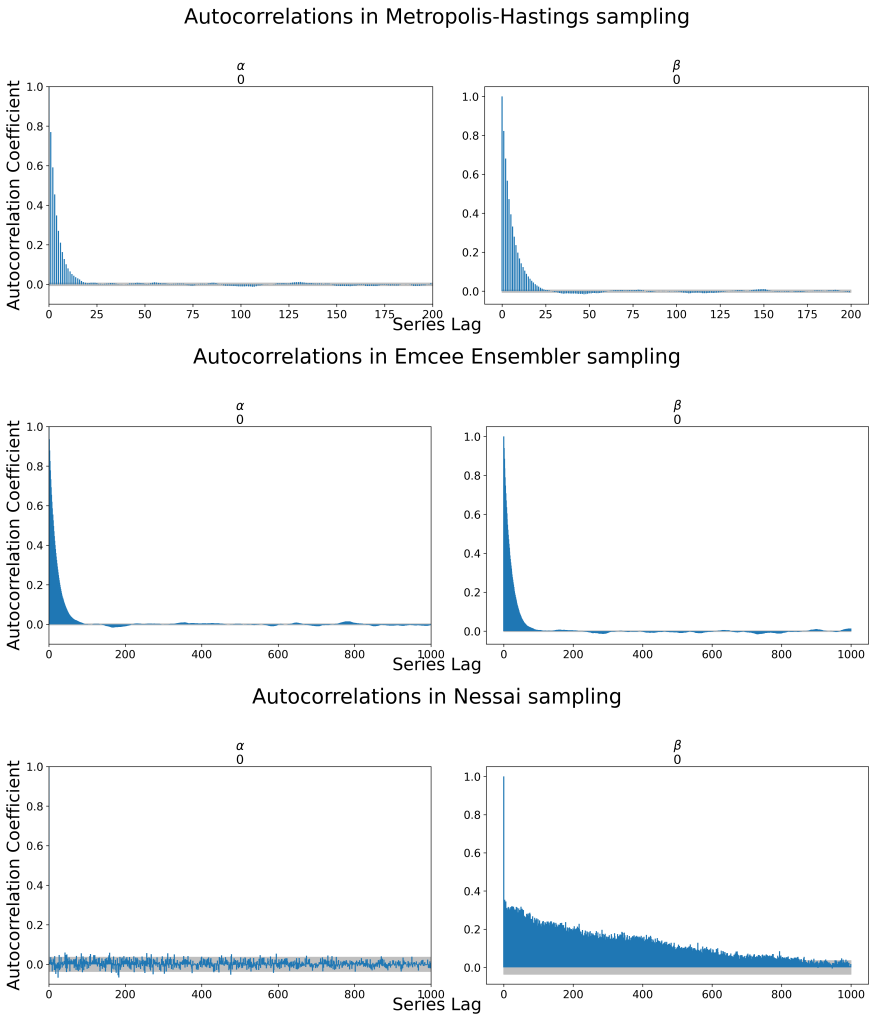


Figure 2.5: *Autocorrelation measured for different lags. Note that all autocorrelation essentially disappears at a lag of 20 for MH and 100 for ES. These numbers perfectly correspond to the efficiency reported in 2.2 - namely the rates of around 1/20 and 1/100 respectively.*

for measuring chain quality of multiple chains. The latter compares the between-chain variance with the in-chain variance, such that a metric closer to 1 shows a better convergence rate. In summary, satisfying convergence results were obtained for all 3 algorithms, as detailed in Table 2.2.

Remark: The Gelman-Rubin statistic for a set of chains $C_1, C_2, ...C_N$ and corresponding samples $C_i = x_{i,1}, x_{i,2}, ...x_{i,L}$ is given by:

$$R = \frac{\frac{L-1}{L}W + B}{W} \tag{2.8}$$

$$W = \frac{1}{N(L-1)} \sum_{i=1}^{N} \sum_{j=1}^{L} (x_{ij} - \bar{x}_i)^2, B = \frac{1}{N-1} \sum_{i=1} (\bar{x}_i - \bar{x})^2$$

Here $\bar{x}_i$ refers to the mean value of chain C_i and $\bar{x}$ - to the global mean. We further assume that each chain is of the same length. This is not a problem, as we can always split a chain into multiple ones, which produces a valid set due to the independence assumption in Markov chains ($P(x_n|x_{n-1}, x_{n-2}, ..., x_1) = P(x_n|x_{n-1})$. That is in fact what we do in the case of the long Metropolis-Hastings chain we generate.

It is further notable that this statistics does not make sense for any non-MCMC methods, such as Nested Sampling, which is why it is omitted from the report.

Method	**Initial number of samples**	**Final Number of samples**	**Accepted fraction**	**GR Statistic** α / β
Metropolis-Hastings	100,000	5,264	5.26%	1.004 / 1.004
Ensemble Sampling	500,000	6,667	1.33%	1.001 / 1.001
Nested Sampling	2,974	2,974	100%	1.34 / 1.307

Table 2.2: *Convergence and pre-processing information about the different methods. The Gelman-Rubin statistics for Metrolopolis-Hastings and Ensemble Sampling show good convergence, with a value very close to 1. The Nested Sampling should be disregarded, because the samples are not drawn in the form of chains and neighbouring samples are highly correlated due to the nature of the algorithm. It is notable that even though the ensemble sampler is "less effective", it is more efficient in generating a larger number of samples*

In addition to checking each sample set separately, we check if the samples converge to the same distribution. First of all, in the case of nested sampling, most aforementioned statistics are infeasible, which requires us to use some form of external validation. This form of agreement also makes it more likely that all sampling methods are representative given positive test results. As a univariate measurement, we utilise the Kolmogorov-Smirnov test, applied pairwise between the algorithms on each parameter sample. While this test does not take the joint probability into account, it gives us a statistical significance figure of whether the marginal distributions are the same. As expected, the test confirms our algorithms have converged to the same marginals, with results detailed in Table 2.3. Namely, because none of the p-values is below 5%, it is unlikely that we can reject that they are from the same distribution.

	α			β		
Statistic/p-value	**MH**	**ES**	**NS**	**MH**	**ES**	**NS**
Metropolis-Hastings (MH)	N/A	0.43	0.84	N/A	0.29	0.83
Ensemble Sampling (ES)	0.016	N/A	0.67	0.018	N/A	0.39
Nested Sampling (NS)	0.014	0.016	N/A	0.014	0.02	N/A

Table 2.3: *The results for the Kolmogorov-Smirnov test for each parameter, containing both the value of the statistic, as well as the corresponding p-value. The underlying null hypothesis is that both of the compared samples stem from the same distributions.*

As further confirmation which takes into account the multivariate nature of the problem, we utilise the Kullback-Leibler divergence (KLD) to measure the distances between distributions.

Definition 2.5.1 Monte Carlo estimation of KLD: The Kullback-Leibler divergence is given by:

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} P(x) \log(\frac{P(x)}{Q(x)}) dx$$

This integral seems infeasible to compute, however, we can simplify the calculation by using Monte Carlo integration:

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} P(x) \log(\frac{P(x)}{Q(x)}) dx = \mathop{\mathbb{E}}_{P(x)}[\log(\frac{P(x)}{Q(x)})] \approx \frac{1}{N} \sum_{x \in S_1} \log(\frac{P(x)}{Q(x)})$$

Here S_1 is the sample representing $P(x)$ from one of our algorithms.

Now we only need to approximate $P(x)$ and $Q(x)$ for the different samples, which we can achieve using *Kernel Density Estimation* [14] (with a Gaussian kernel). Because this distance metric is asymmetric, we also introduce the symmetric KLD measure:

$$KL_{sym} = \frac{1}{2}(D_{KL}(P||Q) + D_{KL}(Q||P))$$

The associated errors from the Monte Carlo integration were also taken into account using the estimated variance given by the Central Limit Theorem: $\sigma_f^2 = \frac{\sum_{i=1}(f(x_i)-\hat{f})^2}{N}$, where $\hat{f}$ is the estimate of the function. As we can see from the summary in Table 2.4, the distributions are confirmed to be very similar, showing further evidence for convergence.

Remark: The KLD between the sample and the true posterior was also considered for evaluation. The metric, however, requires us to normalize the posterior distribution, i.e. requires us to find the evidence, which is computationally infeasible. While the computation with the unnormalized posterior can be useful for comparing the quality of samples between the 3 algorithms, we believe the aforementioned metrics already achieve that.

	Asymmetric		
	MH	**ES**	**NS**
Metropolis-Hastings (MH)	0	0.030 ± 0.005	0.036 ± 0.005
Ensemble Sampling (ES)	0.031 ± 0.005	0	0.043 ± 0.006
Nested Sampling (NS)	0.041 ± 0.009	0.047 ± 0.010	0
	Symmetric		
	MH	**ES**	**NS**
Metropolis-Hastings (MH)	0	0.030 ± 0.004	0.038 ± 0.005
Ensemble Sampling (ES)	0.030 ± 0.004	0	0.045 ± 0.006
Nested Sampling (NS)	0.038 ± 0.005	0.045 ± 0.006	0

Table 2.4: *The results for the Kullback-Leibler divergence estimation for each pair of models. Both the asymmetric (top) and symmetric (bottom) versions of the metric are given, both showing a small KLD. This shows the distributions are likely very similar.*

Chapter 3

Parameter Estimation using the Light Intensity

3.1 Part (vi)

We add another separable component to our prior for I_0. The intensity of light should be a scale-invariant parameter because the rate of photons arriving at a given area should increase by the same factor as the source intensity. Therefore, a suitable prior would be Jeffrey's prior, or the log-uniform distribution, given by:

$$\pi_{I_0}(I_0) = \frac{\mathbb{1}_{(I_{min}, I_{max})}}{I_0 \log(\frac{I_{max}}{I_{min}})}$$

Here we elect $I_{min} = 10^{-2}$, $I_{max} = 10^3$ to cover a wide range of values. The final prior takes the form:

$$\pi(\alpha, \beta, I_0) = \frac{\mathbb{1}_{\alpha \in (\alpha_{min}, \alpha_{max})} \mathbb{1}_{\beta \in (\beta_{min}, \beta_{max})} \mathbb{1}_{I_0 \in (I_{min}, I_{max})}}{I_0(\log(I_{max}) - \log(I_{min}))(\alpha_{max} - \alpha_{min})(\beta_{max} - \beta_{min})}$$

3.2 Part (vii)

Similar to the previous inference task, we need to first define a posterior:

$$P(\alpha, \beta, I_0, \|\{x_k\}, \{\log(I_k)\}) = \frac{\mathcal{L}_{x,I}(\{x_k\}, \{\log(I_k)\}|\alpha, \beta, I_0)\pi(\alpha, \beta, I_0)}{\int_0^\infty \int_0^\infty \int_{-\infty}^\infty \mathcal{L}_{x,I}(\{x_k\}, \{\log(I_k)\}|\alpha, \beta, I_0)\pi(\alpha, \beta, I_0)\ d\alpha\ d\beta\ dI_0} \tag{3.1}$$

Here $\mathcal{L}_{x,I}(\{x_k\}, \{\log(I_k)\}|\alpha, \beta, I_0)$ can be rewritten as:

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

$$= \prod_{i=1} \frac{\beta}{\beta^2 + (x_i - \alpha)^2}$$

We can the perform a similar transformation as before to use the log-posterior instead for numerical stability and easier computation. Similar to Part (v), we perform identical experiments, with small changes that will be further elaborated below.

3.2.1 Sampling

First of all, when running the Metropolis-Hastings algorithm, it is noticeable that the identity matrix $\mathbf{I_3}$ will no longer perform well for efficiently exploring the parameter space. This was confirmed by obtaining a final sample size of less than 100/100,000 with the default setting. Instead, we correct for this effect by choosing a covariance matrix of the form $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \sigma_I^2 \end{pmatrix}$. After experimentation, a satisfactory value of $\sigma_I^2 = 10$ was reached.

The rest of the algorithms did not need to be finetuned and we can continue with the analysis as previously.

3.2.2 Results

After repeating the experiments, we report the findings as previously detailed. The marginal and joint distributions can be seen in Figure 3.1, while the numerical summaries are given in Table 3.1. A curious observation is that the standard deviation associated with the α parameter has significantly decreased, which will be discussed in Part (viii). The results we obtained are quite different to what we got in the previous section. That shows how important the intensity information is to our prediction. It is also notable that even though our choice of prior for I_0

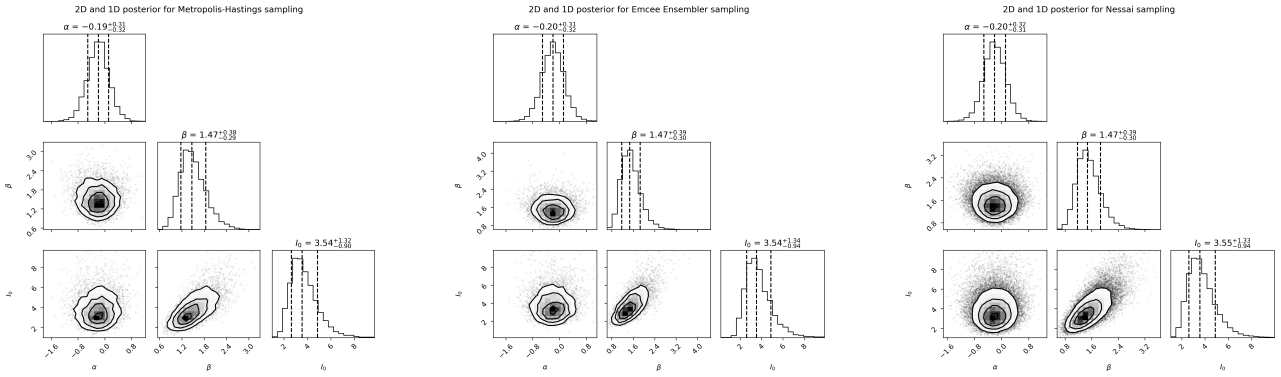


Figure 3.1: *Corner plots for each of the 3 methods, which again show a good agreement. The parameters are given as the median with the distances to the 16- and 84-percent quantiles.*

Method	α	β	I_0
Metropolis-Hastings	-0.20 ± 0.33	1.51 ± 0.35	3.75 ± 1.19
Ensemble Sampling	-0.20 ± 0.33	1.52 ± 0.37	3.75 ± 1.25
Nested Sampling height	-0.20 ± 0.33	1.52 ± 0.36	3.75 ± 1.22

Table 3.1: *The results from each of the 3 methods, in the form mean ± std.*

generally skews the posterior distribution towards smaller values, this bias has been overcome when comparing the resulting sample to the original prior. This demonstration can be found in Figure 3.2, where we show this relationship for the Ensemble Sampler. The choice was made arbitrarily, as the equivalent plots are visually similar for both of the other methods, as can be seen in the Appendix.

3.2.3 Convergence and stability

We repeat the same experiments as detailed in Part (v). As before, we achieve satisfactory convergence results, reflected both in the trace and correlation plots shown in Figures 3.3 and

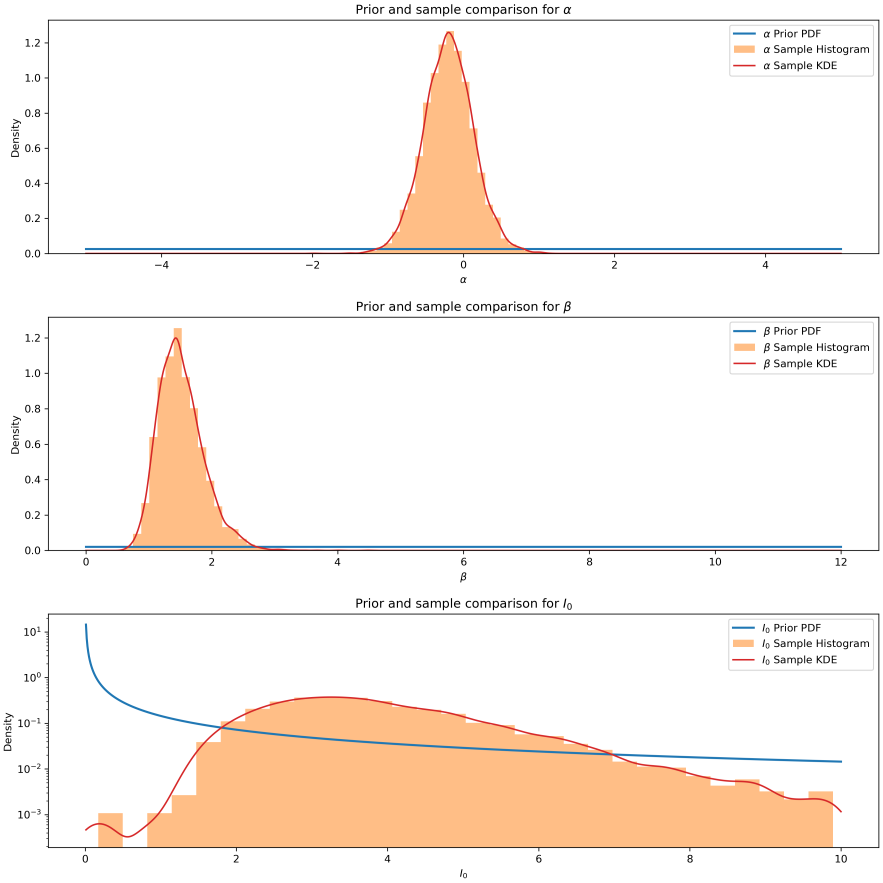


Figure 3.2: *The posterior distribution estimation. For α and β, the posterior is much more concentrated around the likely value, in contrast to the priors, which have wider ranges. For I_0, despite the high initial bias for smaller values, the posterior clearly discards said bias. A log scale was used for this parameter's y-axis for better clarity.*

3.4, as well as the summary statistics given in Table 3.2.

Method	**Initial number of samples**	**Final Number of samples**	**Accepted fraction**	**GR Statistic** α / β / I_0
Metropolis-Hastings	500,000	5,618	1.12%	1.001 / 1.001 / 1.001
Ensemble Sampling	500,000	5,748	1.15%	1.002 / 1.002 / 1.002
Nested Sampling	33,601	33,601	100%	1.19 / 1.17 / 1.17

Table 3.2: *Convergence and pre-processing information about the different methods. We observe the same conclusions as we did in part (v) - with good convergence for the MCMC methods. There is, unfortunately, a decrease in efficiency due to the higher complexity of the model.*

However, in this experiment, we observe higher values for the KLD measurements, as seen in Table 3.3. This should, nevertheless, not be a point of concern for convergence. Due to us using the KDE estimation for a smooth posterior, the now 3-dimensional nature of our data makes it very likely that some regions are highly undersampled, giving us significant differences in distributions. If given higher computational power, running the algorithms with 10 times more samples will likely result in a much smaller KLD. This is further backed up by the Kolmogorov-Smirnov test results in Table 3.4 that show a high probability that the marginal distributions are the same.

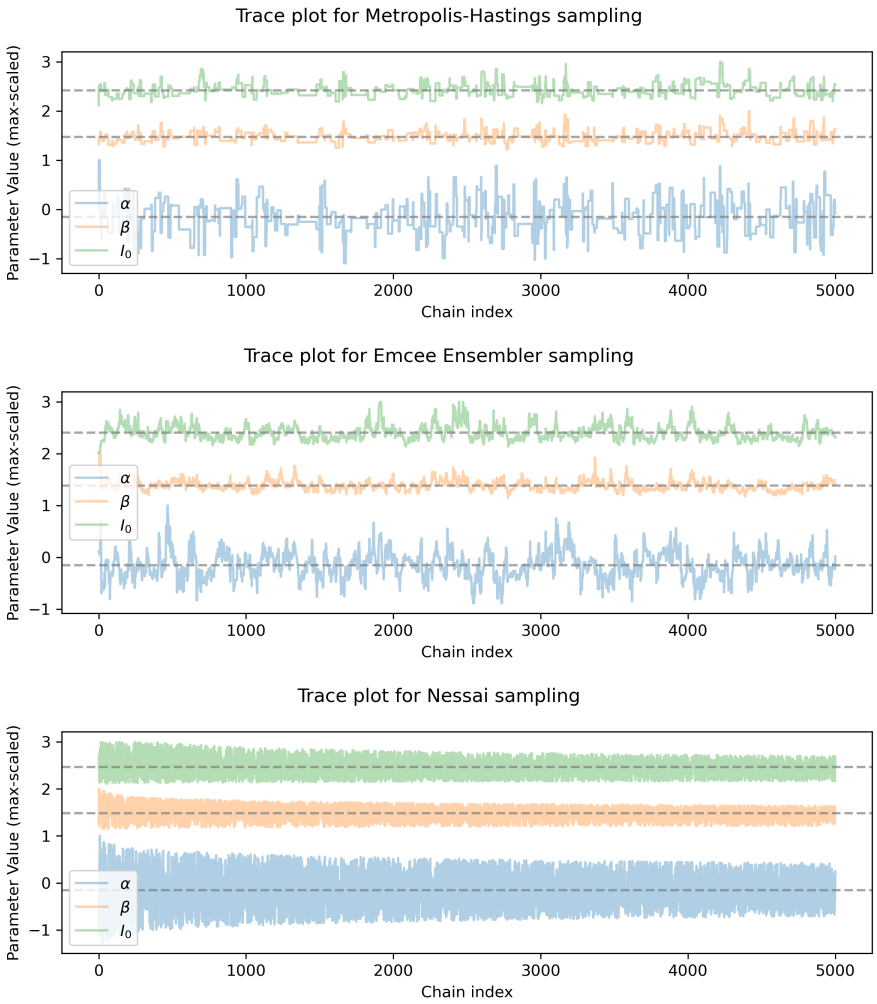


Figure 3.3: *Similarly to part (v), the samples were correspondingly scaled and shifted for visibility.*

	Asymmetric		
	MH	**ES**	**NS**
Metropolis-Hastings (MH)	0	0.11 ± 0.006	0.083 ± 0.006
Ensemble Sampling (ES)	0.12 ± 0.011	0	0.09 ± 0.010
Nested Sampling (NS)	0.07 ± 0.003	0.07 ± 0.002	0
	Symmetric		
	MH	**ES**	**NS**
Metropolis-Hastings (MH)	0	0.11 ± 0.006	0.08 ± 0.003
Ensemble Sampling (ES)	0.11 ± 0.006	0	0.08 ± 0.005
Nested Sampling (NS)	0.08 ± 0.003	0.08 ± 0.005	0

Table 3.3: *The results for the Kullback-Leibler divergence estimation for each pair of models. Both the asymmetric (top) and symmetric (bottom) versions of the metric are given.*

3.3 Part (viii)

The results we obtain for the parameter α differ significantly between the two approaches. That said, this does not necessarily mean we have a worse estimation, as incorporating additional information represents the physics better. This is further complemented by a lower standard

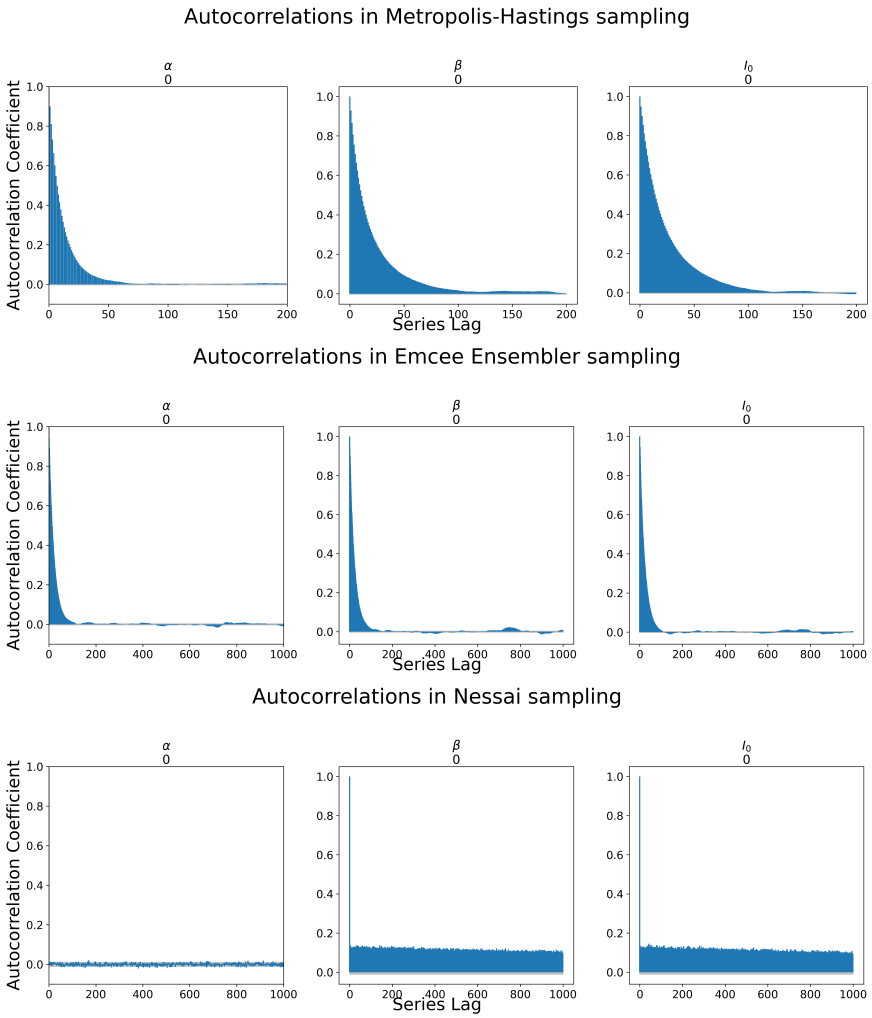


Figure 3.4: *Autocorrelation measured for different lags. Analogous to the comparison in Part (v), the points at which the correlation becomes low enough perfectly match the rate at which points are accepted for the MCMC methods.*

	α			β			$\mathbf{I_0}$		
Statistic/p-value	**MH**	**ES**	**NS**	**MH**	**ES**	**NS**	**MH**	**ES**	**NS**
MH	N/A	0.73	0.08	N/A	0.35	0.07	N/A	0.35	0.36
ES	0.013	N/A	0.45	0.017	N/A	0.64	0.017	N/A	0.45
NS	0.018	0.012	N/A	0.019	0.011	N/A	0.013	0.012	N/A

Table 3.4: *The results for the Kolmogorov-Smirnov test for each parameter, containing both the value of the statistic, as well as the corresponding p-value. Here the abbreviations stand for Metropolis-Hastings (HS), Ensemble Sampling (ES) and Nested Sampling (NS).*

deviation, which shows our uncertainty has significantly decreased.

In any case, both values of α include each other in their 68% confidence interval, so they do not show any concerning disagreement. Hence, we can conclude that the second model likely describes the data better and the estimate would be more reliable.

Chapter 4

Conclusion

In this report, we described a thorough framework for Bayesian inference in the context of the Lighthouse problem. We demonstrated several techniques for assessing convergence and commented on how they complement and confirm one another. Furthermore, through the use of additional information in the form of light intensity, we validated the effectiveness of our framework to perform well without much finetuning. Finally, we demonstrated an improvement in the predictions which better capture the reality of our small dataset.

Bibliography

[1] D. S. Sivia and J. Skilling, *Data Analysis - A Bayesian Tutorial*. Oxford Science Publications, Oxford University Press, 2nd ed., 2006.

[2] J. Jacod and P. Protter, *Properties of Characteristic Functions*, pp. 111–116. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004.

[3] S. G. Kwak and J. H. Kim, "Central limit theorem: the cornerstone of modern statistics," *Korean J Anesthesiol*, vol. 70, pp. 144–156, Feb. 2017.

[4] C. P. Robert, "The metropolis-hastings algorithm," 2016.

[5] D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman, "¡tt¿emcee¡/tt¿: The mcmc hammer," *Publications of the Astronomical Society of the Pacific*, vol. 125, p. 306–312, Mar. 2013.

[6] M. J. Williams, "nessai: Nested sampling with artificial intelligence," Feb. 2021.

[7] M. J. Williams, J. Veitch, and C. Messenger, "Nested sampling with normalizing flows for gravitational-wave inference," *Phys. Rev. D*, vol. 103, no. 10, p. 103006, 2021.

[8] M. J. Williams, J. Veitch, and C. Messenger, "Importance nested sampling with normalising flows," 2 2023.

[9] J. Goodman and J. Weare, "Ensemble samplers with affine invariance," *Communications in Applied Mathematics and Computational Science*, vol. 5, pp. 65–80, Jan. 2010.

[10] J. Skilling, "Nested sampling for general Bayesian computation," *Bayesian Analysis*, vol. 1, no. 4, pp. 833 – 859, 2006.

[11] D. J. Rezende and S. Mohamed, "Variational inference with normalizing flows," 2016.

[12] A. D. Sokal, "Monte carlo methods in statistical mechanics: Foundations and new algorithms note to the reader," 1996.

[13] A. Gelman and D. B. Rubin, "Inference from Iterative Simulation Using Multiple Sequences," *Statistical Science*, vol. 7, no. 4, pp. 457 – 472, 1992.

[14] S. J. Sheather, "Density Estimation," *Statistical Science*, vol. 19, no. 4, pp. 588 – 597, 2004.

Appendix

Pseudocodes

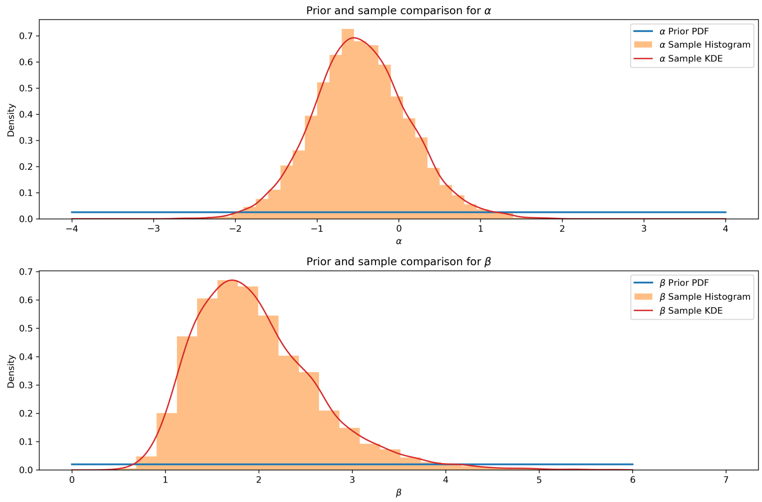
Metropolis-Hastings

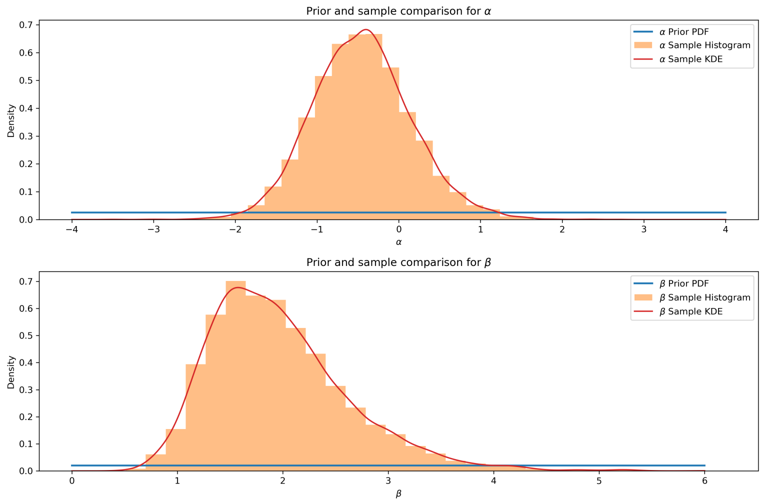
Initialize $x^{(0)}$ arbitrarily
for $i = 1$ to N iterations **do**
  Generate a candidate x' from a proposal distribution $q(x'|x^{(i-1)})$
  Calculate acceptance ratio $\alpha = \min\left(1, \frac{p(x')q(x^{(i-1)}|x')}{p(x^{(i-1)})q(x'|x^{(i-1)})}\right)$
  Draw $u \sim \text{Uniform}(0, 1)$
  if $u \leq \alpha$ **then**
    Accept the candidate: $x^{(i)} = x'$
  else
    Reject the candidate: $x^{(i)} = x^{(i-1)}$
  end if
end for=0

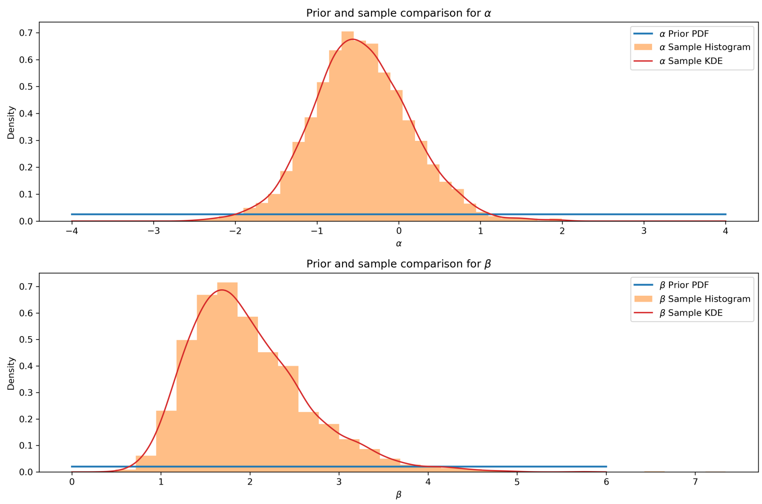
Nested Sampling

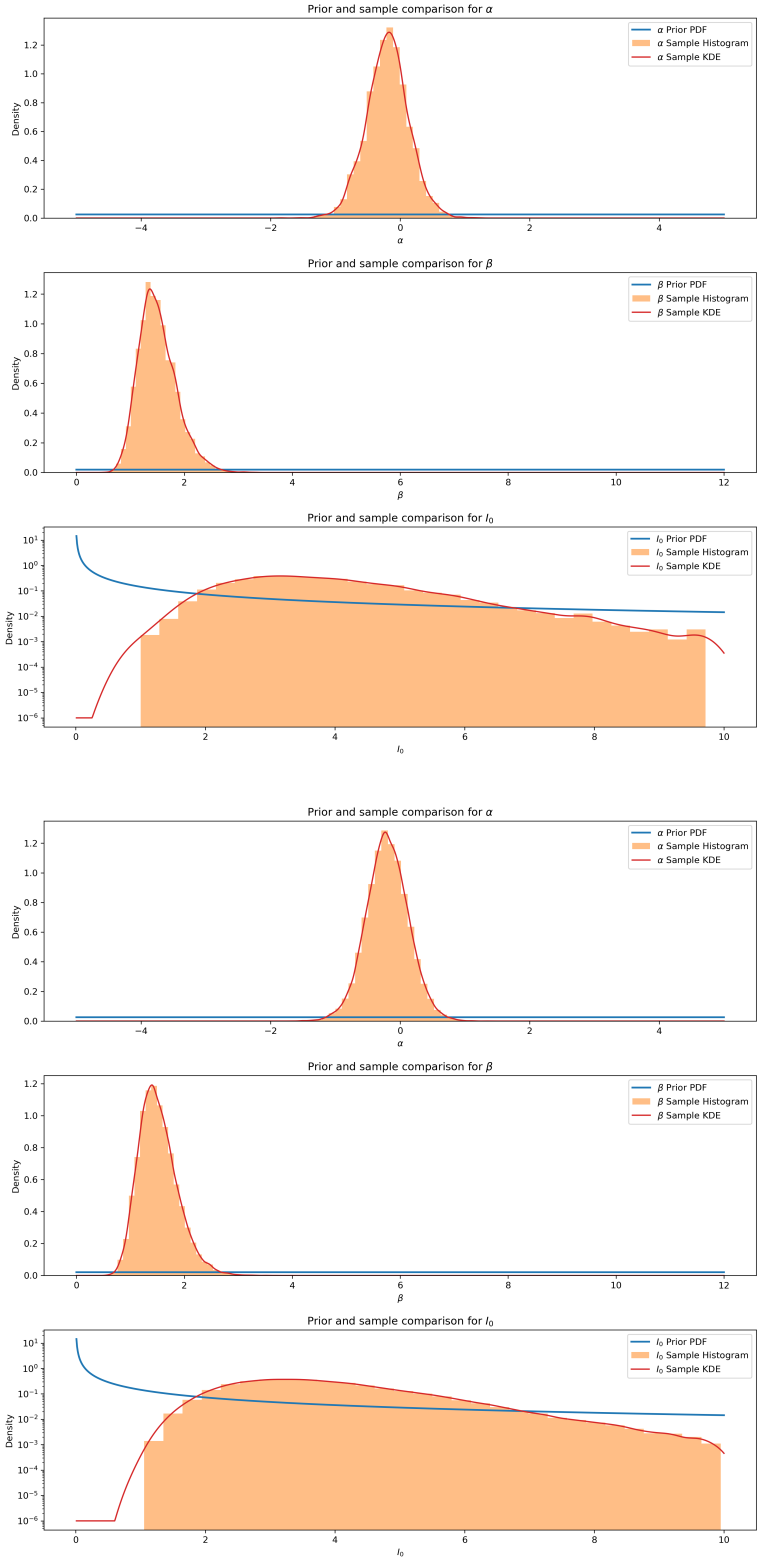
Initialize a set of N live points sampled from prior π
Evaluate the likelihoods $\{\mathcal{L}_k\}$ for each live point
Initialize $X = 1$, $\mathcal{L}_{\max} = 0$, and an empty list of samples
while stopping criterion not met **do**
  Find, $k = argmin_k \mathcal{L}_k$ and let $\mathcal{L}_{\min} = \mathcal{L}_k$
  Select a new point with likelihood $> \mathcal{L}_{\min}$ by randomly sampling from $\pi(x|\mathcal{L}_x > \mathcal{L}_{\min})$
  Update the evidence and other quantities of interest
  Replace x_k with the new point
  Shrink the prior volume X according to the rule $X \leftarrow X \exp(-1/N)$
end while
Return the estimated evidence, posterior samples, and other quantities =0

Posterior to prior comparisons









Ackowledgement for the use of generative tools

We acknowledge the utilization of ChatGPT, developed by OpenAI, for purposes related to generating documentation. ChatGPT was employed as a tool to assist in creating Doxygen-complacent documentation for describing the source files, contributing to the completion of the project's objectives. No other generative tools were used, and ChatGPT was not used for any purpose outside of the aforementioned tasks.
Reference: OpenAI. ChatGPT. 2023, https://openai.com/chatgpt.