S2 Coursework

Cailley Factor

March 2024

(i) The lighthouse problem concerns a lighthouse, which is at position α along the coastline and β out to sea. The lighthouse rotates and emits flashes at uniformly-distributed angles θ. An array of detectors spread along the coastline record the locations x_k, where N flashes are received at the respective angle θ_k ($-\pi/2 < \theta < \pi/2$) for the kth flash. The trigonometric relationship between α, β, θ, and x for the kth flash is:

$$\beta tan\theta_k = x_k - \alpha \tag{1}$$

(ii) Using the fact that θ is uniformly distributed, the probability density function (PDF) of the kth angle θ_k is:

$$prob(\theta_k|\alpha, \beta, I) = \frac{1}{\pi} \tag{2}$$

In order to write the likelihood in terms of the location for a single flash x_k, this requires a change of variables. Differentiating both sides of equation 1 yields:

$$\beta sec^2\theta_k \frac{d\theta_k}{dx_k} = 1 \tag{3}$$

Plugging in the identity $tan^2\theta + 1 \equiv sec^2\theta$ and utilising equation 1 again results in the following expression.

$$\frac{d\theta_k}{dx_k} = (\beta(1 + tan^2\theta_k))^{-1} = (\beta(1 + \left(\frac{x_k - \alpha}{\beta}\right)^2))^{-1} \tag{4}$$

Utilising the equation for change of variables, results in the following:

$$prob(x_k|\alpha, \beta) = prob(\theta_k|\alpha, \beta) \times \left|\frac{d\theta_k}{dx_k}\right| = \frac{\beta}{\pi(\beta^2 + (x - \alpha)^2)} \tag{5}$$

(iii) Assuming that the distance out to sea β is fixed, using Bayes' theorem, the posterior PDF for the flashes can be written as:

$$prob(\alpha|\{x_k\}, \beta) \propto prob(\{x_k\}|\alpha, \beta) \times prob(\alpha|\beta) \tag{6}$$

We can assign a simple uniform PDF to the prior, such that:

$$prob(\alpha|\beta) = prob(\alpha) = \begin{cases} \frac{1}{\alpha_{max} - \alpha_{min}} & \text{for } \alpha_{min} \leq \alpha \leq \alpha_{max}, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Given that the recording of the flashes are independent events, the likelihood function is the product of the probabilities for the N independent flash recordings:

$$prob(\{x_k\}|\alpha, \beta) = \prod_{k=1}^{N} prob(x_k|\alpha, \beta) = \prod_{k=1}^{N} \frac{\beta}{\pi(\beta^2 + (x - \alpha)^2)} \tag{8}$$

We can substitute equations 7 and 8 into equation 6 and take the natural logarithm of this posterior PDF to yield:

$$ln(prob(\alpha|\{x_k\},\beta)) = c - \sum_{k=1}^{N} ln(\beta^2 + (x_k - \alpha)^2) \tag{9}$$

where c is a constant [1]. The maximum likelihood estimator for α can be derived from equation 9 by differentiating with respect to α, where α_o is the maximum likelihood estimator of α [1]. This yields:

$$\left.\frac{dL}{d\alpha}\right|_{x=\alpha_o} = \sum_{k=1}^{N} \frac{2(x_k - \alpha_o)}{\beta^2 + (x_k - \alpha_o)^2} = 0 \tag{10}$$

Equation 10 is difficult to evaluate analytically, but it can be seen that the maximum likelihood estimator of α is not the sample mean of the positions $\hat{x} = \frac{1}{N}\sum_k x_k$ [1]. The explanation behind this is that the Cauchy distribution does not satisfy a key assumption of the central limit theorem (CLT), which assumes finite variance of the distribution [1]. The variance of the Cauchy distribution σ^2 is infinite, due to the very wide wings of the distribution, and the mean is undefined [1]. The variability of the mean does not decrease with increasing numbers of measurements, i.e., taking more samples x_k does not provide a better estimator for $\bar{x}_k$ [1]. Thus, samples drawn from the Cauchy distribution do not satisfy the CLT, and the sample mean of $\hat{x}$ is accordingly not a good estimator for α. The posterior PDF should be used instead to infer the parameter estimates.

(iv) Given that no information is known about the unknown parameters α and β, suitable priors could be uniform priors to reflect a state of ignorance about the position of the flashes along the coast α and the position of the lighthouse out to sea β. The priors can be written accordingly as:

$$prob(\alpha) = \begin{cases} \frac{1}{\alpha_{max} - \alpha_{min}} & \text{for } \alpha_{min} \leq \alpha \leq \alpha_{max}, \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

$$prob(\beta) = \begin{cases} \frac{1}{\beta_{max} - \beta_{min}} & \text{for } \beta_{min} \leq \beta \leq \beta_{max}, \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

(v) The prior on α and β can be assumed to be independent and thus, the posterior distribution can be defined by multiplying the likelihood times the priors on α and β (13). Stochastic sampling evaluates the integral $E_{p(\theta)}[f(\theta)] = \int f(\theta)P(x)dx$ by generating N samples θ_n and then approximating the integral as $E_{p(\theta)}[f(\theta)] \approx \frac{1}{N}\sum_{n=1}^{N} f(\theta_n)$ [2]. This can be used to calculate the expectation of parameters as the formula for the expectation of parameters is equal to $E(\Theta) = \int \theta p(\theta) d\theta \approx \frac{1}{N}\sum_{n=1}^{N} \theta_n$, i.e., the sample mean.

$$prob(\alpha, \beta|\{x_k\}) = prob(\{x_k\}|\alpha, \beta) \times prob(\alpha) \times prob(\beta) \tag{13}$$

For numerical stability for the computational Markov chain Monte Carlo, the logarithm of the posterior (13) is calculated (14). The likelihood function for the N independent flash recordings (8) and the priors in equations 11 and 12 were substituted into equation 14, where $\alpha_{min} \leq \alpha \leq \alpha_{max}$ and $\beta_{min} \leq \beta \leq \beta_{max}$, yielding equation 15.

$$\log(prob(\alpha, \beta|\{x_k\})) = \log(prob(\{x_k\}|\alpha, \beta)) + \log(prob(\alpha)) + \log(prob(\beta)) \tag{14}$$

$$\log(prob(\alpha, \beta|\{x_k\})) = \log(\prod_{k=1}^{N} \frac{\beta}{\pi(\beta^2 + (x_k - \alpha)^2)}) + \log(\frac{1}{\alpha_{max} - \alpha_{min}}) + \log(\frac{1}{\beta_{max} - \beta_{min}}) \tag{15}$$

The log posterior distribution (15) was sampled from using the emcee package, which utilises an affine-invariant ensemble of samplers, such that the performance of the algorithm is independent of the aspect ratio of highly anisotropic distributions, i.e., the algorithm can sample from distributions regardless of how

skewed they are [3, 4, 2]. K walkers are simultaneously evolved in the ensemble [3]. Each walker X_k is updated by drawing a walker X_j from the complementary ensemble of the remaining K-1 walkers, $S_{[k]}(t)$ [3]. The overall algorithm for a single stretch move update step is:

1. for k = 1, ..., K do
2. Draw a walker X_j at random from the complementary ensemble $S_{[k]}(t)$
3. $z \leftarrow Z \sim g(z)$, defined in equation 16
4. $Y \leftarrow X_j + Z[X_k(t) - X_j]$
5. $q \leftarrow Z^{N-1} p(Y)/p(X_k(t))$
6. $r \leftarrow R \sim [0, 1]$
7. if $r \leq q$, as in equation 17
8. $X_k(t+1) \leftarrow Y$
9. else
10. $X_k(t+1) \leftarrow X_k(t)$
11. end if
12. end for [3]

Z is a random variable drawn from the distribution g(Z=z), defined in equation 16 [3]. With this g(Z=z), the proposal is symmetric and thus the probability that a proposal is accepted is defined as in equation 17 [3].

$$g(z) = \begin{cases} \frac{1}{\sqrt{z}} & \text{if } z \in \left[\frac{1}{a}, a\right] \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

$$q = min(1, Z^{N-1} \frac{p(Y)}{p(X_K(t)}) \tag{17}$$

The parameters of the prior distributions were set to be $\alpha_{min} = -100$, $\alpha_{max} = 100$, and $\beta_{min} = 0$, $\beta_{max} = 100$. This is because the flash locations are found in the range of -18.9 to 70.5 and thus a range slightly larger than this was selected as a bound for the lighthouse location along the coast, as it should fall within this range given that the light beams intersect the coast at an angle of $-\pi/2 < \theta < \pi/2$, and the same upper range was chosen for β.

The samples were initialised randomly within the parameter space in these bounds. 100 walkers were chosen and a sample size of 10,000 used and the sampler initialised and run. The flattened samples from the different walkers was used to plot a trace plot, shown in figure 1. The characteristic appearance of the plot without clear trends and periodicity demonstrates that the chain is effectively exploring the parameter space. The acceptance fraction was found to be 70.5%. The chain appeared to converge at around 100 samples based on the trace plot, however, the first 1000 samples were discarded to ensure that the chain had converged.

The variance of independent samples can be calculated as $\sigma^2 = \frac{1}{N} Var_{p(\theta)}[f(\theta)]$, such that the variance decreases proportionally to the sample size by $\frac{1}{\sqrt{N}}$ [2]. However, the samples are correlated and the variance accounting for the samples' lack of independence is given by $\sigma^2 = \frac{\tau_f}{N} Var_{p(\theta)}[f(\theta)]$, where τ_f is the integrated auto-correlation time [2]. This was calculated using an in-built method of the emcee package, and the maximum autocorrelation time for each of the parameters extracted before thinning and 'burn-in' [2]. The autocorrelation times for the parameters α and β were found to be 21.7 and 33.9 (3 s.f.). As well as discarding the first 1000 samples for the Monte Carlo chain, the chain was thinned by the maximum autocorrelation time, i.e., 33.9, as the chain was sufficiently long to enable an adequate number of samples with this thinning, resulting in 27,200 uncorrelated samples.

The 2D histogram showing the joint posterior on α and β and the 1D histograms of the marginalised posterior distributions are displayed in figure 2 in the form of a corner plot. The plots are labelled with the median and upper and lower errors supplied by the corner package [5]. To perform inference on the posterior distribution, the collection of all the simulated draws after burn-in and thinning were used to calculate the MAP estimate, mean and standard deviation of the parameters. The MAP estimate was calculated by

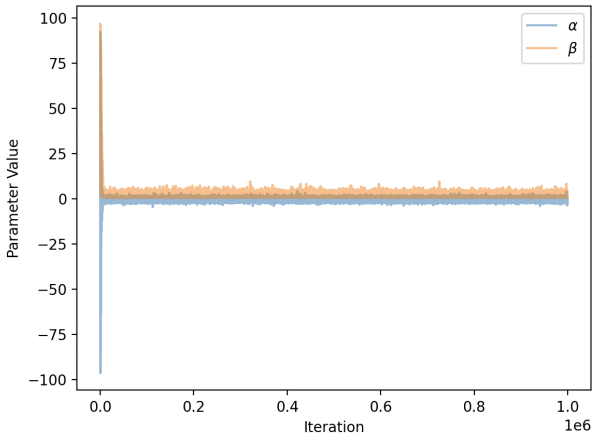

Figure 1: Trace plot of samples from the posterior distribution $P(\alpha, \beta|\{x_k\})$

determining the mode of the parameter values for the two-dimensional posterior, and was -0.510 for α and 1.533 for β (3 d.p.). The mean and standard deviation were -0.449 $\pm$ 0.601 and 1.954 $\pm$ 0.660 for α and β respectively to 3 decimal places.

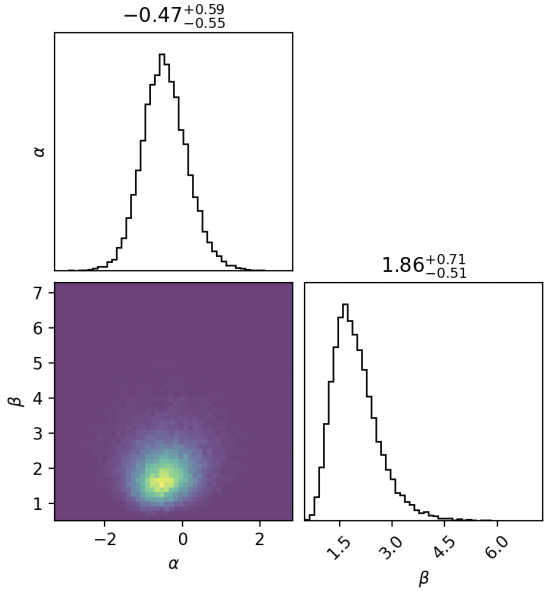

Figure 2: Corner plot of samples from the posterior distribution $P(\alpha, \beta|\{x_k\})$

The Gelman-Rubin statistic is a very useful statistic to monitor convergence and requires simulating m $\geq$ 2 sequences with burn-in [6]. Two variance components are calculated: the within-chain variance (W) with n-1 degrees of freedom and the variance between the m sequence means ($\bar{x}_i$) each based on n values of X (B/n) [6]. Let $\bar{X}_i$ denote the sample mean for chain i, $\hat{\mu}$ denote the overall mean, and s_i^2 denote the sample variance for chain i [7]. The within-chain variance is therefore $W = \frac{1}{m}\sum_{i=1}^{m} s_i^2$. B/n is calculated as $B/n = \frac{1}{m-1}\sum_{i=1}^{m}(\bar{x}_i - \hat{\mu})^2$ [7]. The Gelman-Rubin statistic $\sqrt{\hat{R}}$ is shown in equation 18 [7]. This statistic is not in-built in the emcee package and thus was hand-coded. In order to utilise the statistic, two chains were run with different randomly initialised starting parameters, as multiple chains are required. The Gelman-Rubin statistic was found to be for 1.000 α and for 1.000 β (to 4 d.p.) indicating substantial convergence as these values are very close to 1.

$$\sqrt{\hat{R}} = \sqrt{\frac{\frac{n-1}{n}W + \frac{1}{n}B}{W}} \tag{18}$$

(vi) An array of detectors measures the intensity of each flash I_k for the kth flash. The likelihood of an intensity flash follows a log-normal distribution (19) with uncertainty $\sigma = 1$, and an expectation of the log intensity $\mu = \log\left(\frac{I_0}{d^2}\right)$ where $d^2 = \beta^2 + (x-\alpha)^2$. The likelihood function can be rewritten in terms of the parameter I in equation 20. A suitable prior for I_0 is a log normal prior as intensity is a physical parameter that is scale invariant (21).

$$\mathcal{L}_I(\log I|\alpha,\beta,I_0) = \frac{\exp\left(-\frac{(\log I-\mu)^2}{2\sigma^2}\right)}{\sqrt{2\pi\sigma^2}} \tag{19}$$

$$\mathcal{L}_I(I|\alpha,\beta,I_0) = \frac{\exp\left(-\frac{(\log I-\mu)^2}{2\sigma^2}\right)}{I\sqrt{2\pi\sigma^2}} \tag{20}$$

$$prob(I_0) = \begin{cases} \frac{1}{I_0 log(I_{0max}/I_{0min})} & \text{for } I_{0min} \leq I_0 \leq I_{0max}, \\ 0 & \text{otherwise.} \end{cases} \tag{21}$$

(vii) The location and intensity measurements are independent, so the likelihood for a combined location and intensity measurement is given by equation 22 or equation 23. The combined likelihood of the location and intensity measurements for the flashes from 1, 2,..., N with location x_k intensity I_k and is given by equation 24. For calculating the posterior, Iterative Bayesian update can be used such that the posterior $prob(\alpha,\beta|x_k)$ can be used as the prior for α and β. The resulting posterior for the flashes is given by equation 25. In order to sample from the posterior using emcee, the logarithm of the posterior is computed by substituting equation 24 into equation 25 and taking the logarithm (26).

$$\mathcal{L}_{x,I}(x,\log I|\alpha,\beta,I_0) = \mathcal{L}_x(x|\alpha,\beta)\mathcal{L}_I(\log I|\alpha,\beta,I_0) \tag{22}$$

$$\mathcal{L}_{x,I}(x,I|\alpha,\beta,I_0) = \mathcal{L}_x(x|\alpha,\beta)\mathcal{L}_I(I|\alpha,\beta,I_0) \tag{23}$$

$$prob(\{x_k\},\{I_k\}|\alpha,\beta,I_0) = \prod_{k=1}^{N} prob(x_k|\alpha,\beta) \times \prod_{k=1}^{N} prob(I_k|\alpha,\beta,I_0) \tag{24}$$

$$prob(\alpha,\beta,I_0|\{x_k\},\{I_k\}) = prob(\{x_k\},\{I_k\}|\alpha,\beta,I_0) \times prob(\alpha,\beta|x_k) \times prob(I_0) \tag{25}$$

$$\begin{aligned} log(prob(\alpha,\beta,I_0|\{x_k\},\{I_k\})) = \sum_{k=1}^{N} log(prob(x_k|\alpha,\beta)) + \sum_{k=1}^{N} log(prob(I_k|\alpha,\beta,I_0)) \\ + log(prob(\alpha,\beta|x_k)) + log(prob(I_0)) \end{aligned} \tag{26}$$

To determine the bounds for the prior for I_0, a minimum bound was used as 0.001, as the intensity will be non-zero. The absolute intensity of the lighthouse will be much greater than the measured intensity of each flash by the detectors. Given that the range of recorded intensities was from 0.00140 to 7.61 (3 s.f.), a maximum bound of 100 was used as it is an order of magnitude greater than the highest recorded intensity. The same bounds for α_{min}, α_{max}, β_{min}, and β_{max} were used as in question v, for the same rationale as above.

In order to sample from the posterior, a greater number of walkers were utilised due to the higher dimensionality. 150 walkers were used for 10,000 steps. As before, the walkers were initialised randomly in the parameter space based on the parameter bounds. The trace plot for the samples is shown in figure 3. As above, based on the plot the chain appears to take around 100 samples to converge, so the first 1000 were discarded. The overall acceptance rate was slightly lower at 63.7%. The autocorrelation times

were calculated using the in-built method and were 20.3, 40.0, and 63.1 for α, β, and I_0, respectively. The maximum autocorrelation time, i.e., 63.1, was used to thin the chain, resulting in 21,300 uncorrelated samples.

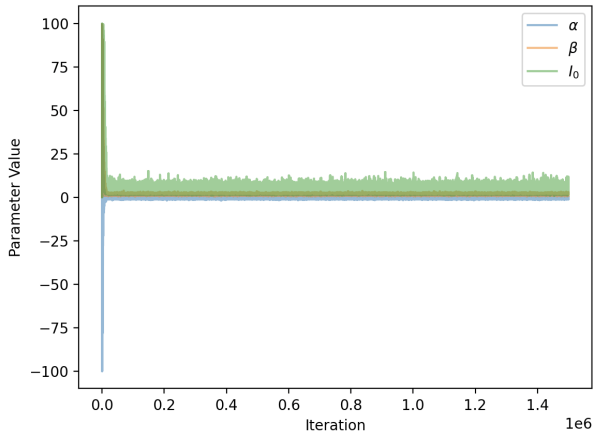

Figure 3: Trace plot of samples from the posterior distribution $P(\alpha, \beta, I_0|\{x_k\}, \{I_k\})$

The 2D histogram showing the joint posterior on α, β and I_0 and the 1D histograms of the marginalised posterior distributions are displayed in figure 4 in the form of a corner plot, as in question v. As before, the plots are labelled with the median and upper and lower errors supplied by the corner package [5]. To perform inference on the posterior distribution, the collection of all the simulated draws after burn-in and thinning were used to calculate the MAP estimate, mean and standard deviation of the parameters. As before, the MAP estimate was calculated based on the mode of the parameter from sampling the three-dimensional posterior, resulting in a MAP of -0.303 for α, 1.536 for β, and 3.360 for I_0. The mean and standard deviation were -0.286 $\pm$ 0.282, 1.566 $\pm$ 0.291, and 3.822 $\pm$ 1.119 for α, β, and I_0 to 3 decimal places, respectively. Another chain was run in order to utilise the hand-coded Gelman-Rubin method to assess convergence. The Gelman-Rubin statistic for the chains with burn-in was found to be 1.000 for α, 1.000 for β, and 1.000 for I_0 (to 4 d.p.) indicating substantial convergence as these values are very near 1.

(viii) The mean and standard deviation of α was found to be -0.449 $\pm$ 0.601 in question v, and -0.286 $\pm$ 0.282 in question vii, sampling from the two-dimensional and three-dimensional posterior distributions, respectively. Including the intensity data and sampling from the three-dimensional posterior $P(\alpha, \beta, I_0|\{x_k\}, \{I_k\})$ with the added intensity data rather than from the two-dimensional posterior $P(\alpha, \beta|\{x_k\})$ led to a reduced standard deviation and a less negative value for α. The estimate for alpha is improved by providing greater confidence.

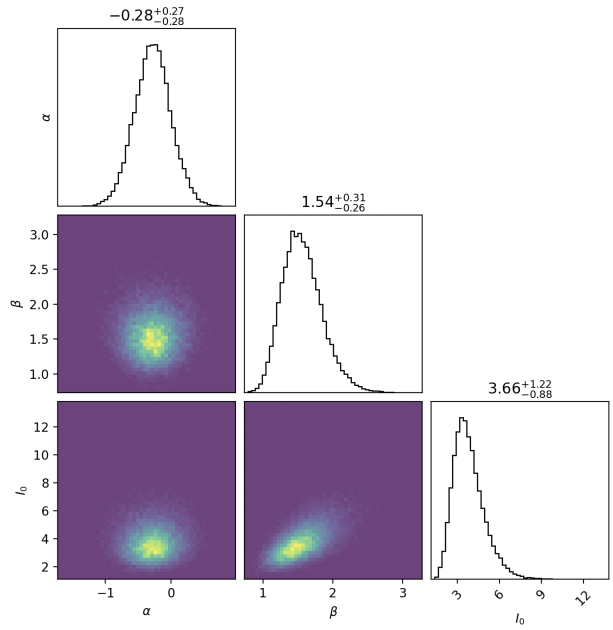

Figure 4: Corner plot of samples from the posterior distribution $P(\alpha, \beta, I_0|\{x_k\}, \{I_k\})$

References

[1] D. Sivia. *Data Analysis: A Bayesian Tutorial*. Oxford Science Publications. New York: Oxford University Press, 1996.

[2] Dan Foreman-Mackey et al. *Emcee Documentation*. Copyright 2012-2021, Dan Foreman-Mackey & contributors. version 3.1.4. 2021. URL: `https://emcee.readthedocs.io/en/stable/` (visited on 03/10/2024).

[3] Daniel Foreman-Mackey et al. "emcee: The MCMC Hammer". In: *Publications of the Astronomical Society of the Pacific* 125.306 (Mar. 2013), pp. 306–312. DOI: `10.1086/670067`.

[4] Jonathan Goodman and Jonathan Weare. "Ensemble Samplers with Affine Invariance". In: *Communications in Applied Mathematics and Computational Science* 5.1 (2010), pp. 65–80. DOI: `10.2140/camcos.2010.5.65`.

[5] Daniel Foreman-Mackey. "corner.py: Scatterplot matrices in Python". In: *The Journal of Open Source Software* 1.2 (June 2016), p. 24. DOI: `10.21105/joss.00024`. URL: `https://doi.org/10.21105/joss.00024`.

[6] Andrew Gelman and Donald B. Rubin. "Inference from Iterative Simulation Using Multiple Sequences". In: *Statistical Science* 7.4 (1992), pp. 457–511.

[7] Dootika Vats and Christina Knudson. *Revisiting the Gelman-Rubin Diagnostic*. Preprint. Department of Mathematics and Statistics, Indian Institute of Technology Kanpur, Kanpur, IN - 208016, India; Department of Mathematics, University of St. Thomas, St. Paul, Minnesota 55105. Sept. 2020.

Auto-generation Citations

ChatGPT version 4.0 was used for: - Setting the labels on the sides of the corner plots from sampling from both the two-dimensional and three-dimensional posterior. The following prompt was used alongside the code for the corner and histogram plot: 'How to set the labels on the edges of the corner plot'. - Removing the tick marks on the histograms added to the corner plot for sampling. The following prompts were used alongside the code for the corner and histogram plot: 'How to hide x-axis labels and ticks for all but the bottom row" and "How to hide y-axis labels and ticks for all but the first column'. - Initially, my docker container did not flush in real time to the terminal. I debugged this with ChatGPT and modified my docker code, accordingly. I used the following prompts, alongside my dockerfile script: 'Why isn't the output of my dockerfile flushing in real time on my Mac?' and 'What can I change such that my dockerfile flushes the output - I know it has something to do with conda'.

GitHub Copilot was used to help write documentation for Doxygen and comments within the code.