Imperial College London

IMPERIAL COLLEGE LONDON

DEPARTMENT OF MATHEMATICS

Markov Chain Monte-Carlo Methods in application to Astrophysics

Author:
Samuel Korsen
Christine Zhang
Haley Lo
Mingze Ma
Sally Marshall

Supervisor:
James Martin

June 2021

Contents

Abstract

Change point models are applicable in many real-world settings where the distribution of a stochastic process changes at a specific point. Properties of such models, and specifically the posterior distributions surrounding the change point, can be determined via Markov Chain Monte-Carlo (MCMC) methods, which we will explore within this report.

In this report we study and formulate 4 successful models for data from the emission of photons from stars. In the first 2 models, we assume the number of change points = 1. We will then use 2 popular MCMC algorithms - the Gibbs Sampler and Metropolis Hastings - to generate Markov Chains which will allow us to appropriately model our observed data. Model 2 is an improvement of model 1. Following this, we will go on to use several statistical analysis and model-checking methods to assess the suitability of our models as well as the efficiency of the corresponding algorithms by inspecting the convergence properties.

In the last 2 models, we relax the assumption of the number of change point being equal to 1. In the first model, we increase the number of change points of the model up till a point where two parameters converge. Then in the second model, we adopt Reversible Jump Markov Chain Monte Carlo (RJMCMC) to model our data.

Chapter 1

Introduction

1.1 Physical Problem

This report will explore techniques used in modelling photon arrivals from the Orion Nebula. Stars constantly emit photons which can be observed by telescopes, and the frequency of such observations over given time steps can be modelled as a Markov Chain. Throughout a stars lifetime, the distribution of such emissions changes. This can be mathematically viewed as a change to our Markov Chain at a specific 'change point' in the time series. The data analysed in this report is binned data of the number of UV photons emitted from stars on the Orion Nebula over a 2 week period. Using Markov Chain Monte-Carlo (MCMC) methods, we will explore the change points over the given time period.

1.2 Bayesian Analysis

The fundamental statistical approach taken throughout this report is Bayesian Analysis. The flexibility of a Bayesian approach makes it appropriate in this setting, specifically with regards to applications in Markov Chain Monte Carlo methods where data sets may be complex, with missing entries or underlying errors. The eventual aim in using Bayesian Analysis is to find a suitable probability distribution which is assigned to a random variable dependent on observations from our experimental data - known as the posterior distribution (1).

The core principle behind Bayesian statistics can be divided into 3 key steps. Initially, we select a distribution we believe best represents how we expect the parameters to behave in the model based on our preliminary analysis prior to the addition of data into the model - this distribution is known as the prior distribution. Using the parameters derived from the distribution, we calculate the probability density function of observing our given data given the parameters used - this function is the likelihood. Finally, we test our model using the observed data, the results of which we use to update the prior distribution via the likelihood function. (1)

1.3 Bayes Theorem

The eventual outcome of Bayesian inference is to establish our posterior distribution. In this approach, both parameters and observations are considered to be random quantities. Let D denote the observed data and θ denote model parameters. We can then express Bayes' Theorem in terms of probability distribution (2):

$$f(\theta|D) = \frac{f(D|\theta)f(\theta)}{f(D)}$$

Where $f(\theta|D)$ is the posterior distribution for parameter θ; $f(\theta)$ is the prior distribution for our parameter; $f(D)$ is the marginal probability of the data and $f(D|\theta)$ is the sampling density for our observed data, which is proportional to the Likelihood function. Note, we can calculate the marginal probability of the data by:

$$f(D) = \int f(D|\theta)f(\theta)d\theta$$

From the combination of the properties of these distributions, we deduce that the posterior distribution is directly proportional to the product of our prior distribution and the likelihood function.

In order to use Bayes' Theorem to our advantage, we can follow a set method. First, we require the assumption that our model parameters are fixed quantities throughout, which will allow us also to assume our parameter θ is the realisation of a random variable which follows the prior distribution - $p(\theta)$. We can then calculate our likelihood function $L(D|\theta)$, using our observed data D given parameter θ. From our results above, we can calculate the posterior distribution. Noting $p(D)$ is independent of θ, we require only to normalise our equation to find:

$$f(\theta|D) = kf(D|\theta)f(\theta)$$

where k is our constant of proportionality. (2)

We can further our statistical investigation by analysis of this new distribution and it's propoerties, for example, obtaining credible intervals or posterior parameter, which are incredibly useful when refining our model.

Chapter 2

Markov Chain Monte Carlo

2.1 Markov Chains

A stochastic process on the state space Ω is a collection of Ω-valued random variables $(X_t)_{t\in T}$, indexed by a set T which represents time. A Markov chain is a type of stochastic process, in which the defining property is the fact that the chain has 'no memory' of where it has been. Formally, the probability distribution of the subsequent state only depends on the current state, so the future is conditionally independent of the past, given the present. So given random variables $X_1, X_2, X_3, ...$ we can say:

$$Pr(X_n|X_{n-1}, X_{n-2}, ..., X_0) = Pr(X_n|X_{n-1})$$

for all $n \in T$ and all state spaces X_i. We also assume the Markov chain is time homogeneous; the probability of one particular state transitioning to another state remains the same throughout time:

$$Pr(X_{n+1} = j|X_n = i) = Pr(X_1 = j|X_0 = i)$$

for all $n \in T$ and all $i, j \in \Omega$. (3)

2.2 Monte-Carlo

Monte Carlo methods approximate the solution of a problem as a parameter of a distribution, then make statistical estimates about the solution by using a random sequence to generate a sample of that distribution (4). The stochastic nature of our problem means these methods are very applicable; we can use Monte-Carlo methods to characterise the distribution of the photon emissions before and after the change point, and also calculate expectations of the parameters at hand (5).

2.3 Markov Chain Monte-Carlo Methods

We can combine Markov chains and Monte-Carlo as defined above to form Markov Chain Monte-Carlo (MCMC). In MCMC methodology, we construct a Markov chain so that the distribution of a random sample from the chain, $x_1, x_2, ..., x_n$, converges to the real posterior distribution. (6) In other words, the posterior distribution is the stationary distribution of the Markov chain. Samples drawn in correct proportions throughout the support of the posterior can be used to calculate expectations (7).

2.4 Improving MCMC Methods

Later in this report in Chapter 4, we look at ways to improve on the initial MCMC methods. In this section, we will briefly look at the theory behind some of these improvements.

2.4.1 Changing proposal and prior distributions

Changing the proposal and prior distribution is an efficient way to improve the acceptance rate and the convergence speed. However, this method needs lots of mathematical analysis and intuition.

Changing the proposal distribution increases acceptance rate because the new variable waiting for acceptance or rejection needs to be taken from the proposal distribution. If the expectation of the proposal is roughly the same as real value of the variable, and the variance is low, then it will have a high chance of accepting the proposal value. Conversely, if the variance is large or the expectation of proposal distribution is far from the real value, then it may have a low acceptance rate. In order for the expectation to be close to the real value, we usually take an updating expectation for proposal distribution and obviously, a large acceptance rate will improve the convergence rate.

We can also improve the model by changing the prior distribution to improve the posterior distribution. Since the posterior distribution has changed, the full conditional distribution will also change. By changing the full conditional distribution, we can modify it into a known distribution and use Gibbs sampler, which is more efficient than the normal Metropolis sampler.

2.4.2 Burn-in

The burn-in is a number of iterations we remove from the beginning of the sample, as the convergence of the sample may initially be slow. The expectation calculated after the removal of these burn-in samples is called the ergodic average, and by the Ergodic theorem, this is guaranteed to converge to the expectation (7). Convergence diagnostics are tools for finding out the length of the burn-in, but as this length is likely to be less than 1% of the total length, Geyer suggests that this calculation is unnecessary (8).

Chapter 3

Change Point Model/ Model Selection

3.1 Bayesian Change Point Models

A change point model is a model for dealing with data where certain characteristics of the underlying distribution of the data undergo occasional changes. More specifically, this leads to a situation where we treat the data as initially coming from a certain distribution, and then beyond a certain time point, as coming from a different distribution (or from the same distribution family but with different parameters). In fact, we are not limited to having only one change point; the characteristics of the underlying distribution of the data may change twice or more. Both single change point models and multiple change point models will be discussed in this report.

As discussed in (9), one may create a change point model through the introduction of a dummy variable, which takes the value of 0 up to the change point and the value of 1 after the change point until the following change point where it returns to 0. A simple approach in finding our change point k involves modelling a range of values for k and evaluating each model's goodness of fit. A Bayesian approach to change point estimation can provide a much more efficient identification, specifically in the case where the change point isn't easily apparent in the data. Analysis of log likelihood functions for k; derivation of a marginal posterior distribution for k and the use of Markov Chain Monte Carlo methods including the Metropolis-Hastings Sampler (Subsection 4.1.1) and The Gibbs Sampler (Subsection 4.1.2) - the ones which we will cover in this report - are all examples of such methods which can be used.

3.2 Selecting our change point model

We will choose to model the data of photon emission counts using a Bayesian Change Point Model. Exploratory analysis on the data suggests that this model is appropriate as the rate of photon emission does not seem to be constant throughout the time period in which data was collected. This matches with the theory of our physical problem as it is known that the rate of photon emission by a star varies depending on several factors, such as the evolution and the spectral type of the star (10). This pattern is of importance in astrophysics as it can be indicative of the lifetime of the star. Looking at Figure 6.5 below, we initially predict a change point at approximately time bin 10 (and potentially another one at approximately time bin 30).

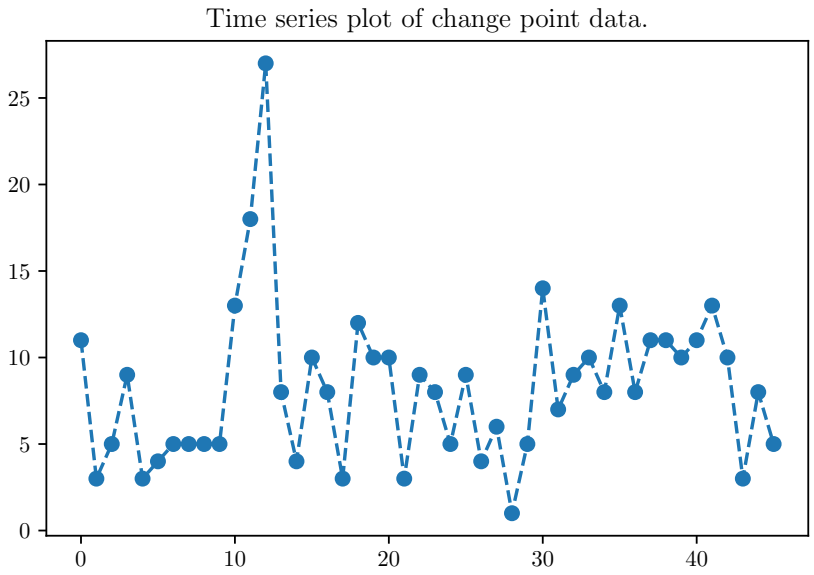


Figure 3.1: Binned Count Data of number of photon emissions in each time bin of length 10,000 seconds

The times at which photons arrive can be viewed as a Poisson process, so we will incorporate this into our change point model by modelling the photon count as a Poisson random variable with a rate parameter that changes at some point (or at several different points) in the process.

Following the change point model suggested by Penn State Astrostatistics MCMC tutorial (11), and letting C_i be the count of photon emissions in time bin i, initially with just a singular change point in time bin k, this can be written as:

$$C_i|k,\theta,\lambda \sim \text{Poisson}(\theta) \text{ for } i = 1,...,k$$
$$C_i|k,\theta,\lambda \sim \text{Poisson}(\lambda) \text{ for } i = k+1,...,n$$

The parameters of main interest in our model are the change point location k, as well as θ and λ, the Poisson rate parameters of the data distribution before and after the change point respectively.

Chapter 4

Our Method

4.1 Sampler Methods

4.1.1 Metropolis Hastings

The Metropolis-Hastings algorithm is the general form of the algorithm used in Markov Chain Monte Carlo methods. The goal of this algorithm is to construct a Markov chain whose stationary distribution is our distribution of interest. In our case, this is the full conditional posterior distribution of the model parameters.

Let $\pi(x)$ be the full conditional posterior distribution of the parameters. Let P be our vector of parameters and say we have values for our parameters at time t. The state of P at time t is P_t The method for determining the state of our Markov chain at time $t+1$ is as follows (7):

1. Sample a candidate point Y from a proposal distribution, $q(x|P_t)$ which in fact does not have a required form, but the choice of this proposal distribution can affect the convergence behaviour of the chain.
2. The proposed point Y is then accepted with a carefully calculated acceptance probability according to: $\alpha(P_t, Y) = \min\left(1, \frac{\pi(Y)q(P_t|Y)}{\pi(P_t)q(Y|P_t)}\right)$
3. If the candidate point is accepted, then the next state of the Markov chain becomes $P_{t+1} = Y$, otherwise we stay at our previous value of the chain and $P_{t+1} = P_t$.

Single component Metropolis Hastings (7) refers to the Metropolis Hastings algorithm, but instead of updating all parameters at once, at each iteration of the algorithm we update the components of our vector of parameters one by one. We use the most recent values of all the other parameters when sampling from the proposal distribution and when calculating the acceptance probability.

4.1.2 Gibbs Sampler

The Gibbs sampler is a special case of single component Metropolis-Hastings sampling. It is only possible to use this method when we know the full conditional posterior distribution of our parameter, and it is easy to sample from. For each parameter, we assume the other parameters are known and consider the uni-variate conditional distribution when conditioning on the so-called "known" parameters. By developing these uni-variate distributions for each parameter, we can easily deduce a multivariate distribution which makes the Gibbs Sampler

incredibly useful to many applications of MCMC. (12)

In the Gibbs Sampler, the acceptance probability is always 1. By expressing the updated proposal distribution for the i$^{\text{th}}$ component of our sample X:

$$q_i(Y_i|X_i, X_{-i}) = \pi(Y_i|X_{-i})$$

where $\pi(Y_i|X_{-i})$ is the full conditional distribution of the i$^{\text{th}}$ component of X, conditioning on all the remaining components of X; we find this to be true. From the above expression, we see samples are only taken from full conditional distributions, where we condition on the other parameters which we have previously assumed to be known.

We outline the Gibbs Sampler algorithm where we consider m unknown parameters:

1. We fix $\theta_1^0, \theta_2^0, ..., \theta_m^0$ which represent our estimates for the parameters at time 0.
2. For each parameter, we condition on the other parameters from the previous time step. This establishes the relationship at time t as:
   - Sample $\theta_1^t \sim \pi(\theta_1|\theta_2^{(t-1)}, ..., \theta_m^{(t-1)}, Y)$
   - $\vdots$
   - Sample $\theta_i^t \sim \pi(\theta_i|\theta_1^{(t)}, ..., \theta_{(i-1)}^{(t)}, \theta_{(i+1)}^{(t-1)}, ..., \theta_m^{(t-1)}, Y)$
   - $\vdots$
   - Sample $\theta_m^t \sim \pi(\theta_m|\theta_m^{(t)}, ..., \theta_{(m-1)}^{(t)}, Y)$

4.2 Our First Model

As stated in Section 3.2, our initial model is based on the tutorial (11). We implemented this model using a function in Python which we coded, based on a function given in the tutorial. A brief overview of the method used in the function is as follows.

We initialise 5 Markov chains, one for each parameter, and set the first value in each chain to be the initial value for our parameters. Then we update each parameter in turn, according to the process outlined in Section 4.2.3, and add the new value to the corresponding chain. This process is repeated for our number of iterations and we are left with 5 Markov Chains, one for each parameter in our model.

4.2.1 Prior Distributions

Our initial prior distributions are as follows:

$$\begin{aligned}
\theta|b_1 &\sim \text{Gamma}(0.5, b_1) \\
\lambda|b_2 &\sim \text{Gamma}(0.5, b_2) \\
b_1 &\sim \text{Inverse Gamma}(0, 1) \\
b_2 &\sim \text{Inverse Gamma}(0, 1) \\
k &\sim \text{Uniform}(1, ..., n)
\end{aligned}$$

Despite having a rough idea of where the change point in the data is, we have decided to use a uniform prior distribution as we cannot be certain of the change point location, and from a Bayesian perspective we should therefore treat it as random.

4.2.2 Posterior Distributions

We are interested in sampling from the posterior distributions. As outlined in Section 1.3, we can obtain the full posterior distribution of our parameters up to a normalising constant by multiplying the likelihood function of the data with the prior distributions of the parameters. As detailed in fullcond.pdf (13), given the data $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$, the posterior distribution is:

$$f(k, \theta, \lambda, b_1, b_2|\mathbf{Y}) \propto \prod_{i=1}^{k} \frac{\theta^{Y_i} e^{-\theta}}{Y_i!} \prod_{i=k+1}^{n} \frac{\lambda^{Y_i} e^{-\lambda}}{Y_i!} \times \frac{1}{\Gamma(0.5) b_1^{0.5}} \theta^{-0.5} e^{-\theta/b_1}$$
$$\times \frac{1}{\Gamma(0.5) b_2^{0.5}} \lambda^{-0.5} e^{-\lambda/b_2} \times \frac{e^{-1/b_1}}{b_1} \frac{e^{-1/b_2}}{b_2} \times \frac{1}{n}$$

Using the above expression we can obtain marginal full conditional distributions for each parameter by ignoring all terms that are constant with respect to the parameter. This gives us the following full conditional distribution for θ, λ, k, b_1 and b_2 respectively:

$$f(\theta|k, \lambda, b_1, b_2, \mathbf{Y}) \propto \text{Gamma}\left(\sum_{i=1}^{k} Y_i + 0.5, \frac{b_1}{k b_1 + 1}\right)$$
$$f(\lambda|k, \theta, b_1, b_2, \mathbf{Y}) \propto \text{Gamma}\left(\sum_{i=k+1}^{n} Y_i + 0.5, \frac{b_2}{(n-k) b_2 + 1}\right)$$
$$f(k|\theta, \lambda, b_1, b_2, \mathbf{Y}) \propto \theta^{\sum_{i=1}^{k} Y_i} \lambda^{\sum_{i=k+1}^{n} Y_i} e^{-k\theta - (n-k)\lambda}$$
$$f(b_1|k, \theta, \lambda, b_2, \mathbf{Y}) \propto \frac{1}{b_1^{0.5}} e^{-\theta/b_1} \times \frac{e^{-1/b_1}}{b_1} \propto b_1^{-1.5} e^{-(1+\theta)/b_1} \propto \text{IG}(0.5, \frac{1}{\theta + 1})$$
$$f(b_2|k, \theta, \lambda, b_1, \mathbf{Y}) \propto \frac{1}{b_2^{0.5}} e^{-\lambda/b_2} \times \frac{e^{-1/b_2}}{b_2} \propto b_2^{-1.5} e^{-(1+\lambda)/b_2} \propto \text{IG}(0.5, \frac{1}{\lambda + 1})$$

4.2.3 Sampler methods used in this model

Here we have used the single-component Metropolis Hastings algorithm to sample from our posterior distribution and within this we have used the Gibbs sampler.

As derived above, the posterior full conditional distributions for θ, λ, b_1 and b_2 were known and easy to sample from. We were therefore able to use a Gibbs update for these parameters at each iteration of the algorithm. The change point k is different as its posterior full conditional distribution is known, but it is not easy to sample from. This leads us to use the Metropolis-Hastings update on this parameter. The proposal distribution for k used in this step was a uniform distribution over the range of time bins.

4.2.4 Plots and Results

The Markov chains generated by these sampler methods can be used for inference as they are sampled approximately from the posterior distributions of each parameter. Our output was 5 Markov chains each of length 100,000. Something to note here for our change point parameter k is that the acceptance rate during the process was only 8.6%. This is a flaw in this model as it is too low and it leads to the data having high auto-correlation, an effect

which we wish to reduce as much as possible. Auto-correlation of the models will be further discussed in Chapter 5.

Trace Plots

Trace plots show the values that each parameter takes over the course of the iteration process. These trace plots are good to get an initial understanding of the values the parameters are taking, and the convergence behaviour of each Markov chain. We can see that the values of θ and λ are concentrated in fairly tight intervals (4-9 and 8-10 respectively) with occasional fluctuations much further away from this range. The change point k fluctuates in value much more in the region between 0 and 40, however we can see there is a darker patch at $k = 10$ suggesting this is roughly its most common value. The trace plots for b_1 and b_2 were not included as they were uninformative, due to a lack of convergence by these parameters. See Figure 4.1.

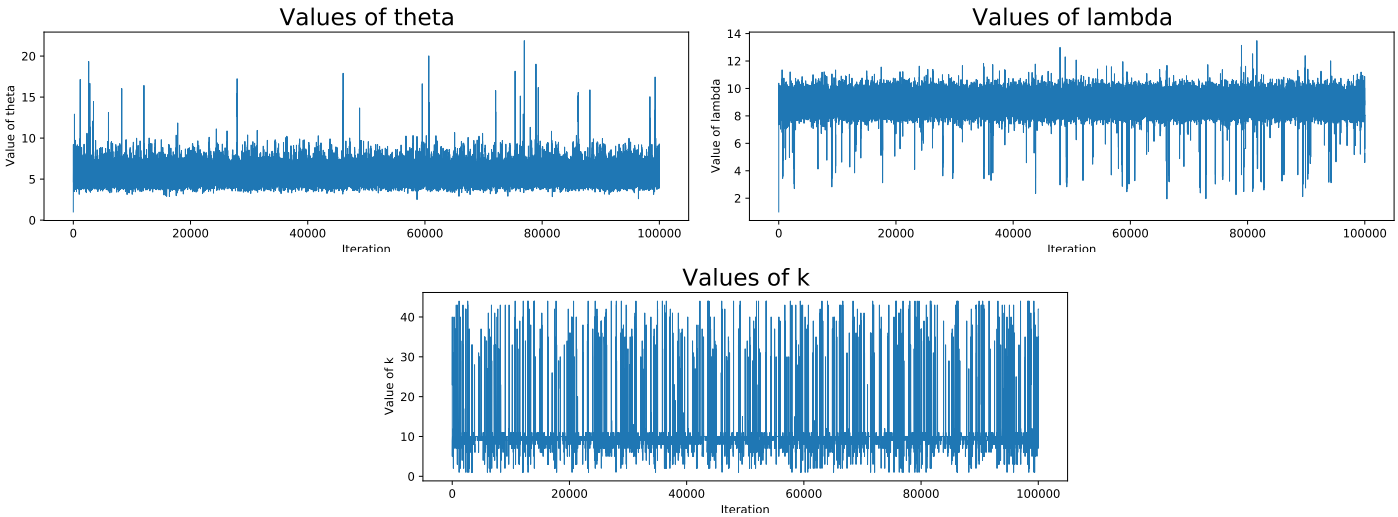


Figure 4.1: Trace plots

Smoothed density plots/Histograms for posterior distributions

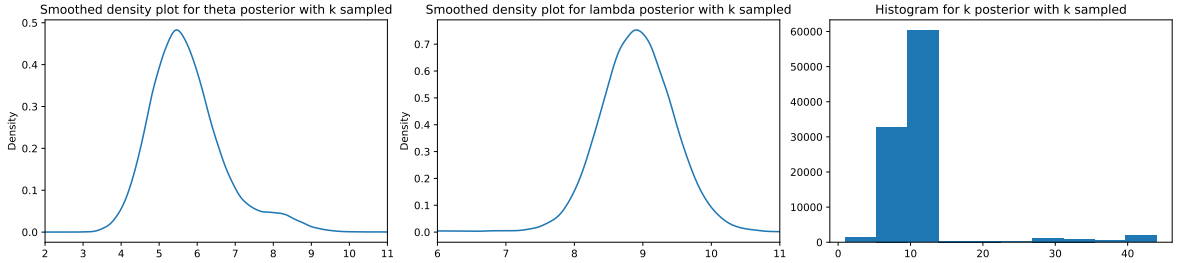
Based on the Markov chains we have generated, we can approximate the posterior probability density functions of each parameter as shown in the smoothed density plots for θ and λ, and a histogram for k in Figure 4.2.

Monte Carlo estimates vs sample size plots for

We also coded a function based on the tutorial (11) to calculate how the sample size affected the Monte Carlo estimate of the mean of the posterior distribution of each parameter. This is shown in the following figures; the figures for b_1 and b_2 were not very informative (due to extremely large fluctuations away from a modal value near zero) so we have only included the plots for θ, λ and k in Figure 4.3.

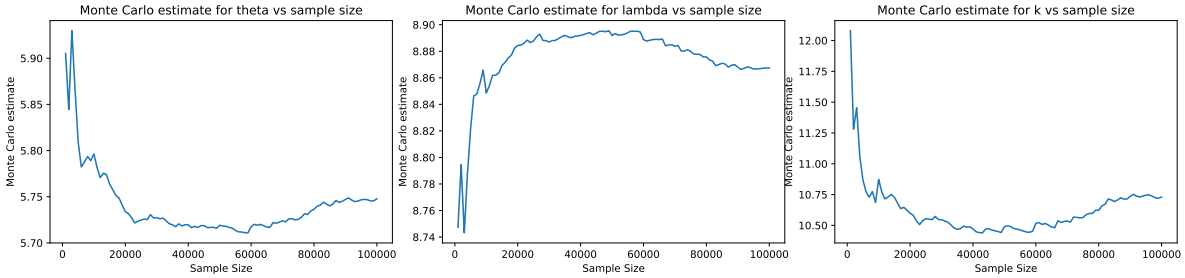
Results

The expected values of the parameters, based on our Markov chains of length 100,000, are as follows:



(a) Theta (b) Lambda (c) Change point k

Figure 4.2: Posterior densities and histogram



(a) Theta (b) Lambda (c) Change point k

Figure 4.3: MC estimate vs sample size

	θ	λ	k	b_1	b_2
Expected Value	5.75	8.87	10.73	36866121	46585080

So we can deduce that the data ($E(k)$) has a change point approximately between time bins 10 and 11. The data ($E(\theta)$) before the change point is coming from a Poisson distribution with rate parameter approximately equal to 6, and after the change point the data ($E(\lambda)$) is coming from a Poisson distribution with rate parameter approximately equal to 9.

4.3 Our Second Model

A second initial model is motivated by looking at which basic changes we can make to our first model. A simple improvement we will focus on is to improve the acceptance rate for our change point parameter k. An acceptance rate of approximately 15-35% is what we aim for in MCMC methodology. To do this, we will change the proposal distribution for k. We will now use a Gamma distribution centered at time bin 10. We will also allow for a burn-in period of length 5000 iterations, in order to allow the chain to settle and to ignore any early behaviour where the chain may not be convergent.

Finally, we will also change the prior distributions of b_1 and b_2, in order to see if the behaviour of the Markov chains corresponding to these parameters improves and see if we get better convergence than in our first model, where the mean was undefined as the α parameter in the posterior Inverse Gamma distributions of b_1 and b_2 was 0.5, and the mean of an inverse gamma distribution is undefined for $\alpha < 1$.

4.3.1 Prior Distributions

Our prior distributions are now as follows:

$$\begin{aligned}\theta|b_1 &\sim \text{Gamma}(0.5, b_1)\\ \lambda|b_2 &\sim \text{Gamma}(0.5, b_2)\\ b_1 &\sim \text{Gamma}(0.01, 100)\\ b_2 &\sim \text{Gamma}(0.01, 100)\\ k &\sim \text{Uniform}(1, ..., n)\end{aligned}$$

4.3.2 Posterior Distributions

We now derive the posterior distribution in the same way as in Section 4.2.2, by taking the product of the likelihood of the Poisson distributed data with the prior distributions of the parameters. As detailed in chptmodel.pdf (14), given the data $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$, the posterior distribution is:

$$\begin{aligned}f(k, \theta, \lambda, b_1, b_2|\mathbf{Y}) \propto \prod_{i=1}^{k} \frac{\theta^{Y_i} e^{-\theta}}{Y_i!} \prod_{i=k+1}^{n} \frac{\lambda^{Y_i} e^{-\lambda}}{Y_i!} \times \frac{1}{\Gamma(0.5) b_1^{0.5}} \theta^{-0.5} e^{-\theta/b_1} \times \frac{1}{\Gamma(0.5) b_2^{0.5}} \lambda^{-0.5} e^{-\lambda/b_2}\\ \times \frac{1}{\Gamma(0.01) 100^{0.01}} b_1^{-0.99} e^{-b_1/100} \times \frac{1}{\Gamma(0.01) 100^{0.01}} b_2^{-0.99} e^{-b_2/100} \times \frac{1}{n}\end{aligned}$$

Again, using this posterior distribution, we can obtain marginal full conditional posterior distributions for each parameter by ignoring all terms that are constant with respect to the parameter:

$$\begin{aligned}f(\theta|k, \lambda, b_1, b_2, \mathbf{Y}) &\propto \text{Gamma}\left(\sum_{i=1}^{k} Y_i + 0.5, \frac{b_1}{k b_1 + 1}\right)\\ f(\lambda|k, \theta, b_1, b_2, \mathbf{Y}) &\propto \text{Gamma}\left(\sum_{i=k+1}^{n} Y_i + 0.5, \frac{b_2}{(n-k) b_2 + 1}\right)\\ f(k|\theta, \lambda, b_1, b_2, \mathbf{Y}) &\propto \theta^{\sum_{i=1}^{k} Y_i} \lambda^{\sum_{i=k+1}^{n} Y_i} e^{-k\theta - (n-k)\lambda}\\ f(b_1|k, \theta, \lambda, b_2, \mathbf{Y}) &\propto \frac{1}{b_1^{0.5}} e^{-\theta/b_1} \times b_1^{-0.99} e^{-b_1/100}\\ f(b_2|k, \theta, \lambda, b_1, \mathbf{Y}) &\propto \frac{1}{b_2^{0.5}} e^{-\lambda/b_2} \times b_2^{-0.99} e^{-b_2/100}\end{aligned}$$

4.3.3 Sampler methods used in this model

We will again use a single component Metropolis-Hastings algorithm on our parameters, and within this we will again use the Gibbs sampler. This time we can use a Gibbs update on θ and λ, as we can easily sample from their posterior distributions and Metropolis-Hastings update on k, b_1 and b_2, as their posterior distributions are not well known. This may not seem like an improvement initially, but we found that the behaviour of the chains was better in this case, as in the first model the means of the Inverse Gamma posterior distributions for b_1 and b_2 were both undefined.

For the proposal distribution of k, b_1, b_2, we assume them to be $k \sim \text{Gamma}(10, 1)$, $b_1 \sim \text{Gamma}(1, 1.1)$ and $b_2 \sim \text{Gamma}(1.5, 1)$. The proposal distributions used in the Metropolis Hastings Algorithm are chosen so the acceptance rate of k, b_1, b_2 are around 15-35%.

4.3.4 Plots and Results

Trace Plots and Smoothed Density Plots

When comparing these trace plots to those in our first model, we see that the traces are more centered about the means. There are fewer fluctuations, especially for the change point parameter k. The trace plots for b_1 and b_2 show that these parameters now take values in a much much smaller range (between 0 and 10 for both of them) than in the first model, which makes sense when we consider b_1 and b_2 as parameters in the distributions of θ and λ respectively. See Figures 4.4 and 4.5.

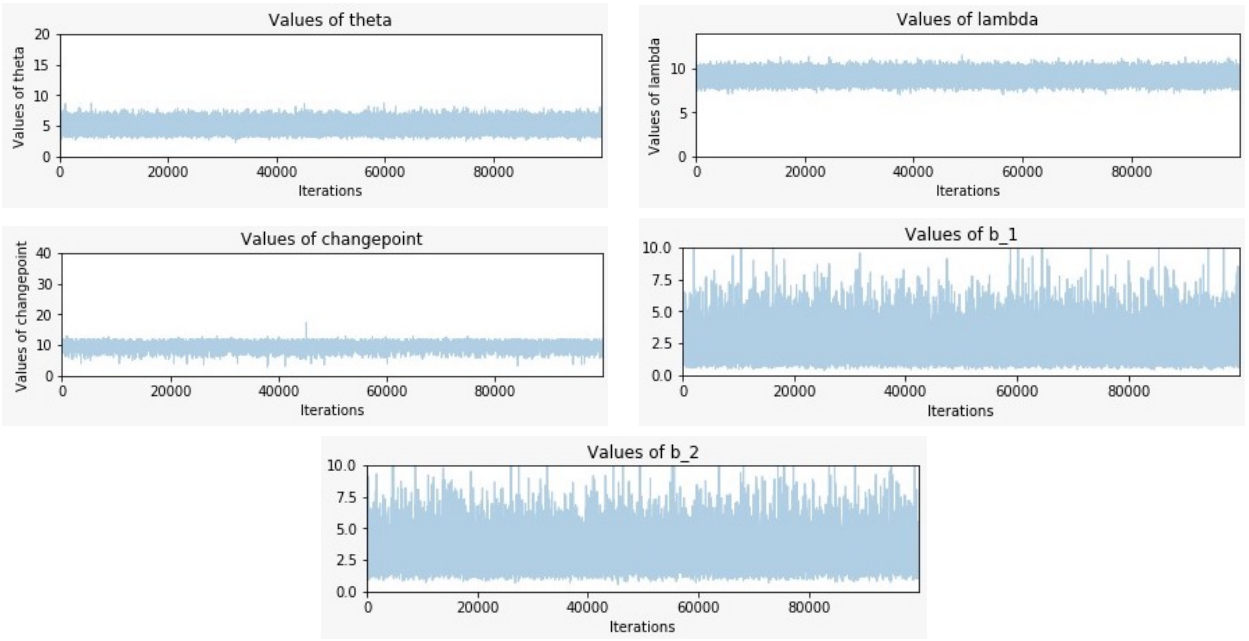
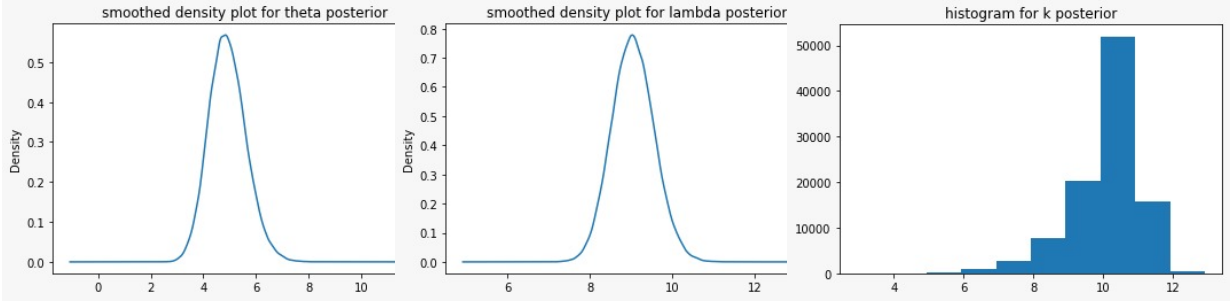


Figure 4.4: Trace Plots

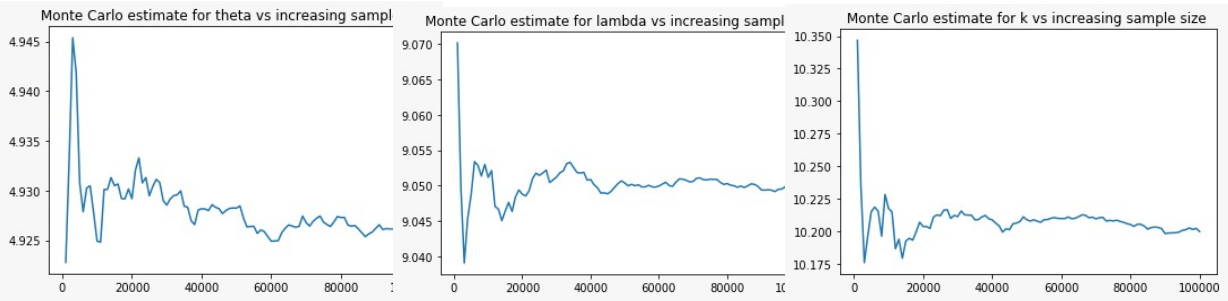


(a) Theta (b) Lambda (c) Change point k

Figure 4.5: Posterior densities and histogram for θ, λ, k

Monte Carlo estimates vs sample size plots

These plots show much better convergence of the Monte Carlo estimates of the means of θ, λ and k, as the scale on the y axis is much smaller than in our first model. See Figure 4.6.



(a) Theta (b) Lambda (c) Change point k

Figure 4.6: MC estimate vs sample size

Results

We can deduce that this new model is an improvement as the new acceptance rates are closer to the desired range. The acceptance rate for the change point parameter k is now 17.5% and the rates for b_1 and b_2 are 34.2% and 27.8% respectively.

The reason for the increased acceptance rate of k is the updated choice of proposal distribution for k at each iteration of the Metropolis-Hastings algorithm. In our first model we used a uniform proposal distribution, but this time we used a Gamma(10,1) distribution.

Another improvement is the trace plots for b_1 and b_2. They now show that the values of these parameters do not have huge fluctuations like in our first model, so the mean values of these posterior distributions now make sense. This will be discussed further in the next chapter. This is to do with the fact that in model 1, the mean of the posterior distribution was undefined but using our second prior distributions, we see an improvement.

The new expected values of of our model parameters are now as follows:

	θ	λ	k	b_1	b_2
Expected Value	4.93	9.05	10.20	2.32	3.27

The means of b_1 and b_2 are much much smaller and can actually be interpreted now. The inference here is similar to that in the first model, except for the fact that according to this model, before the change point the data is now coming from a Poisson distribution with rate parameter approximately equal to 5 instead of 6.

Chapter 5

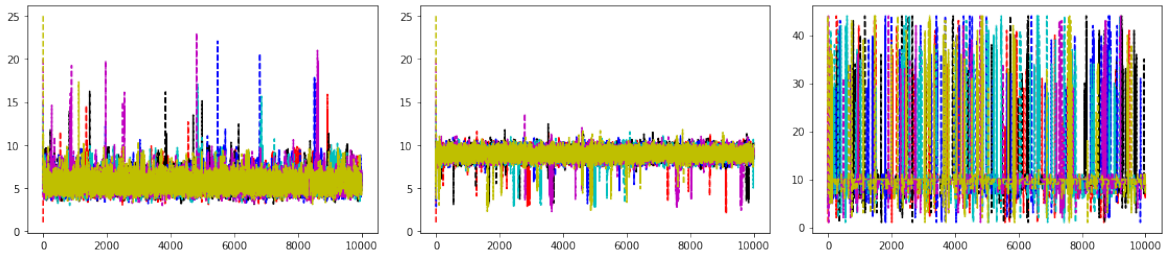
Model Justification

In this chapter, we will focus on some methods for checking the suitability and the goodness of fit of our first 2 models, and we will make comparisons between the 2. All Python functions mentioned are from the `arviz` package.

5.1 Convergence Diagnostics

5.1.1 Different Initial Values

The first method for checking the convergence of a Markov chain is by varying the initial value and observing if the chains still converge to the same value. This is quite an intuitive method to ensure the convergence of the model. Once we know that all chains, regardless of initial value, converge to a similar value, we can confirm that this is a stable point for the Markov chain. Furthermore, when the number of samples becomes very large we can almost ensure that the Markov chain has converged. We will compare the effects of different initial values for θ, λ and k for both models below. The different colours of lines on the plots correspond to different initial values of the Markov chain corresponding to that particular parameter. For Model 1 see Figure 5.1.



(a) Initial value of θ = 1, 5, 10, 15, 20, 25

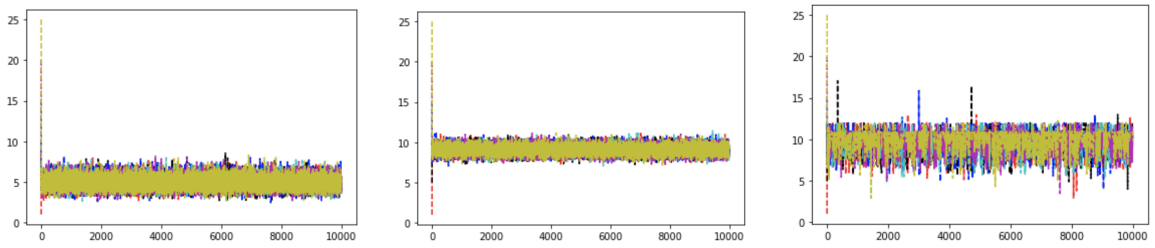
(b) Initial value of λ = 1, 5, 10, 15, 20, 25

(c) Initial value of k = 1, 5, 10, 15, 20, 25

Figure 5.1: Trace plots for model 1

As we can see in Figure 5.2, for our second model the Markov chains corresponding to θ, λ and k fluctuate either the same amount or less than in model 1 for a range of initial values. This is evidence that our second model is an improvement.

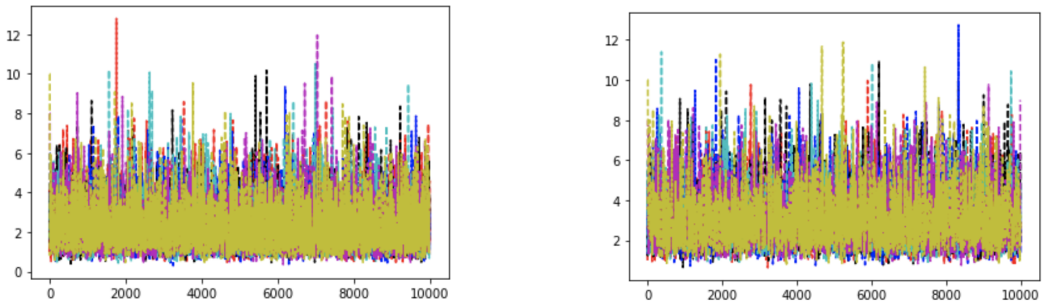
We have also generated these plots for our second model for parameters b_1 and b_2 in Figure



(a) Initial value of θ = 1, 5, 10, 15, 20, 25 (b) Initial value of λ = 1, 5, 10, 15, 20, 25 (c) Initial value of k = 1, 5, 10, 15, 20, 25

Figure 5.2: Trace plots for model 2

5.3. These plots show that unlike in our first model, b_1 and b_2 now converge, for a range of initial values. This further suggests that our second model is an improvement.



(a) initial value of b_1 = 1, 2, 4, 6, 8, 10 (b) initial value of b_2 = 1, 2, 4, 6, 8, 10

Figure 5.3: Trace plots for model 2

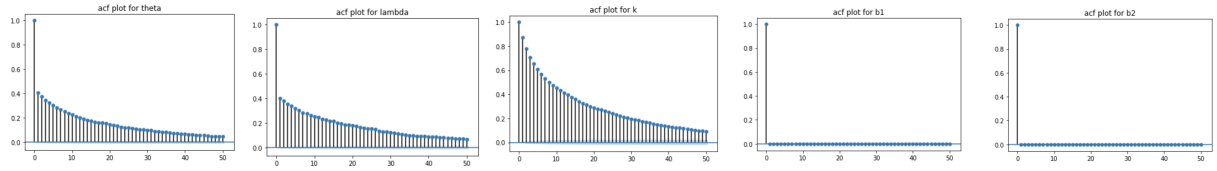
5.1.2 Autocorrelation

Autocorrelation is a measure of the correlation between two samples that are a certain time interval apart from each other. Let us denote this time interval to be k. This autocorrelation is then called the lag k autocorrelation, and is defined as follows (15):

$$ACF_k = \frac{\sum_{i=1}^{N-k}(Y_i - \bar{Y})(Y_{i+k} - \bar{Y})}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$$

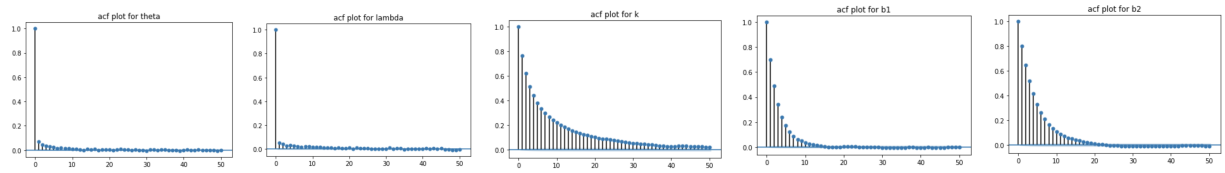
for samples $Y_1, Y_2, ..., Y_N$ at time points $1, 2, ..., N$. This formula assumes the time points are equally spaced, which works with our data (our time bins are of equal length).

We can use autocorrelation plots (plots with various lag k) to check for randomness; if random, the autocorrelation should be near zero. We only plot the positive half as the autocorrelation is symmetric about zero (16). We expect the autocorrelation plots to converge to zero as k increases, as samples further apart should have more independence. The quicker the decay to zero, the more independent the samples are from each other. We can see the autocorrelation plots for Model 1 and Model 2 in Figure 5.4 and Figure 5.5 respectively.



(a) acf for θ (b) acf for λ (c) acf for k (d) acf for b_1 (e) acf for b_2

Figure 5.4: Autocorrelation plots for the parameters in our first model



(a) acf for θ (b) acf for λ (c) acf for k (d) acf for b_1 (e) acf for b_2

Figure 5.5: Autocorrelation plots for the parameters in our second model

Despite the fact that the b_1 and b_2 plots show more autocorrelation, the second model is definitely an improvement. We have almost entirely removed the autocorrelation for θ and λ and it has been reduced for k, shown by the faster decay of the plot.

5.1.3 Effective Sample Size

The effective sample size (ESS) is closely related to the autocorrelation. The ESS compares the length of the Markov chain with the size of independent and identically distributed random sample from the posterior distribution with the same standard error (17). It is defined as:

$$ESS = \frac{N}{\sum_{k=-\infty}^{\infty} ACF_k} = \frac{N}{1 + 2\sum_{k=1}^{\infty} ACF_k}$$

where ACF_k is the lag k autocorrelation as defined above, and N is the sample size.

It is clear that a good estimate of the posterior distribution would have a low ACF, resulting in a high ESS. We can see this below by calculating the ESS using the `ess` function. The effective sample sizes of the Markov chains corresponding to our parameters are as follows:

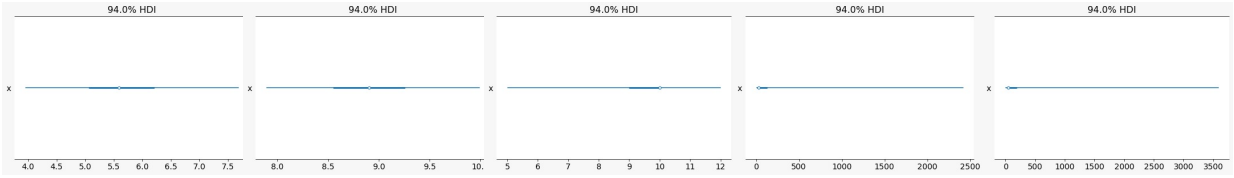
	θ	λ	k	b_1	b_2
Model 1	8025	11762	4250	97555	98117
Model 2	62202	53732	4011	17792	11898

The length of each Markov chain is 100,000, so this is the maximum possible value for effective sample size. We can see that for the chains corresponding to θ and λ, Model 2 significantly increases the ESS which is a good improvement; for the chain corresponding to k, the ESS remains very similar; and for the chains corresponding to b_1 and b_2, the ESS has significantly decreased, which is not an improvement. However, we are more interested in the parameters θ and λ, as these have a more important interpretation in the model (rates of photon releases before and after change point). Therefore overall we can class Model 2 as an improvement on Model 1 with respect to the effective sample size.

5.1.4 Credible Intervals

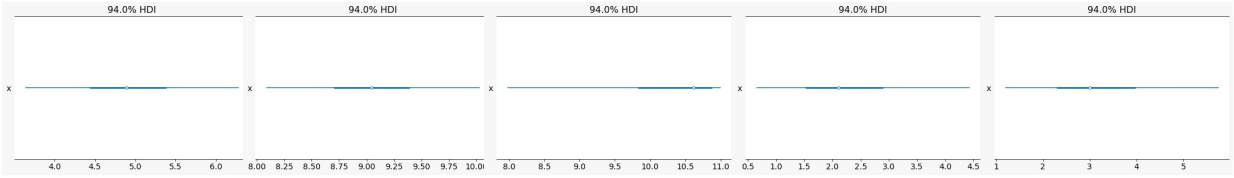
A credible interval is the Bayesian equivalent of the frequentist confidence interval. We can view it as an interval in the domain of the posterior distribution, in which the estimated parameter falls with a particular probability (18). The interval is also dependent on the choice of the prior distribution.

Credible intervals are not unique; here, we use the Highest Density Interval (HDI). This is sometimes referred to as the 'most plausible Bayesian confidence interval' (19), calculated using the `plot_forest` function. The plots are shown for Model 1 in Figure 5.6 and for Model 2 in Figure 5.7.



(a) HDI for θ (b) HDI for λ (c) HDI for k (d) HDI for b_1 (e) HDI for b_2

Figure 5.6: HDI plots for the parameters in our first model



(a) HDI for θ (b) HDI for λ (c) HDI for k (d) HDI for b_1 (e) HDI for b_2

Figure 5.7: HDI plots for the parameters in our second model

you can see that the intervals for θ and λ have remained largely unchanged, apart from a slight shift in location. The plot for k has both shifted and become slightly wider, but the intervals for b_1 and b_2 have narrowed significantly (you can see this by the change of scale) which is an improvement.

5.1.5 Gelman-Rubin statistic

The Gelman-Rubin statistic, also known as the $\hat{R}$ statistic, is another convergence statistic. It compares two estimators for the target variance: one estimator is an overestimate, and the other is an underestimate. As each estimator converges to the required variance, by taking the ratio of the two estimators, the statistic will converge to 1. The statistic should be as close to 1 as possible, with any values larger than 1.1 suggesting lack of convergence (20).

We calculated the $\hat{R}$ statistic using the `rhat` function from the `arviz` package in Python, which computes it as (21):

$$\hat{R} = \frac{\hat{V}}{W}$$

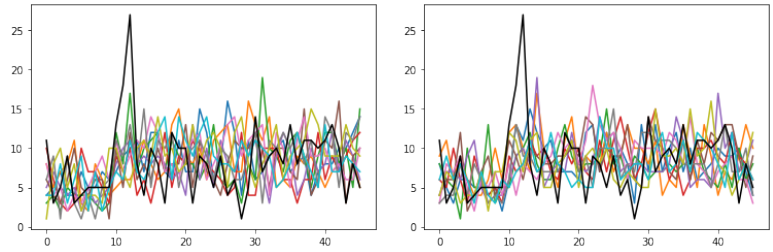
where $\hat{V}$ is the estimate for the posterior variance, and W is the within-chain variance. The results are below:

	θ	λ	k	b_1	b_2
Model 1	1.001	1.000	1.007	1.000	1.000
Model 2	1.000	1.000	1.003	1.001	1.001

Again, the $\hat{R}$ statistic for b_1 and b_2 have increased marginally, but the statistic for θ is now optimally at 1.000, and the statistic for k has also decreased closer to 1, further supporting model 2 being a better fit.

5.2 Goodness of Fit

The first key step in evaluation of MCMC model outputs is to check for model convergence. This is done in the previous sections. It is entirely possible for a unsuitable model to converge, so we need to ensure that the estimated model fits the data adequately. One way of evaluating model fit is to compare model predictions with the observations used to fit the model. This method is outlined by Chris Fonnesbeck. (22). In other words, the fitted model can be used to simulate data, and the distribution of the simulated data should resemble the distribution of the actual data.



(a) plot of goodness for fit for model 1

(b) plot of goodness for fit for model 2

Figure 5.8: goodness of fit plots

For both models, we use the results (λ, θ, k) we get to generate the number of photon before and after the change point. In this case, we generate 10 such lines and compare the result with original data(the black line). We can observe that almost all points are able to be included in the possible region, which means our result is acceptable to express the data. However, the very exceptional point is not being included, which means we need to modify our assumption and models.

Chapter 6

Model Improvements

In this chapter, we will focus on 2 obvious ways that can improve our model – changing our number of turning points from 1 to several, and also introducing the elegant Reversible Jump Markov Chain Monte Carlo methods proposed by Peter J. Green (23) in 1995.

6.1 Increasing the Number of Change Points

6.1.1 Background and methodology

In previous chapters, we have assumed that the number of change points for our data is equal to 1. This might not always be the case and it is therefore not ideal to have such an assumption. To find out the number of change points in our data, we increase it from 1 to i, where $i < 46$, the number of data points that we have. When the two rate parameters of the Poisson distribution are approximately the same between two adjacent change points, then we are done and we can conclude that there are only i change points.

6.1.2 Prior Distributions

In this model, we assume without loss of generality that when there are n change points, θ_i is gamma distributed, b_i is inversely gamma distributed, and k_i is uniformly distributed, where $i = 1, 2, ..., n + 1$.

$$
\begin{aligned}
Y_1^k | k_1, .., k_n, \theta_1, .., \theta_n &\sim \text{Poisson}(\theta_1), \text{for k} = 1, .., k_1 \\
&\vdots \\
Y_{n+1}^k | k_1, .., k_n, \theta_1, .., \theta_n &\sim \text{Poisson}(\theta_{n+1}), \text{for k} = k_n, .., 45 \\
\theta_1 | b_1 &\sim \text{Gamma}(0.5, b_1)(\text{pdf} = g_1(\theta_1 | b_1)) \\
&\vdots \\
\theta_{n+1} | b_{n+1} &\sim \text{Gamma}(0.5, b_{n+1})(\text{pdf} = g_{n+1}(\theta_{n+1} | b_{n+1})) \\
b_1 &\sim \text{Inverse Gamma}(0, 1)(\text{pdf} = h_1(b_1)) \\
&\vdots \\
b_{n+1} &\sim \text{Inverse Gamma}(0, 1)(\text{pdf} = h_n(b_{n+1})) \\
k_1 | k_2 &\sim \text{Uniform}(1, ..., k_2)(\text{pdf} = u_1(k_1 | k_2))
\end{aligned}
$$

$$k_2|k_1, k_3 \sim \text{Uniform}(k_1+1, ..., k_3)(\text{pdf} = u_2(k_2|k_1, k_3))$$
$$\vdots$$
$$k_n|k_{n-1} \sim \text{Uniform}(k_{n-1}+1, ..., 45)(\text{pdf} = u_n(k_n|k_{n-1}))$$

The reason why we choose b_1, b_2, ... to be Inversely Gamma distributed is because the full conditional probability distribution is known, as shown in the previous chapter.

6.1.3 Posterior Distributions

Our goal is to draw samples from the posterior distribution when the number of change points is equal to n. The posterior distribution is:

$$f(\mathbf{k}, \theta, \mathbf{b}|\mathbf{Y}) \propto \prod_{i=1}^{k_1} \frac{\theta_1^{Y_i^1} e^{-\theta_1}}{Y_i^1!} \prod_{i=k_1+1}^{k_2} \frac{\theta_2^{Y_i^2} e^{-\theta_2}}{Y_i^2!} \cdots \prod_{i=k_n+1}^{45} \frac{\theta_{n+1}^{Y_i^{n+1}} e^{-\theta_{n+1}}}{Y_i^{n+1}!} \times \prod_{i=1}^{n+1} \frac{1}{\Gamma(0.5) b_i^{0.5}} \theta^{-0.5} e^{-\theta_i/b_i}$$
$$\times \prod_{i=1}^{n+1} \frac{1}{\Gamma(0.01) 100^{0.01}} b_i^{-0.99} e^{-b_i/100} \times \frac{1}{k_2 - 1} \times \frac{1}{k_3 - k_1} \cdots \times \frac{1}{45 - k_{n-1}}$$

Similar to in our earlier models, using the above expression, we are able to obtain marginal full conditional distributions for each parameter by ignoring all terms that are constant with respect to the parameter.

This gives us the following full conditional distribution for θ_i and b_i, for $i = 1, 2, .., n+1$:

$$f(\theta_i|\theta_{-i}, \mathbf{k}, \mathbf{b}, \mathbf{Y}) \propto \text{Gamma}\left(\sum_{i=k_i-1}^{k} {}_iY_i + 0.5, \frac{b_i}{(k_i - k_{i-1}) * b_i + 1}\right)$$

$$f(b_i|\mathbf{k}, \theta, b_{-i}, \mathbf{Y}) \propto \frac{1}{b_i^{0.5}} e^{-\theta/b_i} \times \frac{e^{-1/b_i}}{b_i} \propto b_i^{-1.5} e^{-(1+\theta_i)/b_1} \propto \text{IG}(0.5, \frac{1}{\theta_i + 1})$$

and the full conditional for k is

$$f(k_j|\mathbf{k_{-j}}, \theta, \mathbf{b}, \mathbf{Y}) \propto \prod_{i=k_{j-1}+1}^{k_j} \frac{\theta_j^{Y_i^j} e^{-\theta_j}}{Y_i^j!} \prod_{i=k_j+1}^{k_{j+1}} \frac{\theta_{j+1}^{Y_i^{j+1}} e^{-\theta_{j+1}}}{Y_i^j!} \times \frac{1}{k_j - k_{j-2}} \times \frac{1}{k_{j+2} - k_j}$$

6.1.4 Sampler Methods

We will again use a single component Metropolis-Hastings algorithm on our change point parameters k_i, and for other parameters we will use the Gibbs sampler. This time we can use a Gibbs update on θ_i and b_i and Metropolis-Hastings update on k_i.

6.1.5 Plots and Results

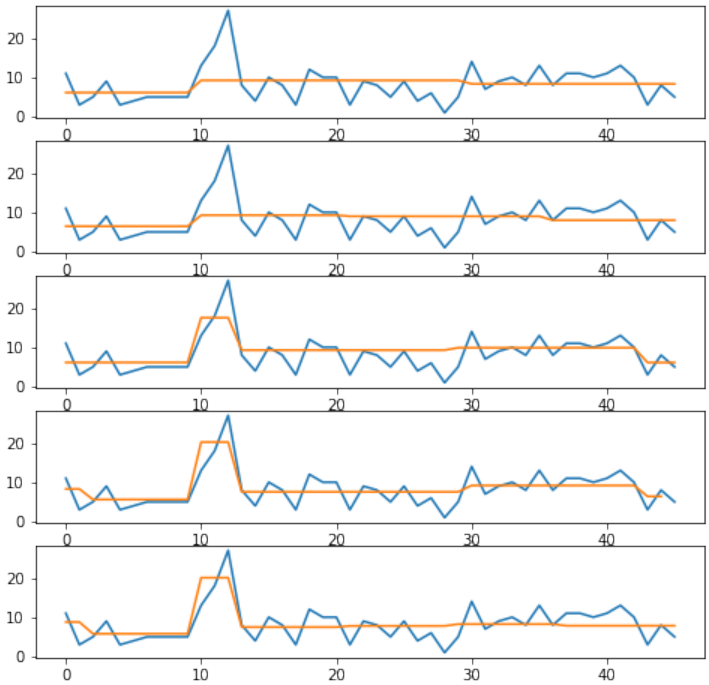


Figure 6.1: the plot compares models with different change points

Number of change points	Values of θ_i	Values of k_i
2	6.17, 9.09, 8.38	8.76, 27.24
3	6.45, 9.07, 8.86, 8.03	8.05, 21.77, 31.94
4	5.95, 18.31, 8.48, 10.01, 6.06	9.44, 12.74, 28.84, 41.52
5	9.16, 5.47, 19.65, 7.54, 8.96, 6.68	2.61, 10.14, 14.13, 28.16, 40.44
6	9.60, 4.97, 19.65, 7.47, 7.71, 8.73, 7.92	2.02, 10.09, 13.02, 21.55, 29.31, 37.02

6.1.6 Explanation of the plots and results

In Figure 6.1, we compare models with different number of change points (from top to bottom is 2, 3, 4, 5, 6 change points respectively). In each graph, we plot the original data, which is the blue line, and the expectation of the θ_i between k_{i-1} and k_i, which is the orange line, by calculating the mean value of the θ_i. In the table above, we calculate the mean of the last 500-1000 samples for each θ_i and k_i in order to get more precise view of the result.

From these figures, we can observe that as the number of change points increases from 2 to 5, the goodness of fit improves. However, when the number of change points is equal to 5, we observe that there is a change point at around $t = 2$. This phenomenon may imply that the first change point is a trivial change point, since this change point separates the first 2 points, and we can observe θ_1 is mainly fitted for the first data point and the second data point is about 7 photon less than the first data point and more similar to the data point from $t = 3$ to $t = 9$. Therefore, there is a large chance that the first data point is over-fitted. However, if we check the case when the number of change points is equal to 6, this "trivial" change point still exists and there is only one more change point with left and right hand side θ very close(implying this is not a change point).

Moreover, observe that when number of change points = 6, the two values of θ (i.e. θ_4 and θ_5) are approximately the same, at around 7.5. This provides further evidence that when the number of change points $\geqslant$ 6, the model is no longer suitable.

Therefore, the above analysis suggests the following conclusion:
Number of change points = 4 or 5 might be the best model in this case. It would be dangerous to quickly conclude that 4 change points is the right answer as there is no strict proof that the model with 4 change points is fitted better than the 5^{th} one. It is also not suitable to come to a conclusion quickly that a certain model is better than the others by only using the above algorithm.
In the following section, we will examine the Reversible Jump Markov Chain Monte Carlo and see what results can we deduce there.

6.2 Reversible Jump Markov Chain Monte Carlo

Peter J. Green (1995) (23) developed a Bayesian multiple change point analysis of the serious coal-mining disaster between 1851 and 1962, given by Raftery and Akman (1986) (24), namely the Reversible Jump Markov Chain Monte Carlo (RJMCMC) method. Jialin Ai (2012) (25) further explores the application of this algorithm in a multiple change point problem. In this section, we will focus on looking at how RJMCMC can be used on our data, especially with Green's method of jumping between different dimensions. We will introduce the basic model framework in the following sections.

6.2.1 Poisson Process

We start RJMCMC by providing a general definition of the Poisson Process.
Definition 1. (26)
A point process is called a Poisson Process with intensity function λ, if
$i.$ $N(t)$ has independent increments;
$ii.$ $N(b) - N(a)$ is Poisson distributed, with rate= $\int_a^b \lambda(t)\, dt$

Unlike previous chapters, here we generalise the definition of the Poisson Process, suggested by Grandell(1997) (26), where λ is treated as a function.

6.2.2 Problem setting and Defining the Random Variables

In this model, we assume the rate function to be a step function $\lambda : \mathbb{R} \to \mathbb{R}_{>0}$ on [s_1,s_{k+1}], where s_1,s_2,...s_{k+1} are the change points. Clearly in our data set, s_1 is fixed to be 0 and s_{k+1} is fixed to be 46. Moreover, define h_j to be the jth piece of $\lambda(.)$, which is a constant function on the interval [s_j,s_{j+1}).

Here in the model, the probability of k counts occur in the interval [a,b] $\subseteq$ [s_1, s_{k+1}] is:

$$Pr(the\ number\ of\ counts\ y_i \in [a,b] = k) = exp(-\Lambda_{(a,b)}) \frac{\Lambda^m_{(a,b)}}{m!}$$

here $\Lambda_{(a,b)} = \int_a^b \lambda(t)\,dt$ because $\lambda(t)$ is now a continuous function.

We now proceed to compute the likelihood function given λ. First note that the individual points are independently distributed with density $\frac{\lambda}{\Lambda_{(a,b)}}$. The likelihood is then:

$$Pr(y_i \in [a,b]|\lambda) = m! \times exp(-\Lambda_{(a,b)})\frac{\Lambda_{(a,b)}^m}{m!} \times \prod_{i=1}^{m}\frac{\lambda(y_i)}{\Lambda_{(a,b)}}$$

where $m!$ is the number of arrangements when m points are ordered.

Consider the interval $[s_1,\ s_{k+1}]$, which there are n photon counts in total. After some simplifications, we obtain:

$$\begin{aligned} Pr(y_i \in [s_1, s_{k+1}]|\lambda) &= n! \times exp(-\Lambda)\frac{\Lambda^n}{n!} \times \prod_{i=1}^{n}\frac{\lambda(y_i)}{\Lambda} \\ &= exp(-\Lambda + \sum_{i=1}^{n}(number\ of\ y_i \in [s_j, s_{j-1})\ log\lambda(y_i)) \end{aligned}$$

where $\Lambda = \int_{s_1}^{s_{k+1}} \lambda(t)\,dt$. Since λ is a step function, $\Lambda = \sum_{j=1}^{k} h_j(s_{j+1} - s_j)$.

Therefore, the likelihood function is:

$$Pr(y_i \in [s_1, s_{k+1}]|\lambda) = exp(-\sum_{j=1}^{k} h_j(s_{j+1} - s_j) + \sum_{i=1}^{n}(number\ of\ y_i \in [s_j, s_{j-1})\ log\lambda(y_i)) \tag{6.1}$$

In this current model above, there are three types of random variables:

1. **The number of steps** k
   Unlike above, here we define k as the number of steps instead of the number of change points. Therefore, the number of change points here in this model is $k-1$.

   Here we suppose that we have k draws from the Poisson Distribution, with the prior probability being below, conditioning on k, with $k \leq k_{max}$.

   $$p(k) = exp(-\lambda)\frac{\lambda^{k-1}}{(k-1)!}$$

2. **The height** h_j**,** $j \in [1,k]$
   According to Green, the heights h_1,...,h_l are Gamma distributed. Green stated that it is not appropriate to select an improper gamma distribution $\Gamma(0,0)$. Here we assume the values of α and β to be 5 and 1.25 respectively. We chose this because there will be an acceptable variance under this setting.

3. **The step** s_j, $j \in [2, k]$
   As stated above, s_1 and s_{k+1} are fixed to be 0 and 46 respectively and s_j denotes the changing points.

Green states that if λ is a step function, there are 4 types of transitions:

1. A change to the height of a randomly chosen step.
2. A change to the position of a randomly chosen step.
3. "Birth" of a new step between a randomly chosen location in[s_1,s_{k+1}]
4. "Death" of a new step between a randomly chosen location in [s_1, s_{k+1}]

At each transition, an independent random choice is made between attempting each of the most four available move types (H, P, k, $k-1$), representing height change, position change, birth or death respectively. We denote the transition probability to be η_k for H, π_k for P, b_k for k, d_k for $k-1$, and $\eta_k + \pi_k + b_k + d_k = 1$.

6.2.3 Change of height of a randomly chosen step

Parameter: h_j, where $h_j \in (0, \infty)$ and $j \in (1, k)$

Prior Distribution of h_1, h_2,...,h_k: They are assumed to be independently drawn from the Gamma(α,β) distribution.

Data: y_i, $i = 1, ..., n \in$ [s_1,s_{k+1}]is a Poisson Process, described above. h_j is the jth piece of $\lambda(.)$, which is a constant function on the interval [s_j,s_{j+1}). m_j is the number of photons counts recorded on the interval [s_j,s_{j+1}). Using 6.1:

$$p(y_1, ..., y_n|h_j) = exp\left(-\sum_{j=1}^{k} h_j(s_{j+1} - s_j) + \sum_{j=1}^{k} m_j + logh_j\right) \tag{6.2}$$

Green chooses h_j at random, and proposes a change to $\widetilde{h_j}$ such that log($\frac{\widetilde{h_j}}{h_j}$) is uniformly distributed on the interval [$-\frac{1}{2}$,$\frac{1}{2}$]. Here, the transition density of the proposing is q(h_j,$\widetilde{h_j}$) = $\frac{1}{\widetilde{h_j}}$.

The acceptance probability for this move is found, by Green, to be:

$$min\left[1, \left((likelihood\ ratio) \times \left(\frac{\widetilde{h_j}}{h_j}\right)^{\alpha} e^{-\beta(\widetilde{h_j} - h_j)}\right)\right]$$

and the explicit formula for the log form of this acceptance probability is found by Ai in his report on page 30 (25).

Metropolis Hastings Algorithm for height move

Here we require the input to be the target density, transition density and initial value. The target density is $\pi(h_j)$, which is proportional to $p(y_1, y_2, ..., y_n|h_j)p(h_j)$ and the transition density is $q(h_j, \widetilde{h_j})$, which is equal to $\frac{1}{\widetilde{h_j}}$. Set initial value h_j^0 to be 1 under this algorithm.

For the output, for $n = 1, 2, 3, ...$, we first generate $\widetilde{h_j^n}$ by log$\widetilde{h_j^n}$ = logh_j^{n-1}+ u_n and generate U_n from the distribution Uniform$[0, 1]$. If $U_n \leq \alpha(h_j^{n-1}, \widetilde{h_j^n})$, where α(h_j^{n-1},$\widetilde{h_j^n}$) is the acceptance probability, then we set h_j^n to be $\widetilde{h_j^n}$, else we reject the value.

6.2.4 Change of position of a randomly chosen step

Parameter: s_j, where $s_j \in (s_{j-1}, s_{j+1})$ and $j \in (2, k)$

Prior Distribution of s_2, s_3,...,s_k: They are assumed to be distributed as the even numbered order statistics from $(2k-1)$ points $t_i, i = 1, 2, ..., 2k-1$ uniformly distributed on (s_1, s_{k+1}). That is, s_2=t_2, s_3=t_4,...,s_j=t_{2j-2},...,s_k=t_{2k-2}, given that the value of k is known.

$$\pi(s_2, ..., s_k|k) = \frac{(2k-1)!}{(s_{k+1} - s_1)^{2k-1}} \mathbb{1}_{s_1<s_2<...<s_k<s_{k+1}} \prod_{j=1}^{k}(s_{j+1} - s_j)$$

Data: This will be the same as 6.2
The acceptance probability for this move is found, by Green, to be:

$$min\left[1, \left((likelihood\ ratio) \times \left(\frac{(s_{j+1} - \widetilde{s_j})(\widetilde{s_j} - s_{j-1})}{(s_{j+1} - s_j)(s_j - s_{j-1})}\right)\right)\right]$$

and the explicit formula for the log form of this acceptance probability is found by Ai in his report on page 34 (25).

Metropolis Hastings Algorithm for position move

Similar to the height move algorithm, we also require the target density, transition density and an initial value as the input. Here, target density is $\pi(s_j|s_2, ..., s_{j-1}, s_{j+1}, ..., s_k)$, which is proportional to $p(y_1, y_2, ..., y_n|s_j)p(s_j|s_2, ..., s_{j-1}, s_{j+1}, .., s_k)$ and the transition density is $q(s_j, \widetilde{s_j})$, which equals $\left(\frac{1}{s_{j+1} - s_{j-1}}\right)$. Set the initial value to be s_j^0, where $s_j^0 \in (s_{j-1}, s_{j+1})$

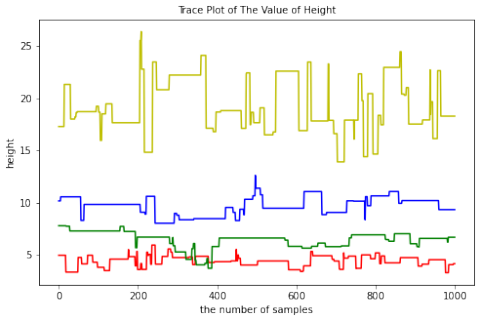
For the output, for $n = 1, 2, 3, ...$, we generate $\widetilde{s_j^n} \sim$ from the distribution Uniform(s_{j-1}, s_{j+1}) and generate U_n from the distribution Uniform$[0, 1]$. If $U_n \leq \alpha(s_j^{n-1}, \widetilde{s_j^n})$, where α(s_j^{n-1},$\widetilde{s_j^n}$) is the acceptance probability, then we accept $\widetilde{s_j^n}$, else $s_j^n = s_j^{n-1}$

6.2.5 Graphs and results for fixing number of steps

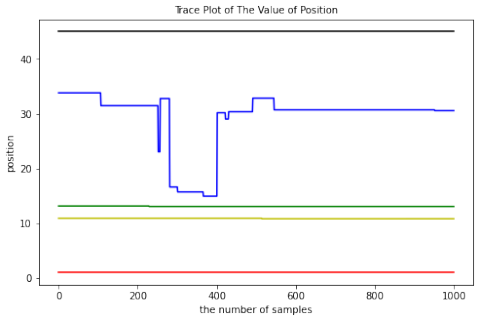
Here we fixed k. With the height and position algorithm, we can plot the graphs and see where the location and height of each changepoint. In the following graphs, we fix k = 4,5,6 respectively.

There will be 4 types of graphs that we will be plotting here. The first will be the path of simulation for heights. The second graph will be the densities for heights. The third one will be the path of simulation for steps and the last one will be the densities for steps.

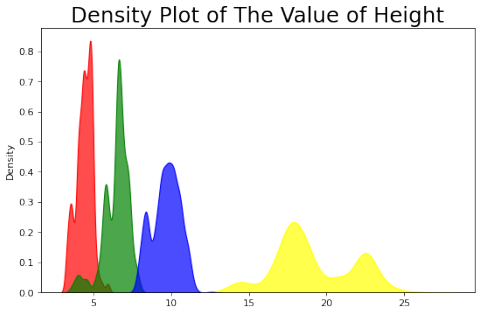
We have also taken the mean of the last 1,000 samples out of the 1,000,000 samples that have been generated using the above algorithm and calculated the mean of the heights and positions.



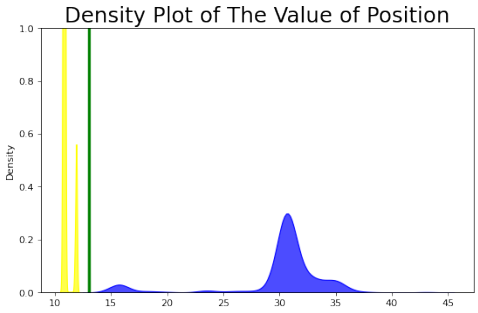
(a) Height trace plot



(b) Position trace plot

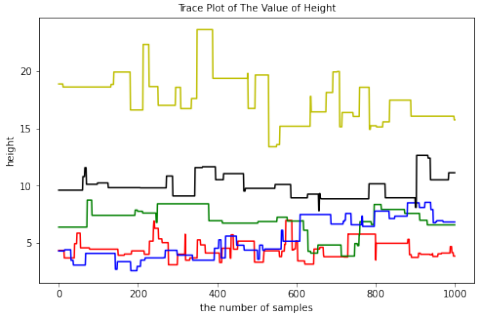


(c) Height density plot

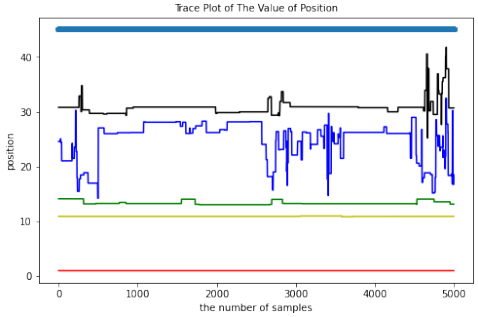


(d) Position density plot

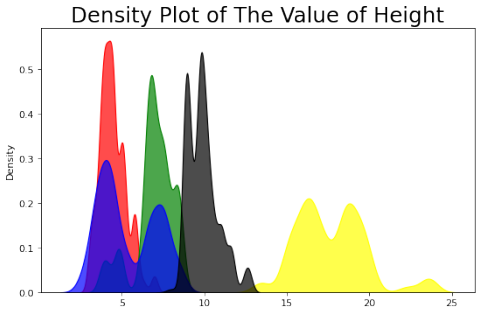
Figure 6.2: Fixing $k = 4$, i.e. number of change points $= k - 1 = 3$



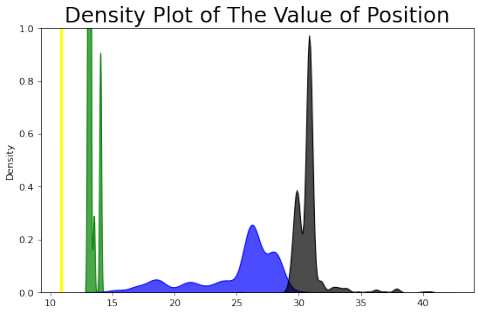
(a) Height trace plot



(b) Position trace plot

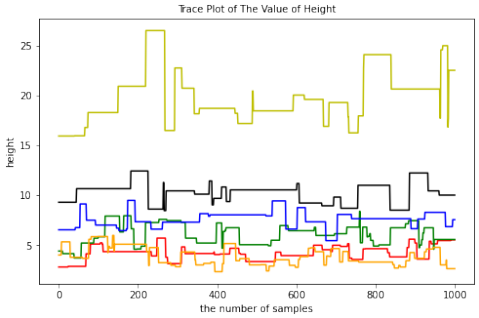


(c) Height density plot

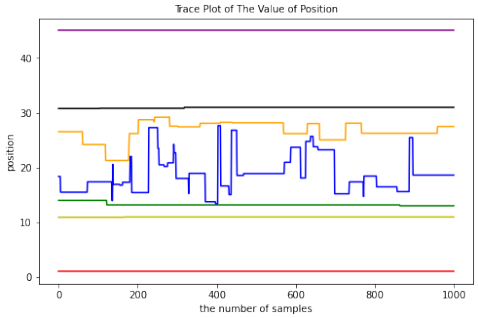


(d) Position density plot

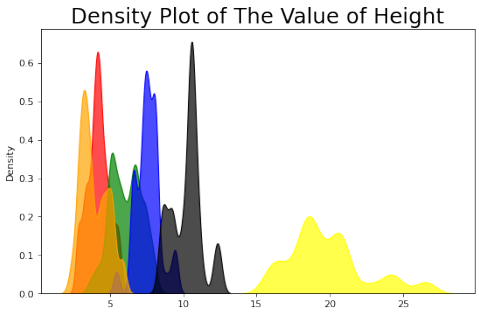
Figure 6.3: Fixing $k = 5$, i.e. number of change points $= k - 1 = 4$



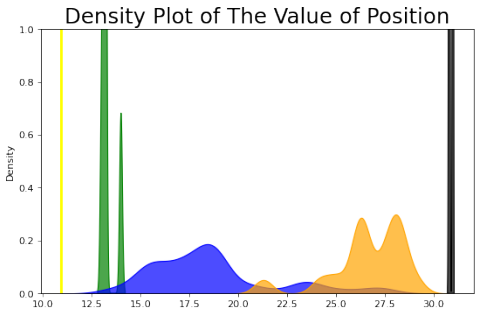
(a) Height trace plot



(b) Position trace plot



(c) Height density plot



(d) Position density plot

Figure 6.4: Fixing $k = 6$, i.e. number of change points $= k - 1 = 5$

Number of change points	Values of the Heights (h)	Values of Positions (s)
3	5.09, 22.12, 6.60, 9.65	10.80, 13.05, 26.73
4	4.50, 20.40, 7.55, 3.59, 9.94	10.80, 13.05, 26.73, 30.18
5	4.21, 19.82, 6.09, 7.58, 3.90, 10.25	10.97, 13.26, 18.63, 26.66, 30.93

6.2.6 Explanation of the graphs and results

1. Position trace and density plots:
Observe the density plot at $k = 4, 5, 6$. The yellow and green curve, corresponding to the position (s) at around 11 and 13 respectively, shows a sharp peak. This indicates that most of the time, the s_1 and s_2 sit at approximately 11 and 13 respectively, which shows the importance of these 2 positions . For the density plot at $k = 5, 6$, the black curve, corresponding to the position at around 31, also shows a sharp peak. This indicates the importance of this position as well. The blue curve, which corresponds to the position at around 26, shows a less sharp peak and therefore indicates that there are certain fluctuations around this position; similarly for the orange curve at $k = 5$.

2. Height trace and density plots:
Observe the density plot at $k = 4, 5, 6$. The yellow curve, which corresponds to the height (h) located at position $s = 11$ is not as sharp as the others, indicating a fluctuation for the height at this point. The green curve, which corresponds to the h at around $s = 11$ gets smoother as k increases. The blue curve, which corresponds to the h located at around $s = 26$, gets sharper and sharper as k increases. Lastly, the black curve, which corresponds to the h located at approximately $s = 31$ maintains its sharpness when k changes.

3. A small conclusion from the above plots
From the plots and analysis above, it is quite clear that $s = 11$, 13 and 29 are important changing positions even though the height density trace plots are smooth for $s = 11$. The model suggests there is a position located at $s = 29$ only when $k \geqslant 5$. Given the significance of this point, we will therefore conclude that k has to be larger than 4. This matches the result from the previous model, which indicates that k is either 5 or 6, i.e.the number of change points is either 4 or 5 for our data.

6.2.7 Birth of a new step

The algorithm for a birth of a step is much more complicated. Since the dimension of the subspace is varying, besides looking at parameters, prior distributions, and data, we will also look at the subspaces in the model with different dimensions.

Subspaces: Define $\mathcal{C}_k = k \times \mathbb{R}^{2k+1}$, where $k \in \mathbb{N}^+$ and $1 \leq k \leq k_{max}$. k is drawn from the Poisson distribution similarly to above, conditioned on k, where $p(k)$ is:

$$p(k) = e^{-\lambda} \frac{\lambda^{k-1}}{(k-1)!}$$

Parameter: $x \in \mathcal{C}_k$, where $x = (k, h_1, ..., h_j, ..., h_k, s_2, ..., s_j, s_{j+1}, ..., s_k)$. Rewrite $\theta^{(k)} = (h_1, ..., h_j, ..., h_k, s_2, ..., s_j, s_{j+1}, ..., s_k)$, so $x = (k, \theta^{(k)})$

Prior:

$$p(x) = p(k) \times \frac{(2k-1)!}{(s_{k+1}-s_1)^{2k-1}}(s_2-s_1)...(s_{j+1}-s_j)...(s_{k+1}-s_k) \times \prod_{j=1}^{k} \frac{\beta^\alpha h_j^{\alpha-1} e^{-\beta h_j}}{\Gamma(\alpha)}$$

Data: This will be the same as 6.2

Here in the birth step, we define $x \in \mathcal{C}_k$ and $\widetilde{x} \in \mathcal{C}_{k+1}$. According to Green, we first choose a position s^* for the proposed new step, uniformly distributed on [0,L]. Here in our data, L=46. Clearly, this lies within an existing interval (s_j, s_{j+1}). If accepted, $s_{j+1}^{'}$ will be set to s^*, and s_{j+1},s_{j+2},... s_k will be relabelled as $s_{j+2}^{'}$, $s_{j+3}^{'}$,..., $s_{k+1}^{'}$ with corresponding changes to the labelling of step heights. We will need to propose new heights $h_j^{'}$, $h_{j+1}^{'}$ for the step function on the subintervals (s_j, s^*) and (s^*, s_{j+1}). The heights therefore, are perturbed in either direction from h_j in such a way that h_j is a compromise between them. Green proposed a weighted mean for this compromise and also defined the perturbation in his paper (23).

After some derivation by Green, the acceptance probability for this move is found to be

$$min[1, (likelihood\ ratio) \times (prior\ ratio) \times (proposal\ ratio) \times (Jacobian)]$$

where prior ratio = $\frac{p(x)}{p(x)}$, proposal ratio = $\frac{d_k(s_{k+1}-s_1)}{b_{k-1}k}$ and Jacobian = $\frac{(h_j^{'}+h_{j+1}^{'})^2}{h_j}$. (Recall b_{k-1} is the probability of moving from $\mathcal{C}_k$ to $\mathcal{C}_{k+1}$ and d_k is the probability of moving from $\mathcal{C}_{k+1}$ to $\mathcal{C}_k$).

The explicit formula for the log form of this acceptance probability is found by Ai in his report on pages 54-55. (25)

Metropolis Hastings Algorithm for step move

In this algorithm, we again need the target density, which is $\pi(k)$ and is proportional to $p(y_1, y_2, ..., y_n|x)p(x)$. The transition density is $q(x, \widetilde{x}) = q(u) = 1$. Here, we set our initial value to be $x^0 \in \mathcal{C}_k$.

For the output, for $n = 1, 2, 3, ...$, first generate s^*, which is the new step from the distribution Uniform(s_j, s_{j+1}) and also generate u from the distribution Uniform$[0, 1]$. Then, calculate $h_j^{'}, h_{j+1}^{'}$ for the step function on the sub-intervals (s_j, s^*) and (s^*, s_j), using the definition of perturbation provided by Green. Generate U_n from the distribution Uniform$[0, 1]$. If U_n is smaller than $\alpha(x, \widetilde{x})$, where $\alpha(x, \widetilde{x})$ is the acceptance probability defined above, then set $x^n = \widetilde{x^n}$, else $x^n = x^{n-1}$.

6.2.8 Death of a new step

The algorithm and acceptance probability for the death of a new step will be very similar to the birth of a new step. They can be calculated by changing the variables appropriately and inverting the ratio terms of the acceptance probability suggested by Green.

6.2.9 Graphs and results for varying number of steps

With the height, position, birth and death acceptance algorithm ready, we can now determine a model with a flexible number of steps. Here we assume the parameter λ of the prior distribution of k to be 3. Figure 6.5 shows the plot of the posterior distribution of k, (frequency against k) after generating 1,000,000 samples using the above algorithm.

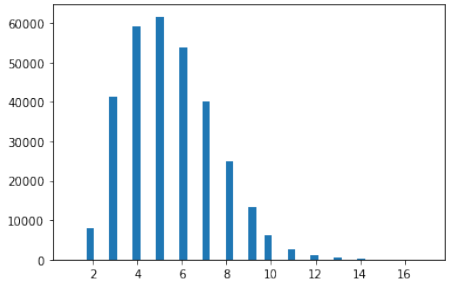


Figure 6.5: Posterior distribution of k (the number of steps)

From the graph, the bar chart achieves the maximum point at $k = 5$. One might suggest that the number of steps in our model is 5. However, at $k = 4$ and 6, it shows a similar height as $k = 5$. The graph therefore does not show clear evidence that $k = 5$.

6.2.10 Conclusion

Number of change points	Values of θ_i	Values of changepoints
4 (model 1)	5.95, 18.31, 8.48, 10.01, 6.06	9.44, 12.74, 28.84, 41.52
5 (model 1)	9.16, 5.47, 19.65, 7.54, 8.96, 6.68	2.61, 10.14, 14.13, 28.16, 40.44
4 (model 2)	4.50, 20.40, 7.55, 3.59, 9.94	10.80, 13.05, 26.73, 30.18
5 (model 2)	4.21, 19.82, 6.09, 7.58, 3.90, 10.25	10.97, 13.26, 18.63, 26.66, 30.93

To summarise, the first and second models in this chapter do not show a precise indication of whether the number of steps k should be 5 or 6 for our data. We therefore conclude that the number of change points is either 4 or 5 for the data given.

Moreover, from the analysis above, we are confident to conclude that the important changepoints are located at around 10, 13, and 29. This is supported by both model 1 and model 2. Looking at the table above (model 1 data), it is observed that these changepoints do not have a big fluctuation as we increase our dimension of the model. In model 2, the sharp peak of the curves representing these points in the density plots further suggest the importance of them. Other changepoints, however, fluctuate as the dimension increase in model 1. The smoothness of the density curves of other points also suggests that they are less important in our model. We therefore do not have full confidence to conclude that other changepoints, suggested in either model 1 or 2, do exist.

Now we look at the rate parameters between the changepoints, assuming that there are no changepoints before the point 10. The rate parameter is found to be around 5 before the changepoint 10. The value 5 is deduced by taking the average of the values of the 2 models. This supports the argument in the first two models in chapter 4. Assuming there are no changepoints in between 10 and 13, the value of the rate parameter is found to be around 19. Between the last two points, since different models suggest that some turning points might exist in between them, it is therefore hard to conclude what the exact value of the rate parameter is.

Chapter 7

Conclusion

Throughout the whole report, we have looked at how the Markov Chain Monte Carlo methods can be applied to our changepoint detection problem. In the first two models, we assume the number of changepoint is equal to 1. Using MCMC, we find that the changepoint is located at around 10, which matches our expectation at the first glance of the data. The values of the rate parameter λ of the Poisson distribution is also found to be approximately 5.3 and 9.0 before and after the changepoint. In the model justification part, we check our model using different statistical tools and are confident that our model fits the data well based on the tools that we introduced in Chapter 5.

One of the most obvious ways to improve our model is to relax the assumption that the number of changepoints is 1. After relaxing this assumption, we looked at two different models to see where the turning points occur. We look at the multiple changepoint model and used the Reversible Jump Markov Chain Monte Carlo method to investigate our data. While there is not an exact number of the changepoints in our data, we have identified 3 important turning points using the two algorithms introduced above. The 3 important changepoints that we observed are located at around time bins 10, 13 and 29. We are also confident to conclude that the total number of changepoints is either 4 or 5 for our data based on the reasoning illustrated above. Combining the results from different improved models, we also have a brief idea of what the rate parameters are between the important changepoints.

To clearly identify what the exact number of change points are and where they are located, we might need to use other changepoint detection techniques such as the Bayesian online detection algorithm, deep learning and neural networks to further investigate our data. Changepoint detection therefore remains a big topic to discover and explore in statistics.

Bibliography

[1] D. B. Dunson, "Commentary: Practical advantages of bayesian analysis of epidemiologic data," *American Journal of Epidemiology*, vol. 153, no. 12, p. 1222–1226, 2001. pages 2

[2] B. Junker, "Basics of bayesian statistics," 2020-2021. pages 2, 3

[3] C. Hallsworth, "Probability for statistics," 2020-2021. pages 4

[4] J. H. Halton, "A retrospective and prospective survey of the monte carlo method," *SIAM Review*, vol. 12, no. 1, p. 2, 1970. pages 4

[5] A. M. Johansen, "Monte carlo methods," in *International Encyclopedia of Education* (P. Peterson, E. Baker, and B. McGaw, eds.), pp. 296–303, Elsevier Science, 2010. pages 4

[6] A. B. Owen, "Monte carlo theory, methods and examples." unpublished book available online, 2013. pages 4

[7] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *Markov Chain Monte Carlo in Practice*. Springer Science+Business Media Dordrecht, 1996. pages 4, 5, 8

[8] C. J. Geyer, "Practical markov chain monte carlo," *Statistical Science*, vol. 7, no. 4, pp. 473–483, 1992. pages 5

[9] B. Western and M. Kleykamp, "A bayesian change point model for historical time series analysis," *Political Analysis*, vol. 12, no. 4, pp. 354–374, 2004. pages 6

[10] X. C. Abrevaya and B. C. Thomas, "Radiation as a constraint for life in the universe," in *Habitability of the Universe Before Earth* (R. Gordon and A. A. Sharov, eds.), pp. 27–46, Academic Press, 2018. pages 6

[11] M. Haran, "A markov chain monte carlo example," 2007. pages 7, 9, 11

[12] B. Walsh, "Markov chain monte carlo and gibbs sampling," 2004. pages 9

[13] M. Haran, "Bayesian change point model: full conditional distributions," 2007. pages 10

[14] M. Haran, "A bayesian change point model," 2007. pages 13

[15] R. E. Kass, B. P. Carlin, A. Gelman, and R. M. Neal, "Markov chain monte carlo in practice: A roundtable discussion," *The American Statistician*, vol. 52, no. 2, p. 99, 1998. pages 17

[16] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*. Prentice-Hall International, 1994. pages 17

[17] L. Gong and J. M. Flegal, "A practical sequential stopping rule for high-dimensional mcmc and its application to spatial-temporal bayesian models," 2014. pages 18

[18] U. of Colorado, "Computational bayesian statistics: Credible intervals," Unknown. pages 19

[19] Unknown, "Bayesian inference," Unknown. pages 19

[20] D. Vats and C. Knudson, "Revisiting the gelman-rubin diagnostic," Master's thesis, 2020. pages 19

[21] Unknown, "Diagnostics: arviz.rhat," Unknown. pages 19

[22] C. Fonnesbeck, "Model checking," 2019. pages 20

[23] P. J. Green, "Reversible jump markov chain monte carlo computation and bayesian model determination," Master's thesis, 1995. pages 21, 24, 31

[24] A. E. Raftery and V. E. Akman, "Bayesian analysis of a poisson process with a change-point," *Biometrika*, vol. 73, no. 1, p. 5, 1986. pages 24

[25] J. Ai, "Reversible jump markov chain monte carlo," Master's thesis, 2012. pages 24, 26, 27, 31

[26] J. Grandell, *Mixed Poisson Process*. Chapman Hall, 1997. pages 24

Chapter 8

Appendix

8.1 Python Code

The code we used for this project can be found at the following link: Our Python Code.

8.2 Presentation

The video of the presentation for our project can be found at the following link: Our Presentation.