BAYESIAN CHANGEPOINT DETECTION IN SOLAR ACTIVITY DATA

James Grant - 2090368

Supervisors - Dr Vladislav Vyshemirsky & Dr Vincent Macaulay

A dissertation submitted to the School of Mathematics and Statistics of the University of Glasgow for the degree of Master of Research

September 2014

Abstract

This project provides an extension to Adams & MacKay (2007)'s existing algorithm for the Bayesian online detection changepoints by introducing Markov Chain Monte Carlo steps. The extension allows the existing algorithm to be applied to data from distributions lacking an associated conjugate prior distribution, by using Monte Carlo integration to estimate probabilities from posterior predictive distributions. The extended algorithm was applied to solar activity data as an example of its new potential and evidence of distinct changepoints was discovered around the years 1790 and 1830.

Acknowledgements

I would like to thank my supervisors Dr Vladislav Vyshemirsky and Dr Vincent Macaulay for the attention and guidance they have both given me throughout this dissertation project. It has been a pleasure to work with them.

Contents

1 Introduction **1**

2 Methods **4**

2.1 Changepoint Data: Concepts and Notation 4

2.2 Review of Literature 7

2.3 Adams and MacKay's Bayesian Online Changepoint Detection Algorithm . 8

2.4 Replicating Adams and MacKay's Algorithm 10

2.4.1 Gaussian Data with Unknown Mean and Known Variance 11

2.4.2 Gaussian Data with Known Mean and Unknown Variance 15

2.4.3 Gaussian Data with Unknown Mean and Unknown Variance 17

2.4.4 Real Datasets 18

2.5 Introducing Markov Chain Monte Carlo steps 19

2.6 Harmonic Data 22

2.6.1 Angular Frequency Known 23

2.6.2 Angular Frequency Unknown 24

2.7 Pre-processing of Sunspot Data 26

3 Results **28**

3.1 Replicating Adams and MacKay's Algorithm 28

3.1.1 Gaussian Data with Unknown Mean and Known Variance 28

3.1.2 Gaussian Data with Known Mean and Unknown Variance 30

3.1.3 Gaussian Data with Unknown Mean and Unknown Variance 30

3.1.4 Real Datasets 33

3.2 Introducing Markov Chain Monte Carlo Steps 37

3.3 Harmonic Data 39

3.3.1 Angular Frequency Known 39

3.3.2 Angular Frequency Unknown 41

3.4 Sunspot Data 44

4 Discussion **47**

1 Introduction

Sunspots are dark patches which have been observed on the surface of the sun by astronomers since the invention of the astronomical telescope in the early 17th Century and their numbers have continued to be recorded ever since. Observing their patterns has taught physicists much about the sun, from understanding the mechanics of its rotation to understanding its electromagnetic activity, as this is the underlying phenomenon to which the presence of sunspots can be accredited (Macaulay, 1992).

After over two centuries of observing sunspots, a regular cycle in their numbers was observed. Wolf identified a link between this cycle and geomagnetic activity and in 1849 developed a *Sunspot Index*, which is still used to this day as a common means of quantifying sunspot levels. The index R, is a function of the number of observed spots f, the number of groups of spots g and a scale factor k (which carries with it some degree of subjectivity) and is given by

$$R = k(10g + f).$$

(Sonnet, 1983)

Wolf Sunspot Index values have continued to be calculated since the creation of the index and historical observations have been used to infer values as far in the past as 1700, although there is a belief that the historical values are less accurate because they are a reconstruction based on data whose daily record had some discontinuities. The accuracy of early records is also somewhat questionable. When monthly averages of Wolf's Sunspot Index are plotted against time, as in Figure 1.1 (WDC-SILSO, 2014), the cycle is very clear (on a daily or weekly scale the index fluctuates too wildly for a pattern to be so recognisable).

Another interesting feature of the data, besides its cyclicity, is the region around 1790-1830 where there is a very noticeable deviation from the otherwise fairly regular pattern. This period has been historically referred to as the *Great Solar Anomaly* or the *Dalton Minimum.* There is a noticeable change in both the amplitude and phase of the cycle during this period and afterwards the pattern seems to return to normal (Sonnet, 1983). There have been links made between the solar activity in this period and lower temperatures on earth, prompting interest in sunspot numbers by certain meteorologists (Wagner, 2005). It has recently been suggested that another such anomaly is imminent or has already begun (Clette et al., 2014).

Looking at the data retrospectively, it is easy to observe the change of 1790-1830 by inspection. Using statistical methods, the exact time at which the parameters of the underlying process abruptly changed, referred to as a *changepoint*, can be identified even more accurately. A question to be asked, however, is whether or not it is possible to use statistical methods to detect a changepoint online (i.e. when it happens, or soon after), rather than looking back on it as a part of a historic data set. The current change in sunspot numbers is a prime example of such a potential changepoint. This question serves as the main inspiration for this project.

There exists an algorithm for the online detection of changepoints using Bayesian statistics as proposed by Adams & MacKay (2007). They point out that there are already a number of existing methods for online changepoint detection within a Classical framework. The Bayesian algorithm which they proposed is only applicable to certain classes of data (that can be modelled by a likelihood with a known conjugate prior) and as such is not suitable for sunspot data, which is too complex to be modelled in such a way. It has been

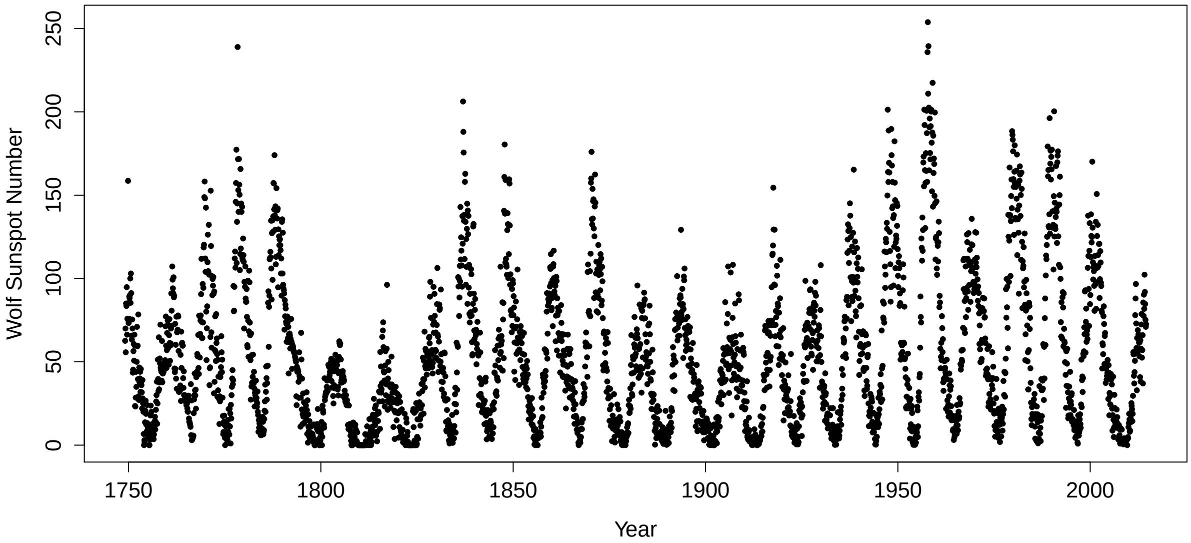

Figure 1.1: Monthly mean values of Wolf's Sunspot Index from February 1749 to July 2014.

an aim of this project to improve their algorithm by introducing alternatives to the existing steps, in the hope that the algorithm will become applicable to more classes of models, including those which could approximate the sunspot data.

The main aims of this project have been to

- replicate the algorithm proposed by Adams and MacKay along with their results, using the computer package R;
- develop a new algorithm suitable for a more general class of data and compare its results to those of Adams and MacKay;
- apply the new algorithm to the Wolf Sunspot Index data and investigate the results.

Section 2 of this dissertation describes the methods involved in the various algorithms used as well as explaining associated theory and notation where appropriate. Section 3 then provides the results of implementing the algorithms and Section 4 gives a discussion and a final conclusion to the dissertation.

2 Methods

This section of the dissertation outlines the theory and methodology which has been applied throughout this project.

Firstly, in Section 2.1 some general concepts and notation associated with the changepoint data sets used throughout the project are defined and discussed. Section 2.2 gives a brief review of existing literature on the topic. Section 2.3 then discusses the notion behind the Bayesian changepoint detection algorithm proposed by Adams & MacKay (2007) along with the concept of *conjugacy* on which it strongly relies. Section 2.4 provides further detail on Adams and MacKay's algorithm. Specifically, in Section 2.4.1 a detailed explanation of the implementation of the algorithm is provided, using a Gaussian likelihood with unknown, changing, mean and known, constant, variance as an example. In Sections 2.4.2 - 2.4.4 the implementation of the algorithm for other types of Gaussian data and finally for several real world examples is considered, concluding the matter on the replication of Adams and MacKay's work.

Section 2.5 describes how I introduced Markov Chain Monte Carlo (MCMC) steps to Adams and MacKay's algorithm, in order to extend its use to likelihoods which do not have a conjugate prior distribution. In Section 2.6 Harmonic Data is discussed, as this is the class to which the Sunspot Index data approximately belong, the section details how the algorithms with and without MCMC steps is implemented both for artificial data and the Sunspot Index data. Finally, Section 2.7 describes how the Sunspot Index data were pre-processed, by applying several transformations to the raw data, so that they are more appropriate for the algorithms developed in Section 2.6.

2.1 Changepoint Data: Concepts and Notation

Figure 2.1 displays data generated from a normal distribution with constant variance $\sigma^2 = 1$ and changing mean θ indicated by the horizontal blue lines. This dataset is typical of the *Changepoint Data* used throughout the project. These data are characterised by abrupt changes in one or more of their generative parameters. The temporal locations of said abrupt changes are referred to as *changepoints*. In this project all of the changepoint datasets used are univariate discrete time series where observations are regularly spaced; however there is no reason that multivariate data or data which does not occur in a regular temporal pattern could not be considered to be changepoint data if there are still abrupt changes present.

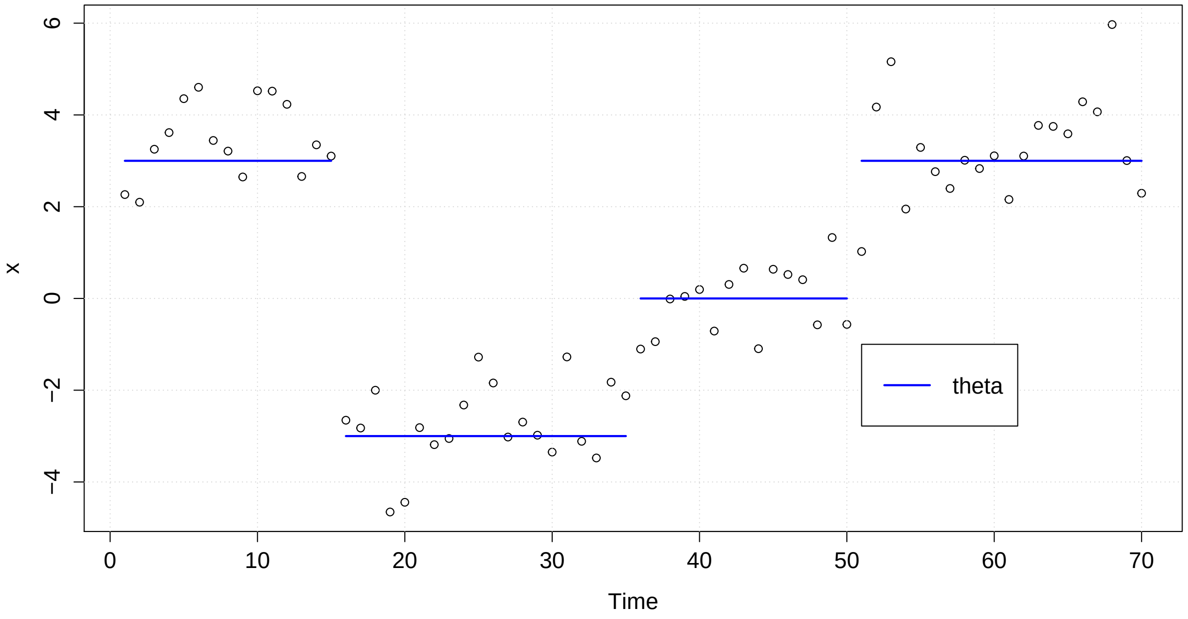

Figure 2.1: 70 observations of Gaussian changepoint data x with abruptly changing mean and known, fixed, variance. Circles represent observations and blue lines represent the underlying mean θ which takes the values $3, -3, 0$ and 3.

Throughout this dissertation, a key quantity of interest will be *run length*. At any given point the run length is the time since the last changepoint. It will be represented by the variable r_t, where t denotes the current time. A given value of the run length therefore implies a changepoint at a particular point in the past, i.e. if $r_t = a$, it implies that the last changepoint occurred at time $t - a$. The current run length only provides information on the location of the most recent changepoint. The distribution of run length values based on observed data has been of interest throughout the project and is calculated as a means of estimating changepoint locations. Since the distribution function assigns probability mass to different potential run lengths at different times it effectively assigns probability mass to different changepoint locations (implied by the run length values) as well.

In principle, a changepoint can occur at any time. However, with the discrete datasets featured in this project, it will never be possible to make any exact inference on their location. All that will be detectable is that a changepoint occurred at some point between

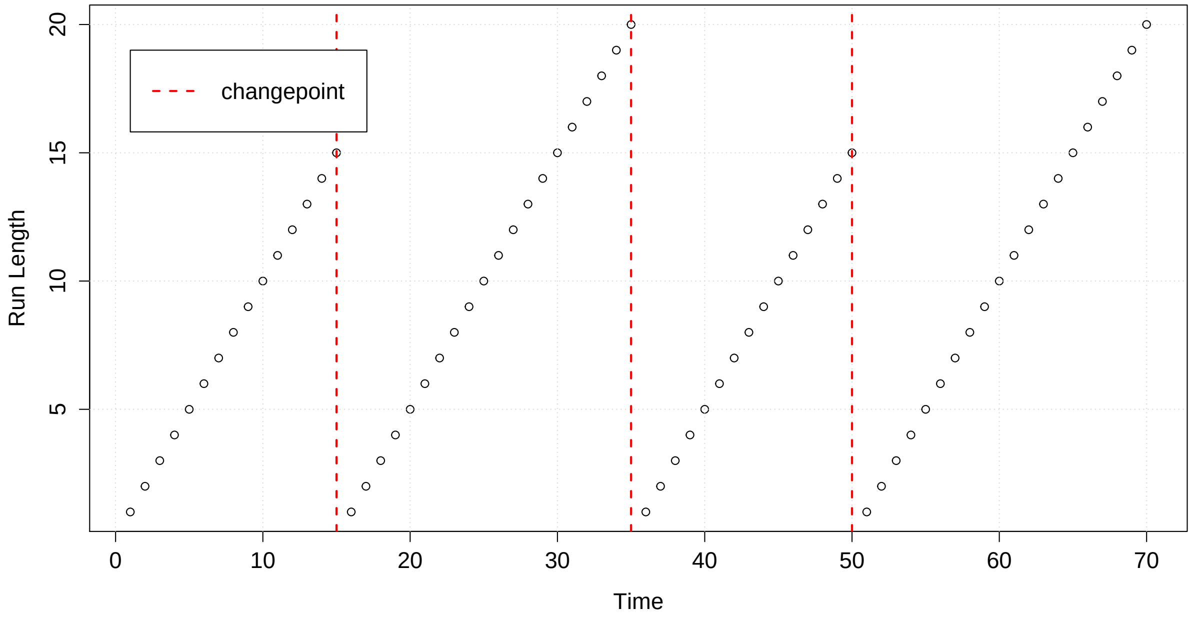

Figure 2.2: Run length against Time for the Gaussian changepoint data with changing mean shown in Figure 2.1. Circles indicate run length values at integer times and red dashed lines indicate the occurrence of a changepoint over the next time unit.

two observations. So for this reason, run length is treated as a discrete variable with the assumption being made that changepoints can only occur at one temporal position between successive observations. While it may seem intuitive to assume that this position is the midpoint between observations, the decision was made to assume that the opportunity for a changepoint came immediately after an observation. The reasoning behind this decision was to ensure that run length always took integer values, which was easier from a computational point of view.

Throughout this dissertation changepoints will be indicated by dashed, vertical red lines as in Figure 2.2 and it is worth bearing in mind that this indicates a changepoint exists at some unknown temporal location between the time point at which the line is placed and the subsequent time point. Figure 2.2 displays the run length values associated with the data displayed in Figure 2.1, with the changepoints highlighted.

The notation $\mathbf{x}_t^{(r)} = \{x_{t-(r-1)}, x_{t-(r-2)}, ..., x_t\}$ will be used to denote the set of observations associated with a given run length r_t. In practice, this is the set of observations since the last changepoint, whose position is implied by the run length r_t. An analogous notation is used for sets of parameters in the statistical models to be developed. For a parameter vector η, the vector $\eta_t^{(r)}$ denotes the vector of posterior parameters calculated using the set of observations $\mathbf{x}_t^{(r)}$.

2.2 Review of Literature

This section gives a review of existing literature, on both Frequentist and Bayesian approaches to changepoint detection. It is by no means exhaustive but gives some context to the work of Adams & MacKay (2007), on which much of this project was based.

From the pioneering work of Page (1954), Frequentist approaches to changepoint detection have generally been online. Page's paper proposed a system for detecting changepoints in an industrial application (such as a manufacturing line), based on monitoring a moving average and classifying a changepoint as the time when this average moved beyond certain limits. This approach tended to focus on a mean parameter θ and could be implemented as a one-sided or two-sided system. Lorden (1971) provides a detailed derivation of the boundaries which should be used for optimal stopping in Page's approach. Page later proposed a test for detecting a change at an unknown point in a parameter whose initial value was known. This test was based on a null hypothesis that the full time series consisted of observations from a single distribution with this initial parameter. The need to know such an initial value for the parameter of interest could be seen as a significant drawback to this approach (Page, 1955). All of these older Frequentist approaches are limited by the need to assume a particular model.

Modern work such as that of Desobry et al. (2005) has often involved on the use of Support Vector Machines. Desobry et al. introduce a nonparametric approach to hypothesis testing that is based on Machine Learning theory and is referred to as Kernel Change Detection. The main advantage of their approach is the freedom from model assumptions.

In contrast, the majority of Bayesian approaches are offline (retrospective). Barry & Hartigan (1993) describe an offline approach where various partitioning methods used to divide the data into contiguous sets, with constant parameter values within each set. The merits of different partitioning criteria are discussed and the authors put their own method forward as one which performs consistently well in a variety of scenarios. Chib (1998), Green (1995) and Stephens (1994) introduce MCMC based methods, such as Chib's use of

MCMC to estimate complex likelihood functions and allow different changepoint models to be compared by Bayes factors.

When Adams & MacKay (2007)'s paper was written it was part of a small minority of Bayesian approaches that were online.

2.3 Adams and MacKay's Bayesian Online Changepoint Detection Algorithm

Adams & MacKay (2007)'s algorithm does not offer a test or an automatic means of detecting a changepoint (at least directly, although it could however be adapted to do so) but rather a means of estimating the run length distribution at each observation time based on the data previously observed.

In principle, this process is just an application of the product rule of probability, combining the predictive probabilities of a new datum given each possible run length with the run length distribution from the previous time step to obtain the current run length distribution. The algorithm is therefore iterative and proceeds in this fashion for as many data as are observed.

As the algorithm is Bayesian, selecting suitable priors both on the unknown parameters and the frequency of changepoints is an important issue. Adams and MacKay exclusively choose conjugate priors for the unknown parameters to simplify the required computations. This will be discussed later in this section. They represent prior belief about the distribution of changepoints using a *hazard function* $H(r_t)$.

The hazard function can take a variety of forms and could be chosen so that changepoints are more likely at small run lengths than large ones or vice versa. However in this project, following the lead of Adams and MacKay, the decision was taken to assume that the length of the gaps between changepoints followed a Geometric distribution with fixed parameter λ, making the hazard function $H(t) = 1/\lambda, \forall t$ i.e. a constant rate of changepoints was assumed. With λ being chosen to reflect some reasonable rate for the data being handled, it can be interpreted as the prior expectation of the mean run length. The justification for this decision was that it is both a noninformative choice and easy to interpret. It was hoped that by choosing the hazard function in this way bias in the results would be limited and the data could speak for themselves.

The algorithm's implementation is described in much fuller detail, in the context of a simple example, in the next section.

The algorithm proposed by Adams and MacKay gains much of its computational efficiency from the use of *conjugate priors* — a class of prior distributions whose use makes determining a posterior (and by extension a posterior predictive distribution) possible in an analytic form.

Definition 2.3.1:
If $\mathcal{F}$ is a class of sampling distributions $p(y|\theta)$ and $\mathcal{P}$ a class of prior distributions for θ, then the class $\mathcal{P}$ is *conjugate* for $\mathcal{F}$ if

$$p(\theta|y) \in \mathcal{P} \ \forall \ p(y|\theta) \in \mathcal{F} \text{ and } \ p(\theta) \in \mathcal{P}.$$

(Gelman et al., 1995)

Using a conjugate prior for a likelihood function therefore means that the posterior distribution will always have a known, closed parametric form — of the same family as the prior distribution. Particular interest is focussed on *natural* conjugate families where $\mathcal{P}$ is the set of all densities with the same functional form as the likelihood. These conjugate families belong exclusively to the *Exponential Family* of distributions and Adams and MacKay focus their work on Exponential family distributions.

Definition 2.3.2:
The class of distributions $\mathcal{F}$ is an *exponential family* if all its members have the form,

$$p(y_i|\theta) = f(y_i)g(\theta)e^{\phi(\theta)^T u(y_i)}.$$

(Gelman et al., 1995)

The likelihood of a sequence of iid observations from an exponential family distribution $y = (y_1, ..., y_n)$ will always be the function of a *sufficient statistic* $\sum_{i=1}^{n} u(y_i)$. A statistic is sufficient if no other statistic which can be calculated from the sample gives any more information as to the value of the parameter being estimated (Fisher, 1922). Exponential family distributions are unique in being the only ones to have a fixed number of sufficient statistics for all n. This is the property that results in them being the only distributions with natural conjugate priors. If the prior density is specified as

$$p(\theta) \propto g(\theta)^{\eta} e^{\phi(\theta)^T \nu},$$

then the posterior density is

$$p(\theta) \propto g(\theta)^{\eta+n} e^{\phi(\theta)^T(\nu+t(y))}.$$

This is of the same distributional form as the prior and shows that the prior density is conjugate (Gelman et al., 1995).

Conjugate priors are popular for many of the same reasons that commonly used distributions (such as the Binomial or Normal) are: they have results that are easy to understand, and more importantly they simplify computations. As the form of the posterior is known in advance all that needs to be done is an update of parameters (based on the fixed number of sufficient statistics) and computationally inefficient integrations are avoided.

Robert (2001) mentions that conjugate priors are seen to be useful because it is argued they provide an objective and systematic means of selecting a prior distribution when the information available does not allow a prior distribution to be fully specified (if it even exists). This same point however is also a source of criticism as it is noted that using a conjugate prior involves making certain assumptions and potentially ignoring aspects of the prior information. Conjugate priors are not always particularly robust, a property that can create serious issues when sample sizes are small and the prior distribution has much more influence on the resulting inference. Additionally, in situations where the likelihood function is more complex a conjugate prior may simply not exist and therefore a method reliant on conjugate priors will not be applicable everywhere.

Clearly conjugate priors have value under a specific set of conditions but it will be worthwhile to extend the algorithm of Adams and MacKay to a version where their use is not necessary. This is what this project attempts to do through introducing sampling by Markov Chain Monte Carlo steps in place of steps that relied on the properties of conjugate priors.

2.4 Replicating Adams and MacKay's Algorithm

To develop the algorithm further, the first step is naturally to understand how to implement it in its existing state. Thus work was done to replicate the results of Adams & MacKay (2007) for several increasingly complex examples. All the examples considered consisted of univariate time series. Initially, artificial data were explored, considering three cases in particular:

1. Gaussian data with unknown mean and known variance,
2. Gaussian data with known mean and unknown variance,
3. Gaussian data with unknown mean and unknown variance.

The idea behind using an artificial dataset is that the true parameter values and changepoint locations are known exactly and the exact accuracy of the algorithm can be assessed. For each of these examples it was necessary to identify an appropriate posterior predictive distribution which results from the use of said conjugate prior. Sections 2.4.1-2.4.3 describe the methods used (in extensive detail for the first case, in outline for the remaining two) and Sections 3.1.1-3.1.3 display the associated results.

The algorithm as developed for artificial data was then applied to real world examples as considered by Adams and MacKay. Sections 2.4.4 and 3.1.4 detail the nature of the datasets, the methods used to analyse them and show a comparison of the results obtained by Adams and MacKay and as a part of this project.

2.4.1 Gaussian Data with Unknown Mean and Known Variance

The first data considered were independent identically distributed observations from a $N(\theta, \sigma^2)$ distribution where σ^2 was known and the value of θ changed at changepoints. The data were generated in R using the `rnorm` command (R Core Team, 2012) and both the true values of θ and the changepoint locations were stored to check the accuracy of the algorithm.

Hoff (2009) describes how we arrive at the posterior predictive distribution required in the algorithm for this data. For some data $\mathbf{x} = \{x_1, ..., x_n | \theta, \sigma^2\} \sim$ i.i.d $N(\theta, \sigma^2)$, the joint sampling density is given by

$$p(x_1, ..., x_n | \theta, \sigma^2) = (2\pi\sigma^2)^{-n/2} \text{exp}\{-\textstyle\sum(x_i - \theta)^2/2\sigma^2\}.$$

It can be shown that in this case the normal distribution is also a conjugate prior for the mean. Suppose $\theta \sim N(\mu_0, \tau_0^2)$ then, by Bayes' Theorem

$$p(\theta|\mathbf{x}, \sigma^2) \propto p(\theta)p(\mathbf{x}|\theta, \sigma^2) \propto \text{exp}\{-\tfrac{1}{2\tau_0^2}(\theta - \mu_0)^2\}\text{exp}\{-\tfrac{\sum(x_i-\theta)^2}{2\sigma^2}\}.$$

After completing the square in θ in the exponent it can be shown that

$$p(\theta|\mathbf{x},\sigma^2) \propto \exp\{-\tfrac{1}{2\tau_n^2}(\theta-\mu_n)^2\},$$

where

$$\tau_n^2 = \frac{1}{\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}}, \quad \mu_n = \frac{\frac{\mu_0}{\tau_0^2}+\frac{\sum x_i}{\sigma}}{\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}}, \tag{1}$$

meaning, $\theta|\mathbf{x},\sigma^2 \sim N(\mu_n,\tau_n^2)$ — another normal distribution, proving the conjugacy in this case. Generally the posterior predictive distribution for a new observation x̃ would be obtained by evaluating the integral

$$p(\tilde{\text{x}}|\mathbf{x},\sigma^2) = \int p(\tilde{\text{x}}|\theta,\sigma^2)p(\theta|\mathbf{x},\sigma^2)d\theta.$$

However in this case Hoff (2009) uses a quicker argument based on the posterior mean and variance of x̃ and the fact that the sum of normally distributed independent random variables is also normal to arrive at the posterior predictive distribution in a closed form:

$$\tilde{\text{x}}|\sigma^2,\mathbf{x} \sim N(\mu_n,\tau_n^2+\sigma^2),$$

where μ_n and τ_n^2 are defined in equation (1). With the posterior predictive distribution and its parameters identified it was then possible to proceed with implementing Adams and MacKay's algorithm. The first step of the algorithm is to make suitable prior assumptions and initialise relevant quantities.

The hazard function should be specified and as described in Section 2.1 the discrete exponential (Geometric) distribution on waiting times for changepoints leading to a constant hazard $H(t) = 1/\lambda \ \forall t$ is a sound, easy to interpret and computationally efficient choice. The simple option for initialising the run length distribution of $\mathbf{P}(r_0 = 0) = 1$ was also chosen for this example, partly because of its simplicity and also as there was no previous data on which to base a hazard function. The last part of the initialisation step was to choose some initial values for the hyperparameters μ_0 and τ_0^2.

The following steps of the algorithm are implemented for times $i = 1, ..., n$ based upon a sample $\mathbf{x} = x_1, ..., x_n$.

1. Evaluate the predictive probability of x_i under the posterior predictive distribution associated with each run length possible at time $t = i-1$

$$\pi_i^{(j)} = \mathcal{N}(x_i|\mu_i^{(j)},\tau_i^{2(j)}+\sigma^2) \text{ for } j = 0, 1, ..., i-1.$$

These probabilities give a measure of the likelihood of the new observation being generated under the parameters associated with each possible run length, which in turn can be used to estimate the likelihood of each of these run lengths being the true one.

2. Growth and changepoint probabilities (essentially the probability of the run length being in a particular state at a particular time) are evaluated by combining the predictive probabilities with the hazard function and growth and changepoint probabilities from the previous iteration. The only way to reach a run length $r_i = j > 0$ is to have $r_{i-1} = j - 1$. So the relevant result is, by the product rule, for $j = 1, ..., i$

$$\mathbf{P}(r_t = i, \mathbf{x}_{1:i}) = \mathbf{P}(r_t = i-1, r_{t-1}, \mathbf{x}_{1:(i-1)}) \cdot \pi_i^{(j)} \cdot (1 - H(i)),$$

with the $(1 - H(r))$ term representing the probability that there is not a changepoint and $\mathbf{x}_{1:i}$ denoting the set of values $\{x_1, ..., x_i\}$.

On the other hand $r_i = 0$ indicates that there has been a changepoint and this can occur from any run length. This means that in order to calculate $\mathbf{P}(r_i = 0, \mathbf{x}_{1:i})$ the Law of Total Probability must be applied summing over every possible run length at time $t = i - 1$:

$$\begin{aligned}\mathbf{P}(r_i = 0, \mathbf{x}_{1:i}) &= \sum_{r_{i-1}} \mathbf{P}(r_i = 0, r_{i-1}, \mathbf{x}_{1:i}) \\ &= \sum_{r_{i-1}} \mathbf{P}(r_i = 0, x_i | r_{i-1}, \mathbf{x}_{1:(i-1)}) \cdot \mathbf{P}(r_{i-1}, \mathbf{x}_{1:(i-1)}) \\ &= \sum_{r_{i-1}} \mathbf{P}(r_i = 0 | r_{i-1}) \cdot \mathbf{P}(x_i | r_{i-1}, \mathbf{x}_i^{(r)}) \cdot \mathbf{P}(r_{i-1}, \mathbf{x}_{1:(i-1)}),\end{aligned}$$

which can be re-expressed in the notation used for growth probabilities as

$$\mathbf{P}(r_i = 0, \mathbf{x}_{1:t}) = \sum_{j=0}^{i-1} \mathbf{P}(r_{i-1} = j, \mathbf{x}_{1:(i-1)}) \cdot \pi_i^{(j)} \cdot H(i).$$

3. To obtain the run length distribution conditional on the observed data for the current iteration, renormalisation is performed (by the product rule in reverse):

$$\mathbf{P}(r_i = j | \mathbf{x}_{1:i}) = \tfrac{\mathbf{P}(r_i = j, \mathbf{x}_{1:i})}{\mathbf{P}(\mathbf{x}_{1:i})} \text{ for } j = 0, 1, ..., i,$$

where $\mathbf{P}(\mathbf{x}_{1:i}) = \sum_{j=0}^{i} \mathbf{P}(r_i = j, \mathbf{x}_{1:i})$. Once renormalisation is performed the probability density associated with run length 0 will be the same fixed constant $1/\lambda$ for all i. This reflects the fact that a changepoint (assumed to occur immediately after a datum) could occur with equal likelihood from any run length, but cannot be detected until some data are generated with the new parameters. So the estimate of the probability of run length zero will always be based solely on the prior information contained in the hazard function and that is why it is fixed.

4. Finally update the model parameters for use in the next iteration

$$\mu_{i+1}^{(0)} = \mu_0$$

$$\tau_{i+1}^{2\ (0)} = \tau_0^2$$

$$\mu_{i+1}^{(j)} = \frac{\frac{\mu_i^{(j-1)}}{\tau_i^{2\,(j-1)}} + \frac{x_i}{\sigma^2}}{\frac{1}{\tau_i^{2\,(j-1)}} + \frac{1}{\sigma^2}} \text{ for } j = 1, ..., i,$$

$$\tau_{i+1}^{(j)} = \frac{1}{\frac{1}{\tau_i^{2\,(j-1)}} + \frac{1}{\sigma^2}} \text{ for } j = 1, ..., i.$$

After some investigation it was found to be desirable to use a slightly altered version of step 2. The set of growth and changepoint probabilities at each time point does not represent a properly normalised distribution (i.e. they do not sum to 1) and as a result it was found that in some examples once run lengths became large the growth and changepoint probabilities became so small that they fell below the minimum precision level in R, causing them to be rounded down to 0 — which would be incorrect.

So a way around this was to use the run length distribution conditional on the observed data from the previous iteration in place of the growth and changepoint probabilities from the previous iteration. The application of Bayes' theorem and the removal of the data term — which is a fixed common factor — shown in the expressions below, proves that this is possible.

For growth probabilities

$$\begin{aligned}\mathbf{P}(r_i = j, \mathbf{x}_{1:i}) &= \mathbf{P}(r_{i-1} = j-1, \mathbf{x}_{1:(i-1)}) \cdot \pi_i^{(j)} \cdot (1 - H(i)) \\ &= \mathbf{P}(r_{i-1} = j-1 | \mathbf{x}_{1:(i-1)}) \cdot \mathbf{P}(\mathbf{x}_{1:(i-1)}) \cdot \pi_i^{(j)} \cdot (1 - H(i)) \\ &\propto \mathbf{P}(r_{i-1} = j-1 | \mathbf{x}_{1:(i-1)}) \cdot \pi_i^{(j)} \cdot (1 - H(i)).\end{aligned}$$

For changepoint probabilities

$$\begin{aligned}\mathbf{P}(r_i = 0, \mathbf{x}_{1:i}) &= \sum_{j=0}^{i-1} \mathbf{P}(r_{i-1} = j, \mathbf{x}_{1:(i-1)}) \cdot \pi_i^{(j)} \cdot H(i) \\ &= \sum_{j=0}^{i-1} \mathbf{P}(r_{i-1} = j | \mathbf{x}_{1:(i-1)}) \cdot \mathbf{P}(\mathbf{x}_{1:(i-1)}) \cdot \pi_i^{(j)} \cdot H(i) \\ &\propto \sum_{j=0}^{i-1} \mathbf{P}(r_{i-1} = j-1 | \mathbf{x}_{1:(i-1)}) \cdot \pi_i^{(j)} \cdot H(i).\end{aligned}$$

As the normalisation in step 3 removes the data term $\mathbf{P}(\mathbf{x}_{1:(i-1)})$, this modification will not impact the performance of the algorithm. Indeed, while they do not mention it explicitly in their paper, this modification is used by Adams and MacKay in supplementary material to their paper.

Commented code illustrating how these steps were implemented in R has been provided as a separate file. By selecting a suitable prior distribution the steps can be applied to any normally distributed data with known (or, more practically, well estimated) variance and changing mean. In Section 3.1.1 the results of applying the algorithm to the data displayed in Figure 2.1 are described.

2.4.2 Gaussian Data with Known Mean and Unknown Variance

A different problem is to suppose that the mean is known and fixed but the variance is neither. The algorithm for this example can be implemented in almost exactly the same way, the only modifications that need to be made are to steps 1 and 4, as a different conjugate prior will be required and as a result the posterior predictive distribution will also be different.

As described by Bishop (2008) it is easier to work with the precision of the normal distribution $\lambda \equiv 1/\sigma^2$ whose likelihood is of the form

$$p(x_1, ...x_n|\theta, \lambda^{-1}) \propto \lambda^{n/2}\text{exp}\{-\tfrac{\lambda}{2}\sum_{i=1}^{n}(x_i - \theta)^2\}.$$

A conjugate prior in this case is a $Gamma(\alpha, \beta)$ distribution: the resulting posterior distribution, when i.i.d. observations $\mathbf{x} = \{x_1, ..., x_n\}$ from $N(\theta, 1/\lambda)$ are observed is $Gamma(\alpha_n, \beta_n)$ where $\alpha_n = \alpha + n/2$ and $\beta_n = \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2$. The posterior predictive distribution which is required for Adams and MacKay's (2007) algorithm is a generalized Student's t-distribution on $\nu = 2\alpha_n$ degrees of freedom with mean θ and variance $\xi = \alpha_n/\beta_n$ (Murphy, 2007). This distribution takes the form

$$St(x|\theta,\xi,\nu) = \frac{\Gamma(\frac{\nu}{2}+\frac{1}{2})}{\Gamma(\frac{\nu}{2})}(\frac{\xi}{\pi\nu})^{1/2}(1+\frac{\xi(x-\theta)^2}{\nu})^{-\frac{\nu}{2}-\frac{1}{2}}. \tag{2}$$

(Bishop, 2008)

In summary, it is assumed a priori that $\lambda \sim Gamma(\alpha_0, \beta_0)$ which gives a posterior predictive distribution $\tilde{\mathrm{x}}|\theta, \mathbf{x} \sim St(\theta, \xi = \frac{\alpha_n}{\beta_n}, \nu = 2\alpha_n)$, where α_n and β_n are defined above.

The steps of the algorithm then become:

1. Evaluate the predictive probability of x_i under the posterior predictive distribution associated with each run length possible at time $t = i - 1$

$$\pi_i^{(j)} = St(x_i|\theta, \tfrac{\alpha_i^{(j)}}{\beta_i^{(j)}}, 2\alpha_i^{(j)}) \text{ for } j = 0, 1, ..., i-1.$$

2. Calculate growth and changepoint probabilities

$$\mathbf{P}(r_i = j, \mathbf{x}_{1:i}) \propto \mathbf{P}(r_{i-1} = j-1|\mathbf{x}_{1:(i-1)}) \cdot \pi_i^{(j)} \cdot (1 - H(r))$$
$$\mathbf{P}(r_i = 0, \mathbf{x}_{1:i}) \propto \sum_{j=0}^{i-1} \mathbf{P}(r_{i-1} = j-1|\mathbf{x}_{1:(i-1)}) \cdot \pi_i^{(j)} \cdot H(r).$$

3. Obtain the run length distribution conditional on the observed data for the current iteration.

$$\mathbf{P}(r_i = j|\mathbf{x}_{1:i}) = \tfrac{\mathbf{P}(r_i=j,\mathbf{x}_{1:i})}{\mathbf{P}(\mathbf{x}_{1:i})} \text{ for } j = 0, 1, ..., i,$$

where $\mathbf{P}(\mathbf{x}_{1:i}) = \sum_{j=0}^{i} \mathbf{P}(r_i = j, \mathbf{x}_{1:i})$.

4. Finally update the hyperparameters for use in the next iteration

$$\alpha_{i+1}^{(0)} = \alpha_0^{(0)}, \quad {\beta_{i+1}^2}^{(0)} = \beta_0^{2(0)}$$

$$\alpha_{i+1}^{(j)} = \alpha_i^{(j-1)} + \tfrac{1}{2}, \text{ for } j = 1, ..., i$$

$$\beta_{i+1}^{(j)} = \beta_i^{(j-1)} + \tfrac{1}{2}\sum_{i=1}^{n}(x_1 - \theta)^2 \text{ for } j = 1, ..., i.$$

By selecting a suitable prior hyperparameters and a hazard function the steps can be applied to any normally distributed data with known (or, more practically, well estimated) mean and changing variance. In Section 3.1.2 an example dataset is shown along with the results of applying the algorithm to it.

2.4.3 Gaussian Data with Unknown Mean and Unknown Variance

The third and final case of Gaussian data considered is the case where both the mean and variance are unknown and subject to changes. Again, the algorithm only requires modifications to steps 1 and 4 because of the use of a different conjugate prior. Again a more convenient parameterisation involves the precision, λ, rather than the variance itself.

In this case a joint conjugate prior on μ and λ is given by a Normal-Gamma distribution, a combination of a Normal prior on μ and a Gamma prior on λ (Murphy, 2007):

$$\begin{aligned} NG(\theta, \lambda|\mu_0, \kappa_0, \alpha_0, \beta_0) &= \mathcal{N}(\theta, \lambda|\mu_0, (\kappa_0\lambda)^{-1}) Gamma(\lambda|\alpha_0, \beta_0) \\ &= \frac{1}{Z_{NG}(\mu_0, \kappa_0, \alpha_0, \beta_0)} \lambda^{\alpha_0 - \frac{1}{2}} \exp(-\frac{\lambda}{2}[\kappa_0(\theta - \mu_0)^2 + 2\beta_0]) \end{aligned}$$

where $Z_{NG}(\mu_0, \kappa_0, \alpha_0, \beta_0) = \frac{\Gamma(\alpha_0)}{\beta_0^{\alpha_0}}(\frac{2\pi}{\kappa_0})^{1/2}$ is a normalising constant. As the distribution is conjugate for the Gaussian likelihood, the posterior will also be a Normal-Gamma distribution. The posterior distribution $NG(\theta, \lambda|\mu_n, \kappa_n, \alpha_n, \beta_n)$'s parameters have the form

$$\begin{aligned} \mu_n &= \frac{\kappa_0\mu_0 + n\bar{x}}{\kappa_0 + n}, \\ \kappa_n &= \kappa_0 + n, \\ \alpha_n &= \alpha_0 + n/2, \\ \beta_n &= \beta_0 + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{\kappa_0 n(\bar{x} - \mu_0)^2}{2(\kappa_0 + n)}. \end{aligned}$$

As a result the posterior predictive will again be a generalized Student's t-distribution (as defined in equation 2) on $\nu = 2\alpha_n$ degrees of freedom with mean μ_n and variance $\xi = \frac{\beta_n(\kappa_n+1)}{\alpha_n\kappa_n}$ (Murphy, 2007).

An obvious, but ultimately trivial, difference in this case is that the prior distribution has four hyperparameters $\mu_0, \kappa_0, \alpha_0$ and β_0. This has no significant impact on the algorithm's steps other than to necessitate the updating of two more parameters than before, indeed a prior distribution with any number of hyperparameters could theoretically be used with Adams and MacKay's algorithm as long as it is conjugate to the assumed likelihood.

The alterations which need to be made to the algorithm from Section 2.4.1 are very similar to those described in Section 2.4.2 so the full steps of the algorithm for this third case are omitted here. Section 3.1.3 does however show example results.

It is possible to apply Adams and MacKay's algorithm to any data set with an exponential family likelihood by identifying a suitable conjugate prior, the associated posterior predictive distribution, and the formulae for the updating of posterior parameters and then making suitable alterations to the full algorithm as described in Section 2.4.1.

2.4.4 Real Datasets

Adams & MacKay (2007) apply their algorithm to three real datasets, which will be described in this section. In section 3.1.4 the results from applying the replicate algorithm to these are given.

The first data set consists of measurements of the nuclear magnetic response of underground rocks (Fearnhead & Clifford, 2003). The data were obtained when drilling through the Earth's crust and are suitable for use with Adams and MacKay's algorithm because the underlying signal is piecewise constant. As new strata of the Earth's crust are reached by drilling downwards, new types of rock are encountered, each with different mean nuclear magnetic responses which cause a jump in the signal. If it is assumed that the nuclear magnetic response measurements are normally distributed around the current level of the signal and their variance can be well estimated this can be treated as an example of Gaussian data with unknown mean and known variance and the algorithm from Section 2.4.1 can be applied to it. The data set consists of 4050 observations and shows where the use of conjugate priors really has its advantage; because of the size of the data set, over 8,000,000 predictive probabilities need to be calculated in the process of applying the changepoint detection algorithm. This already takes some time using the conjugate prior approach and could be unfeasible using a less efficient method.

The data along with the comparison of Adams and MacKay's results and the results obtained as a part of this project are given in Section 3.1.4.

The next data set consisted of daily returns of the Dow Jones Industrial Average from each business day between 5th July 1972 and 30th June 1975. A daily return was calculated as $R_t = \frac{p_t^{close}}{p_{t-1}^{close}} - 2$ where p_t^{close} is the closing price on day t. This time period is significant because of several major events with potential macroeconomic effects which occurred during it. These include the OPEC oil embargo and the Watergate affair which led to the resignation of US President Richard Nixon. Applying the algorithm for Gaussian data with known mean and unknown variance to the daily returns of the Dow Jones Industrial Average over this period would attempt to see if there are points where the volatility (variance) of the market changes suddenly. The timings of such changes could be compared to the timings of certain political events and an expert in the field could use this information to judge whether there is a causative link. As with the nuclear magnetic response data, an exploration of data and results is given in Section 3.1.4.

The third and final dataset (Jarret, 1979) considered by Adams and MacKay consisted of the dates of coal mining disasters. The specific data were the dates of coal mining explosions which resulted in 10 or more fatalities between 15th March 1851 and 22nd March 1962. A point of interest within this time frame is the year 1887 when the Coal Miners Act was introduced by Government, with the intention of reducing the number of such disasters.

Adams and MacKay modelled the data as a Poisson Process by weeks and placed a conjugate prior of a Gamma distribution on the rate parameter ϕ with parameters $\alpha_0 = \beta_0 = 1$. The resulting posterior predictive distribution is a Negative Binomial distribution with parameters α_n and $\frac{1}{1+\beta_n}$, where $\alpha_n = \alpha_0 + \sum_{i=1}^{n} x_i$ and $\beta_n = \beta_0 + n$. They used a constant hazard function with parameter $\lambda = 1000$.

In this project a related approach was used. If the data can be modelled by a Poisson Process with rate ϕ, it means that the waiting times between observations can be modelled as i.i.d. observations from an Exponential distribution with parameter ϕ. The same Gamma prior can be used as a conjugate prior for Exponential data but in this case the posterior predictive distribution is a Lomax distribution, with scale parameter $\beta_n = \beta_0 + \sum_{i=1}^{n} x_i$ and shape parameter $\alpha_n = \alpha_0 + n$. In Section 3.1.4 the waiting times between disasters are plotted as a time series and the results of applying the conjugate prior based algorithm with a Lomax posterior predictive distribution are given.

2.5 Introducing Markov Chain Monte Carlo steps

Having unpicked Adams & MacKay (2007) algorithm and replicated their results by implementing the algorithm in R, it is now possible to go about adding MCMC steps so that the algorithm can be used for data whose likelihood function lacks an associated conjugate prior distribution. Details of this algorithm are given here and in section 3.2 the results generated using the MCMC based algorithm are compared to those generated with Adams and MacKay's original conjugate prior based approach.

The area which requires the most substantial changes is the calculation of predictive probabilities, because if a conjugate prior is not used then the posterior predictive distribution will not necessarily have a known closed form. This means that the posterior predictive distribution can no longer simply be updated by updating the values of some sufficient statistics or hyperparameters.

In both algorithms it is necessary to calculate the predictive probabilities, $\pi_t^{(r)}(x) = \mathbf{P}(x_t|r_t, \mathbf{x}_{t-1}^{(r-1)})$. When a conjugate prior is used the information that is being conditioned upon (the current run length, and the set of observations since the changepoint r_t units in the past, which is implied by the run length) can be summarised by sufficient statistics which in turn give the hyperparameters of a closed form posterior predictive distribution. This distribution can then be used to calculate all desired predictive probabilities.

However without relying on the properties of conjugacy, it cannot be assumed that the posterior predictive distribution has a neat analytical form such as a Normal density function or Student's t-distribution. Instead the predictive probability of an observation should be found by evaluating the following integral:

$$\begin{aligned}\mathbf{P}(x_t|r_t, \mathbf{x}_t^{(r)}) &= \int \mathbf{P}(x_t|\theta, r_t, \mathbf{x}_{t-1}^{(r-1)})\mathbf{P}(\theta|r_t, \mathbf{x}_{t-1}^{(r-1)})d\theta \\ &= \int \mathbf{P}(x_t|\theta, r_t, \mathbf{x}_{t-1}^{(r-1)})\frac{\mathbf{P}(\theta)\mathbf{P}(\mathbf{x}_{t-1}^{(r-1)}|\theta)}{\mathbf{P}(\mathbf{x}_{t-1}^{(r-1)})}d\theta.\end{aligned}$$

The trouble with this is that the integral can often be very difficult or indeed impossible to evaluate, so instead Monte Carlo integration (Hastings, 1970) is introduced to approximate the integral. The approximation below is made

$$\textstyle\int \mathbf{P}(x_t|\theta)\mathbf{P}(\theta|r_t, \mathbf{x}_{t-1}^{(r-1)})d\theta \approx \frac{1}{M}\sum_{i=1}^{M}\mathbf{P}(x_t|\theta^i),$$

where the θ^i are simulated using a Metropolis-Hastings algorithm with $\mathbf{P}(\theta|r_t, \mathbf{x}_t^{(r)})$ being the target distribution. For $r_t = 0, \mathbf{P}(\theta|r_t, \mathbf{x}_{t-1}^{(r-1)}) = \mathbf{P}(\theta)$. The $\mathbf{P}(\mathbf{x}_{t-1}^{(r-1)})$ term that is difficult to calculate will cancel out as a constant of proportionality during the Metropolis-Hastings steps and will no longer pose an issue. As long as M is chosen to be suitably large, the sum can provide an accurate approximation to the integral and in turn the MCMC based algorithm can generate results that are accurate approximations of the results generated by Adams and MacKay's algorithm.

Much like the conjugate prior based algorithm, the steps of the MCMC based algorithm are perhaps best explained in the context of an example. So the simplest example — Gaussian data with unknown mean and known variance — will again be considered. The conjugate prior to this likelihood is also normal and this will be used as the prior in the

MCMC based approach also, but without directly relying on the properties of conjugacy. The same parametrisation from Section 2.2.1 will continue to be used, so that the prior is $\theta \sim N(\mu_0, \tau_0^2)$ and the likelihood is $\mathbf{x} \sim N(\theta, \sigma^2)$

Every time a predictive probability $\pi_t^{(r)}$ needs to be calculated the following Metropolis-Hastings steps are carried out.

- Initialise the sequence of θ values, by choosing a suitable value $\theta^{(1)}$
- for $i = 2, ..., M$
  - Propose a new value
    $$\theta' \sim N(\theta^{(i-1)}, \gamma^2),$$
    where γ^2 is some suitable variance. The exact value will be dictated by the scale of the data but it should generally be smaller than σ^2.
  - Calculate an acceptance probability
    $$\alpha = \frac{p(\theta' | \mu_t^{(r)}, {\tau_t^{(r)}}^2, \sigma^2) q(\theta' | \theta^{(i-1)})}{p(\theta^{(i-1)} | \mu_t^{(r)}, {\tau_t^{(r)}}^2, \sigma^2) q(\theta^{(i-1)} | \theta')},$$
    where q is the proposal distribution. In this situation where the proposal distribution is symmetric, the q terms cancel and the acceptance probability reduces to
    $$\alpha = \frac{\mathcal{N}(\theta' | \mu_t^{(r)}, {\tau_t^{(r)}}^2, \sigma^2)}{\mathcal{N}(\theta^{(i-1)} | \mu_t^{(r)}, {\tau_t^{(r)}}^2, \sigma^2)}.$$
  - Set $\theta^{(i)} = \theta'$ with probability $\min(1, \alpha)$, and $\theta^{(i)} = \theta^{(i-1)}$ otherwise.
- Calculate $\pi_t^{(r)} \approx \frac{1}{M} \sum_{i=1}^{M} \mathbf{P}(x_{t+1} | \theta^i)$.

Otherwise the algorithm works as before, simply using these predictive probabilities estimated based on the Metropolis-Hastings steps in place of those from the known form posterior predictive distribution. In Section 3.2 the MCMC based algorithm is implemented for the data displayed in Figure 2.1 and the results from this method are compared to the result obtained using Adams and MacKay's algorithm.

2.6 Harmonic Data

The Sunspot data is a manifestation of a very complicated magnetohydrodynamic system within the sun, and as a result is far from simple to model. However one of the standout features of the data, and indeed one of the reasons it cannot be modelled by an exponential family distribution is the strong cyclical aspect. Therefore a natural starting point in attempting to implement the changepoint detection algorithm on the Sunspot data is to understand how to apply it to some general harmonic data.

In this section an approach is derived to apply the changepoint algorithm to data from a simple harmonic series in two distinct cases:

1. where the angular frequency, ω, is assumed to be known,
2. where it is unknown.

The form of the likelihood of this data will allow a mixture of the conjugate prior and MCMC based approaches to be used in the second case, as certain parameters which enter the likelihood function linearly will have conjugate priors. The advantage of using this mixture approach is that computing time is saved whenever the conjugate prior based approach can be used. The algorithms for both cases will be applied to artificial data and to the Sunspot data.

A simple model for a harmonic time series is

$$y_i = C\sin(\omega t_i + \phi) + \epsilon_i \quad for i = 1, ..., n, \tag{3}$$

where C, ω and ϕ are constants representing the amplitude, angular frequency and phase of the wave respectively. The ϵ_i terms represent some i.i.d. Normal noise function with a variance σ^2.

Two of the parameters of this model, ω and ϕ, are within the sine function and it is difficult to draw inference on them due to the high level of non-linearity. The problem can be resolved somewhat using the trigonometric identity $\sin(X + Y) = \sin(X)\cos(Y) + \cos(X)\sin(Y)$ to re-express the model as

$$y_i = A\sin(\omega t_i) + B\cos(\omega t_i) + \epsilon_i \text{ for } i = 1, ..., n,$$

where $A = C\cos(\phi)$ and $B = C\sin(\phi)$. This model has two linear parameters, A and B and one non-linear parameter, ω, instead of two non-linear parameters and one linear parameter so should be somewhat easier to work with.

2.6.1 Angular Frequency Known

Assuming that the values of ω and σ^2 are fixed and known it is possible to find a conjugate prior for $\mathbf{A} = \begin{pmatrix} A & B \end{pmatrix}^T$. The likelihood of data $\mathbf{y} = \begin{pmatrix} y_1 & ... & y_n \end{pmatrix}^T$ is normal and analogously to the earlier case a conjugate prior is the normal distribution. To make the proof of this clear, some notation is first defined

$$\begin{aligned} \mathbf{s} &= \begin{pmatrix} \sin(\omega t_1) & ... & \sin(\omega t_n) \end{pmatrix}^T, \\ \mathbf{c} &= \begin{pmatrix} \cos(\omega t_1) & ... & \cos(\omega t_n) \end{pmatrix}^T, \\ D &= \begin{pmatrix} \mathbf{s} & \mathbf{c} \end{pmatrix} \text{ and } \mathbf{A} = \begin{pmatrix} A & B \end{pmatrix}^T. \end{aligned}$$

The likelihood of the i.i.d sample $\mathbf{y}$ will then be

$$\mathbf{y}|\mathbf{A}, \omega, \sigma^2 \sim N_n\left(\begin{pmatrix} \mathbf{s} & \mathbf{c} \end{pmatrix} \mathbf{A}, \sigma^2 I_{n\times n}\right), \tag{4}$$

and as the conjugate prior is normal, it can be written in its simplest form (assuming a zero mean and equal variance τ_0^2 on both random variables and independence) as

$$\begin{aligned} \mathbf{A} &\sim N_2(\mathbf{0}, \tau_0^2 I_{2\times 2}) \\ \epsilon &\sim N_n(\mathbf{0}, \sigma^2 I_{n\times n}). \end{aligned}$$

By the properties of conjugacy the posterior distribution will also be a bivariate normal, $\mathbf{A}|\mathbf{y}, \omega, \sigma^2 \sim N_2(\mu, V)$, whose mean vector μ and covariance matrix V must be identified.

It is possible to determine μ and V by considering only the exponentiated part of the posterior distribution which is

$$Z = -\tfrac{1}{2}\left[\tfrac{1}{\sigma^2}(\mathbf{y} - D\mathbf{A})^T(\mathbf{y} - D\mathbf{A}) + \tfrac{1}{\tau_0^2}\mathbf{A}^T\mathbf{A}\right],$$

by removing a factor of $-\frac{1}{2}$, expanding and gathering like terms this can be re-expressed as

$$\tfrac{\mathbf{y}^T\mathbf{y}}{\sigma^2} + \mathbf{A}^T\Sigma^{-1}\mathbf{A} - \tfrac{2\mathbf{y}^T D\mathbf{A}}{\sigma^2}$$

where $\Sigma^{-1} = \frac{\mathbf{D}^T\mathbf{D}}{\sigma^2} + \frac{I_{2\times 2}}{\tau_0^2}$. As this expression is in a quadratic form, the covariance matrix $V = Cov(\mathbf{A}|\mathbf{y}, \omega, \sigma^2)$ can be identified by inspection as $\Sigma = \left[\frac{\mathbf{D}^T\mathbf{D}}{\sigma^2} + \frac{I_{2\times 2}}{\tau_0^2}\right]^{-1}$.

To find the mean, a property of the normal distribution is used: because of the symmetry in its density function, the mean of a normal distribution is also its mode. This means that μ can be identified by differentiation. Now,

$$\tfrac{dZ}{d\mathbf{A}^T} = 0 + 2\Sigma^{-1}\mathbf{A} - \tfrac{2}{\sigma^2}D^T\mathbf{y}.$$

Setting this expression equal to zero allows the mode $\hat{\mathbf{A}} = \frac{1}{\sigma^2}\Sigma D^T\mathbf{y}$ to be identified. So in summary, the posterior distribution is

$$\mathbf{A}|\mathbf{y}, \omega, \sigma^2, \tau_0^2 \sim N_2(\tfrac{1}{\sigma^2}\Sigma D^T\mathbf{y}, \Sigma).$$

To apply the changepoint algorithm in the conjugate prior based form however, the posterior predictive distribution is required. The new observation $\tilde{y} = \begin{pmatrix}\mathbf{s} & \mathbf{c}\end{pmatrix}\mathbf{A} + \tilde{\epsilon}$ that this distribution aims to predict is a linear transformation of $\mathbf{A}$, so will also be normally distributed. Its exact parameters can then be identified simply by evaluating the expectation and variance of $\tilde{y}$, saving an unpleasant integration. The posterior predictive distribution is found to be

$$\tilde{y}|\mathbf{y}, \omega, \sigma^2, \tau_0^2 \sim N\left(\tfrac{\tilde{D}\Sigma\mathbf{D}^T\mathbf{y}}{\sigma^2}, \tilde{D}\Sigma\mathbf{D}^T + \sigma^2\right),$$

where $\tilde{D} = \begin{pmatrix}\sin(\omega\tilde{t}) & \cos(\omega\tilde{t})\end{pmatrix}$, and where $\tilde{t}$ is the time at which observation $\tilde{y}$ occurs.

The algorithm is then implemented in entirely the same way as described in Section 2.4 except using the univariate normal distribution given above to calculate posterior predictive probabilities. In Section 3.3.1 the use of this algorithm is illustrated for artificial data and in Section 3.4 the results from its use on the Sunspot data are given.

2.6.2 Angular Frequency Unknown

When it cannot be assumed that ω is known a mixed conjugate prior and MCMC based approach should be used. The desired posterior predictive distribution is then no longer conditioned on ω. The posterior predictive distribution is then

$$p(\tilde{y}|\mathbf{y}, \sigma^2, \tau_0^2, \Delta t),$$

where Δt is the time lag between the (equally-spaced) data points. It can be manipulated into integral form below by marginalisation and an application of the product rule

$$\begin{aligned} p(\tilde{y}|\mathbf{y}, \sigma^2, \tau_0^2, \Delta t) &= \int p(\tilde{y}, \omega|\mathbf{y}, \sigma^2, \tau_0^2, \Delta t)d\omega \\ &= \int p(\tilde{y}|\mathbf{y}, \omega, \sigma^2, \tau_0^2, \Delta t)p(\omega|\mathbf{y}, \sigma^2, \tau_0^2, \Delta t)d\omega. \end{aligned}$$

This integral can then be approximated by Monte Carlo integration

$$\int p(\tilde{y}|\mathbf{y}, \omega, \sigma^2, \tau_0^2, \Delta t)p(\omega|\mathbf{y}, \sigma^2, \tau_0^2, \Delta t)d\omega \approx \tfrac{1}{M}\sum_{i=1}^{M} p(\tilde{y}|\mathbf{y}, \omega^{(i)}, \sigma^2, \tau_0^2, \Delta t),$$

where $\omega = \{\omega^{(1)}, ..., \omega^{(M)}\}$ is a set of M Metropolis-Hastings draws from $p(\omega|\mathbf{y}, \sigma^2, \tau_0^2, \Delta t)$, the posterior distribution on ω.

The posterior distribution on ω can be found by a marginalisation of the joint posterior on ω and $\mathbf{A}$, which is a combination of the likelihood of the data and the priors on ω and $\mathbf{A}$. The likelihood remains the same n-dimensional normal distribution (4) from Section 2.6.1., similarly the prior on $\mathbf{A}$ remains the bivariate normal prior $N_2(\mathbf{0}, \tau_0^2 T_{2\times 2})$. For ω, a uniform prior is used $\omega \sim U(0, \frac{\pi}{\Delta t})$ where $\frac{\pi}{\Delta t}$ is the Nyquist frequency (Grenander, 1959). The Nyquist frequency is chosen as the upper limit of this prior to avoid issues of identifiability due to aliasing.

As the prior on ω is uniform the exponential part of the joint posterior (multiplied by a factor of -2) will still be

$$\begin{aligned} Z &= \frac{1}{\sigma^2}(\mathbf{y} - D\mathbf{A})^T(\mathbf{y} - D\mathbf{A}) + \frac{1}{\tau_0^2}\mathbf{A}^T\mathbf{A} \\ &= \frac{\mathbf{y}^T\mathbf{y}}{\sigma^2} - \frac{2}{\sigma^2}\mathbf{y}^T D\mathbf{A} + \mathbf{A}^T\Sigma^{-1}\mathbf{A} \end{aligned}$$

as in Section 2.6.1. It also therefore remains the case that $\hat{\mathbf{A}} = \frac{1}{\sigma^2}\Sigma D^T\mathbf{y}$.

By inspection

$$\begin{aligned} Z &= constant + (\mathbf{A} - \hat{\mathbf{A}})^T\Sigma^{-1}(\mathbf{A} - \hat{\mathbf{A}}) \\ &= constant + \mathbf{A}^T\Sigma^{-1}\mathbf{A} - 2\hat{\mathbf{A}}^T\Sigma^{-1}\mathbf{A} + \hat{\mathbf{A}}^T\Sigma^{-1}\hat{\mathbf{A}}. \end{aligned}$$

Then, by equating coefficients of this expression and the previous expression for Z it is clear that

$$\tfrac{\mathbf{y}^T\mathbf{y}}{\sigma^2} = constant + \hat{\mathbf{A}}^T\Sigma^{-1}\hat{\mathbf{A}} \Rightarrow constant = \tfrac{\mathbf{y}^T\mathbf{y}}{\sigma^2} - \hat{\mathbf{A}}^T\Sigma^{-1}\hat{\mathbf{A}}.$$

Thus the joint posterior distribution (up to a constant of proportionality can be found

$$p(\mathbf{A}, \omega|\mathbf{y}, \sigma^2, \tau_0^2, \Delta t) \propto \exp(-\tfrac{1}{2}(\tfrac{\mathbf{y}^T\mathbf{y}}{\sigma^2} - \hat{\mathbf{A}}^T\Sigma^{-1}\hat{\mathbf{A}}))\exp(-\tfrac{1}{2}(\mathbf{A} - \hat{\mathbf{A}})^T\Sigma^{-1}(\mathbf{A} - \hat{\mathbf{A}})).$$

$\mathbf{A}$ only occurs in the second term which is Bivariate normal in $\mathbf{A}$. The normalising constant of this Bivariate normal will be $\frac{1}{2\pi(det(\Sigma))^{1/2}}$, so when the marginalisation is performed and the second exponential term is integrated out a $2\pi(det(\Sigma))^{1/2}$ term will remain.

The marginal posterior on ω then satisfies

$$\begin{aligned}p(\omega|\mathbf{y},\sigma^2,\tau^2) &\propto \exp(\frac{\mathbf{y}^T\mathbf{y}}{\sigma^2})\exp(\frac{1}{2}\hat{\mathbf{A}}^T\Sigma^{-1}\hat{\mathbf{A}})2\pi(det(\Sigma))^{1/2}\\ &\propto \exp(\frac{1}{2}\hat{\mathbf{A}}^T\Sigma^{-1}\hat{\mathbf{A}})2\pi(det(\Sigma))^{1/2}.\end{aligned}$$

As this distribution is being used in a Metropolis-Hastings algorithm it is sufficient to only know its form up to a constant of proportionality.

With the posterior identified, the mixed algorithm can be implemented in a similar fashion to the algorithm described in Section 2.5 using Monte Carlo integration to calculate the predictive probabilities. Section 3.3.2 shows the use of this mixed algorithm on artificial data and 3.4 gives the results of its application to the Sunspot data.

2.7 Pre-processing of Sunspot Data

Although an algorithm for handling data from a simple harmonic model (3) has now been developed, the raw sunspot data as shown in Figure 1.1 do not resemble data generated from this model very strongly. Model fitting issues could create problems with the algorithm as the assumption of a particular likelihood family is key. Therefore before applying the algorithms of Section 2.6 to the sunspot data some pre-processing was performed.

Firstly smoothed monthly figures from each year were averaged to give smoothed yearly figures for the years 1750 to 2013, a total of 264 observations. This was done because using the full set of monthly observations from this time period would result in a computationally intensive algorithm. The only other way to counteract this and make the algorithm more efficient would be to use fewer iterations in the Metropolis-Hastings steps which would reduce the accuracy of the Monte Carlo integration.

A square root transformation was also applied to the data to smooth out the peaks and troughs, making the pattern of the data more closely resemble a sine wave. The minimum of the transformed data was then subtracted from every observation to bring the minima of the waves down to zero (approximately). The choice of a square root transformation is justified because it is intuitive within the context of the underlying physics. Sunspots are a representation of the energy in the magnetic field which is related to the square of this field. Taking the square root is therefore trying to construct something proportional to the magnetic field, the physical quantity which bears significance (Macaulay, 1992).

Finally, every second cycle was inverted, by changing the signs of observations. This created a 22 year cycle centred on zero which would more closely resemble a sine wave than the 11 year cycle did. This transformation was deemed valid because as described in (Macaulay, 1992) a 22 year cycle is often considered to exist in sunspot numbers because every second 11 year cycle is generated by electromagnetic activity of opposite polarity.

These modifications go some way to manipulating the data into the form of data from the simple harmonic model. The full set of pre-processed data is plotted in Section 3.4, where the results of applying both the algorithms from Section 2.6 to these data are given.

3 Results

3.1 Replicating Adams and MacKay's Algorithm

3.1.1 Gaussian Data with Unknown Mean and Known Variance

The algorithm as described in Section 2.2.1 was applied to the data plotted in Figure 2.1. The 70 data points were generated from a normal distribution with fixed variance $\sigma^2 = 1$ and changing mean θ, which took values $3, -3, 0$ and 3 in the intervals $[1, 15]$, $[16, 35]$, $[36, 50]$ and $[51, 70]$ respectively. The red dashed lines in the run length plot in Figure 2.2 therefore represent the changepoints that the algorithm aims to detect. The algorithm was implemented with the hazard function parameter $\lambda = 18$ and the hyperparameters of the prior distribution chosen as $\mu_0 = 0$ and $\tau_0^2 = 10$.

The heat map plotted in Figure 3.1 represents the values of the run length distribution $\mathbf{P}(r_i = j|\mathbf{x}_{1:i})$. The probability mass associated with each potential run length at each time point is represented by the darkness of the cell, with darker shades of grey corresponding to a higher probability mass. A logarithmic grey scale has been used to make differences more apparent. The changepoints are again indicated by dashed red lines.

If the algorithm is working correctly, between changepoints the estimated most likely run length value (indicated by the darkest cell) should increase by 1 at each time point. After a changepoint happens the estimated most likely run length value should quickly drop to one much nearer zero and start increasing again from this point.

The results in Figure 3.1 indicate that the algorithm is working well, as the correct behaviour described above is being displayed. The position of the darkest cell is increasing linearly in time until changepoints are observed when there is a noticeable drop, illustrated by the triangular patterns in the heat map. Checking against the positions of the red dashed lines it can be seen that the algorithm correctly detects all changepoints and does not falsely report any others.

A point to note is that the algorithm does better at detecting the first changepoint (where θ changes by 6) than the second and third changepoints (where θ only changes by 3). The better quality of detection is shown through a more pronounced differentiation in shade. The result that a larger abrupt change is more easily detected continues to be observed in other examples.

Using a heat map also allows the full run length distribution to be seen, rather than

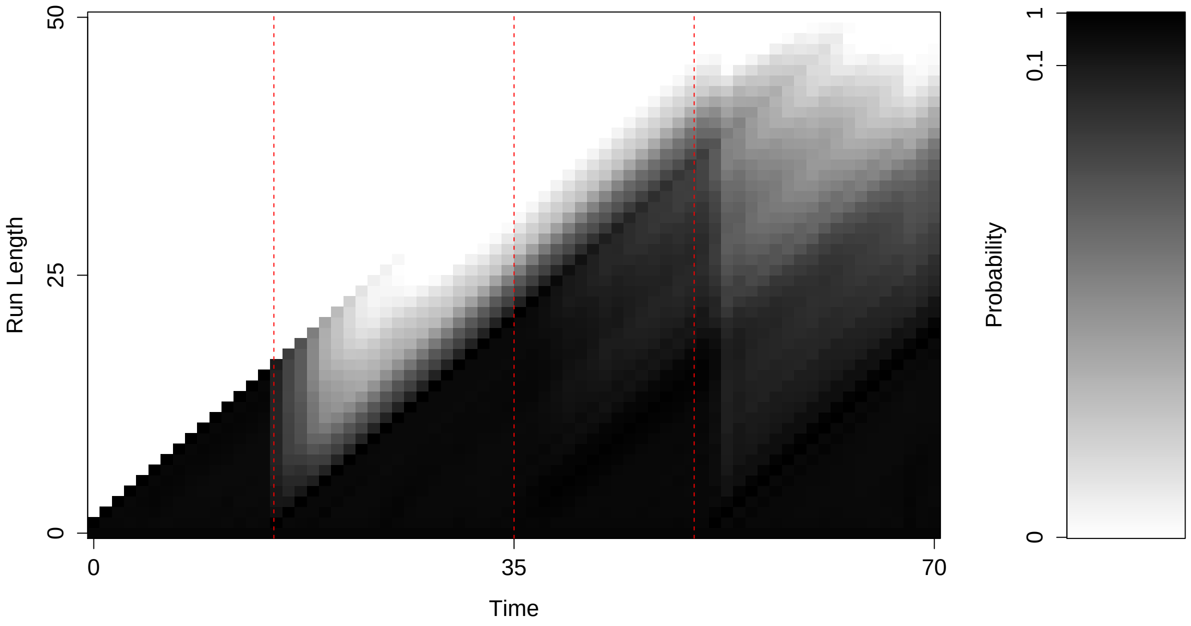

Figure 3.1: Posterior run length distribution at each time step for the changepoint data in Figure 2.1 as estimated by the conjugate prior based algorithm. A logarithmic grey scale is used with darker cells indicating higher probability. Dashed red lines indicate the true changepoints.

just the point estimate of the most likely value. This allows its other features, such as the fact that immediately after a changepoint, run length values corresponding to a growth are still more likely than those values that are lower but not so low that they correspond to a changepoint and how these run length values corresponding to growth gradually decay as more observations generated from the new mean are observed.

It is also a useful tool when analysing data with more complicated likelihoods or where the changepoints are not as pronounced. In these situations it may be more difficult for the algorithm to perform well and the estimated most likely run length value may not be the right one, but the second or third most likely may be; it may also be that the probability masses assigned to these and the most likely value are very similar, so considering only the point estimate would not show this whereas looking at the full distribution would make it clearer. As always, the full Bayesian posterior captures the uncertainty of the inference in rich detail.

3.1.2 Gaussian Data with Known Mean and Unknown Variance

Figure 3.2 displays 500 data generated from a normal distribution with fixed mean $\theta = 0$. The variance changed at the locations marked by the red lines, taking the values $4, 1, 4$, and 1 in the intervals $[1, 150], [151, 240], [241, 440]$ and $[441, 500]$ respectively. The algorithm as described in Section 2.4.2 was applied to these data with hazard function parameter $\lambda = 125$ and prior hyperparameters $\alpha_0 = 3$ and $\beta_0 = 0.5$. Figure 3.3 shows the heat map illustrating the results.

Examining the estimated run length distribution shown in Figure 3.3, it can be said that the changepoints are picked out successfully again. However this detection is done with less confidence than in the first example. This reduced confidence is indicated by a lower degree of differentiation in shade between different possible run lengths immediately after changepoints.

An explanation for this is that in this situation the algorithm is attempting to detect changes in variance, not in a mean. When the variance parameter changes, the most likely values continue to be in much the same region and the changepoint cannot be detected until the algorithm notices values are more (or less) tightly clustered and there is different behaviour at the extremes. When the mean parameter changes on the other hand the region of most likely values shifts too, so it is more likely to see an obviously different pattern of observations at a low run length. Since it is easier to notice the effects of a mean parameter change, than a variance parameter change, the algorithm is able to make the detection with more confidence in the first example.

3.1.3 Gaussian Data with Unknown Mean and Unknown Variance

In Figure 3.4, 200 observations are shown. These observations were generated from a normal distribution with changing mean θ and variance σ^2. In this example, θ takes the values $3, 3, -5$, and -5 and σ^2 takes the values 1, 4, 2, and 0.1, respectively, in the intervals $[1, 50], [51, 100], [101, 150]$ and $[151, 200]$. The algorithm as described in Section 2.4.3 was applied to these data with hazard function parameter $\lambda = 50$ and hyperparameters $\alpha_0 = 3, \beta_0 = 1, \mu_0 = 0$ and $\kappa_0 = 1$. Figure 3.5 gives the heat map illustrating the results.

Figure 3.5 shows that again the three changepoints are detected. However in this case the detection of the first and third changepoints is by no means confident - for some time after the true changepoints the probability densities associated with the run lengths cor-

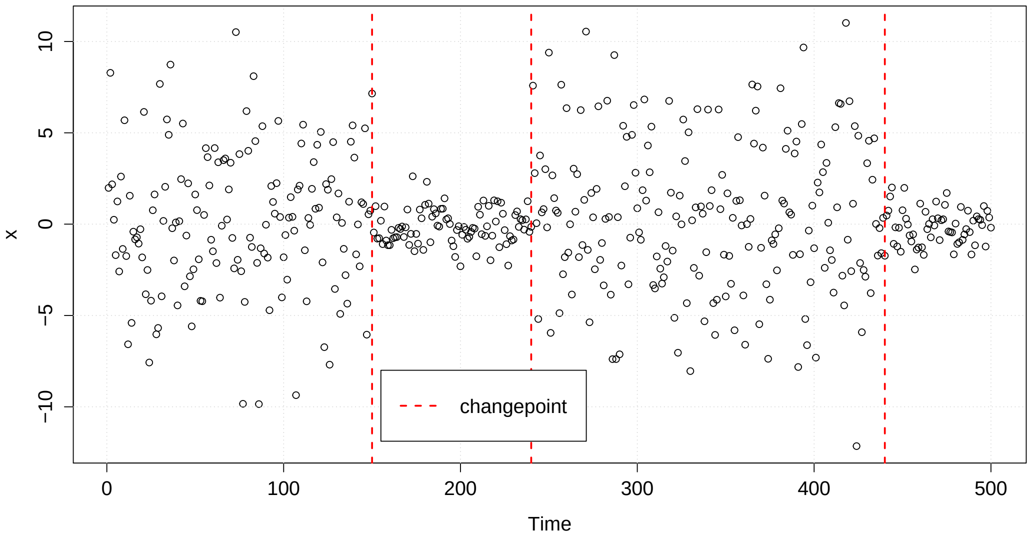

Figure 3.2: 500 observations of Gaussian changepoint data with known, fixed mean θ and changing variance. Circles represent observations and dashed red lines indicate changepoints

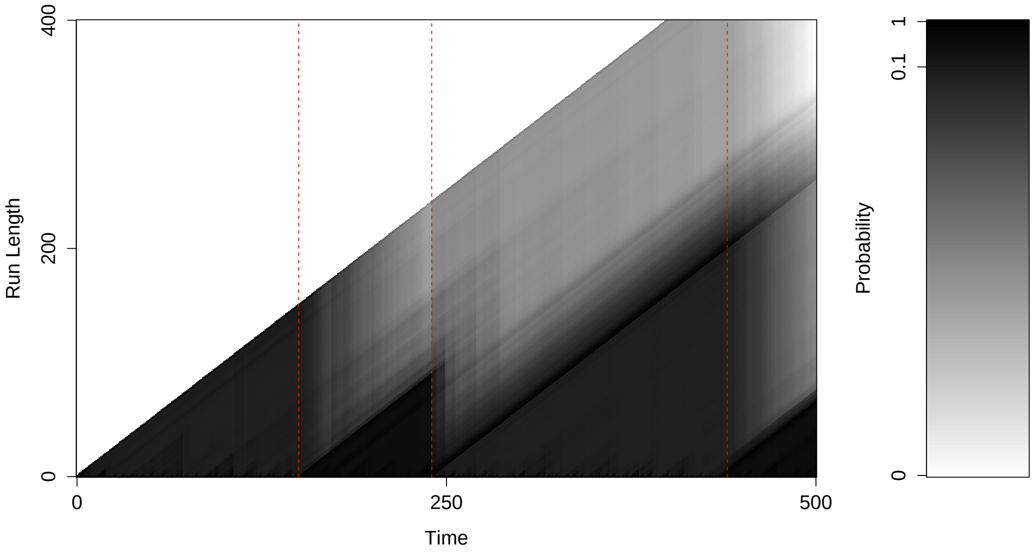

Figure 3.3: Posterior run length distribution at each time step for the changepoint data in Figure 3.2 as estimated by the conjugate prior based algorithm. A logarithmic grey scale is used with darker cells indicating higher probability. Dashed red lines indicate the true changepoints.

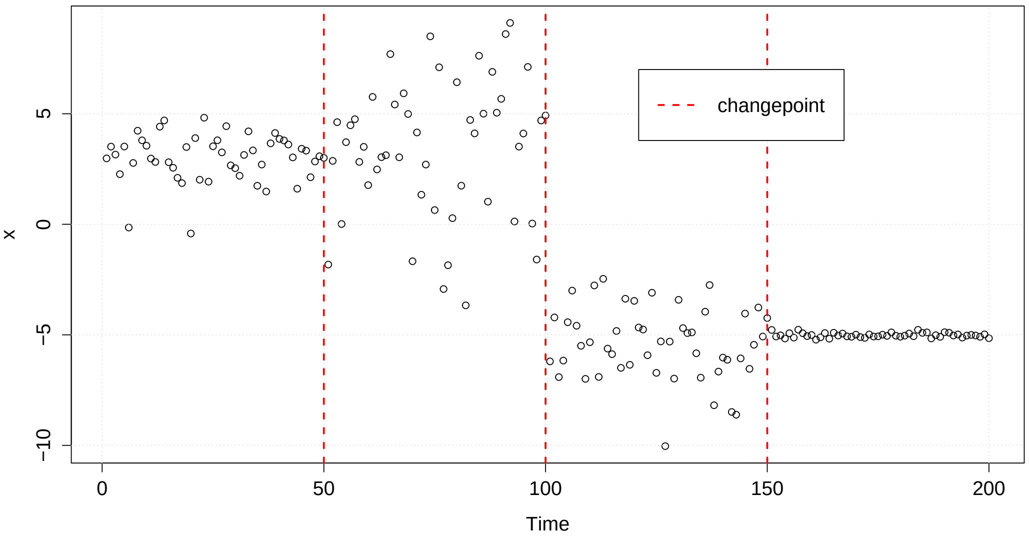

Figure 3.4: 200 observations of Gaussian changepoint data with changing mean and variance. Circles represent observations and dashed red lines indicate changepoints.

responding to these correct changepoints is lower than the probability densities associated with a continued growth. The probability densities associated with continued growth also decay much more slowly than in other examples meaning that the hypothesis that there were no changepoints at times 50 and 150 is still reasonable based on this output. The changepoint at time 100 is detected with much more confidence however. This is the only changepoint at which the mean parameter also changes and this difference in detection power again underlines how the effectiveness of the algorithm is strongly dependent on how extreme the change is and which parameters are changing.

This example again illustrates the value in plotting the full run length distribution. If only the run lengths estimated to be most likely by the algorithm were plotted, as can be seen in Figure 3.6, the output is jagged and somewhat confusing. The changepoints are all detected later by this approach and indeed if there were only 160 observations, the final changepoint would not be recognised at all. When the full run length distribution is plotted it highlights the correct run length as a likely value much sooner and shows how the two possible run lengths have a very similar probability density assigned to them.

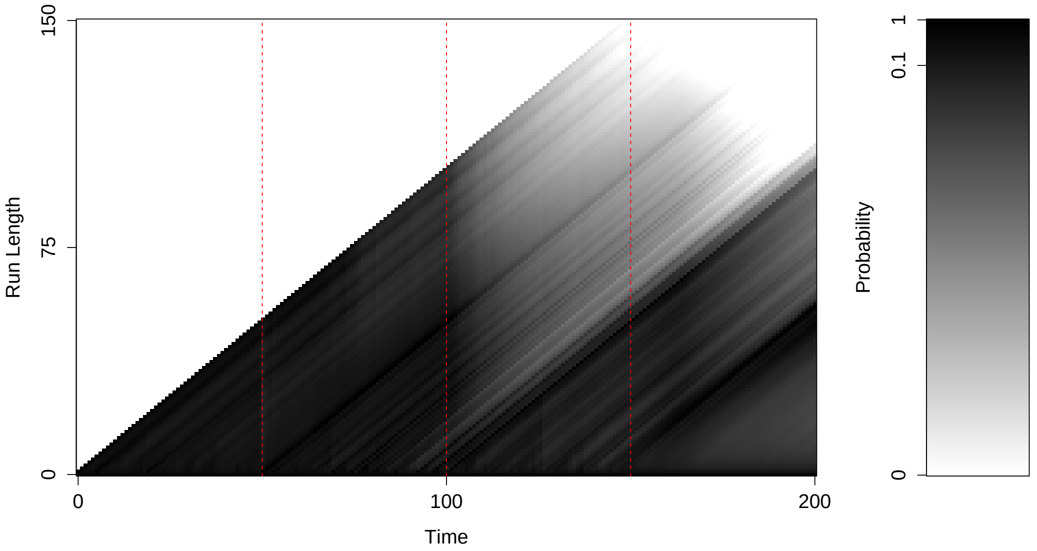

Figure 3.5: Posterior run length distribution at each time step for the changepoint data in Figure 3.4 as estimated by the conjugate prior based algorithm. A logarithmic grey scale is used with darker cells indicating higher probability. Dashed red lines indicate the true changepoints in at least one parameter.

3.1.4 Real Datasets

As described in Section 2.4.4, Adams & MacKay (2007) applied their algorithm to three real data sets. This section describes the results of doing the same with the algorithm developed to replicate theirs. The first data set, of nuclear magnetic responses from underground rocks obtained while drilling the Earth's crust, is shown in Figure 3.7. The changepoints caused by shifts in the underlying signal are very obvious to the eye in this case, which will make it easy to check that the algorithm is working.

To emulate Adams and MacKay's work the algorithm as described in Section 2.4.1 was applied to the data with $\sigma^2 = 2,000,000$, $\lambda = 250$ and prior hyperparameters $\mu_0 = 115,000$ and $\tau_0^2 = 1,000,000$. Figure 3.8 displays the heat map of the run length distribution and in this case the changepoints are very obvious because of the triangular pattern. Comparison can be made to the time series in Figure 3.7 and it can be seen that the positions of changepoints detected by the algorithm do match up with those in the time series.

Adams and MacKay only publish the posterior distribution for an 1100-datum subset

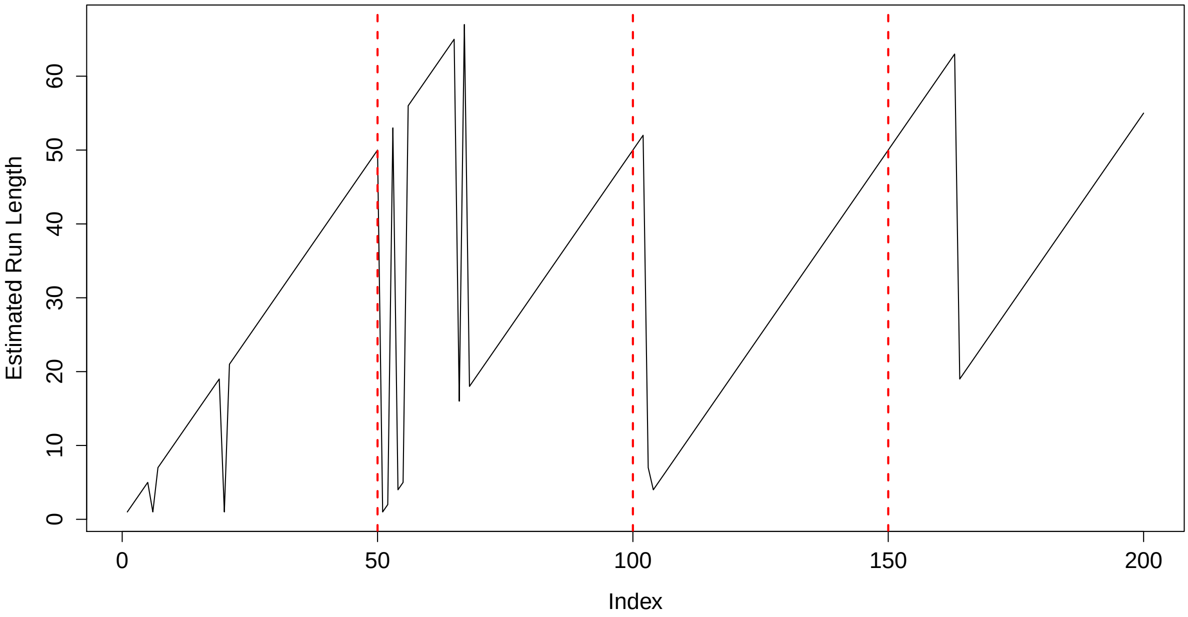

Figure 3.6: Most probable run length at each time step as estimated by the posterior run length distribution for the changepoint data in Figure 3.4. The black line is interpolated between estimates at integer times and dashed red lines indicate the true changepoints.

(between observations 1600 and 2700) but a quick comparison of their Figure and Figure 3.8 shows that the changepoints are detected at the same locations, with similar confidence.

The second dataset considered consisted of 753 daily returns on the Dow Jones between 5th July 1972 and 30th June 1975. The algorithm as described in 2.4.2 was applied to the data in the hope of detecting changepoints that would correspond to certain political events. Figure 3.9 displays a plot of the data against time.

The algorithm was applied with $\theta = 0, \lambda = 250$ (where λ refers to the hazard function parameter here rather than the precision). Prior hyperparameters for the Gamma prior were chosen as $\alpha_0 = 1$ and $\beta_0 = 0.0001$. Figure 3.10 displays the heat map of the posterior run length distribution generated by the algorithm with three times corresponding to major political events highlighted. A indicates the 30th January 1973 when three of President Nixon's former aides were convicted, B indicates the 19th October 1973 when the OPEC oil embargo began and C indicates the 9th August 1974 when President Nixon resigned.

Figure 3.10 shows that the results are lacking in confidence - indicated by low differ-

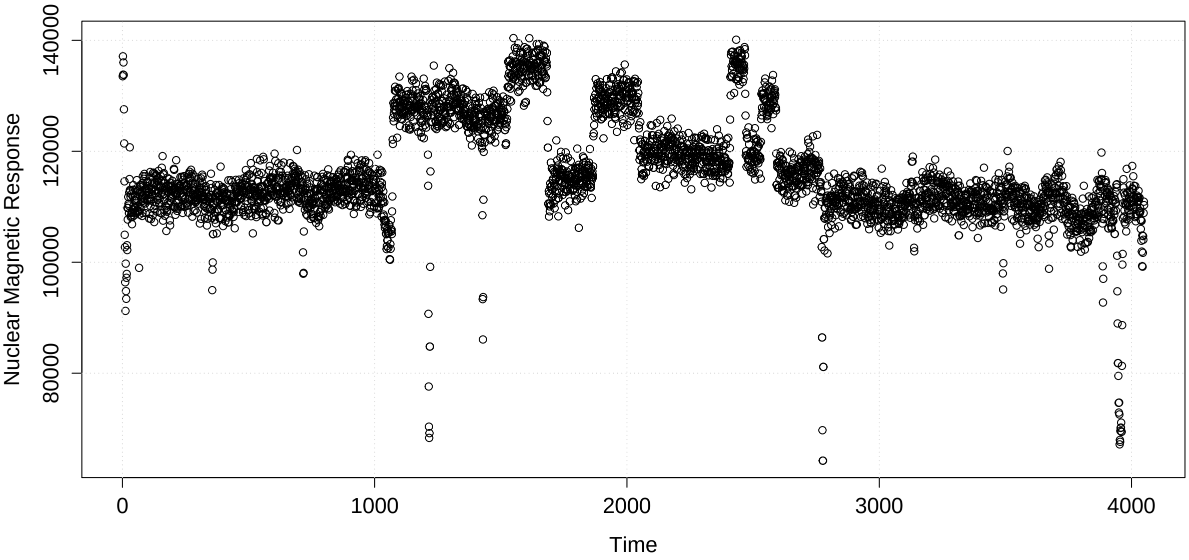

Figure 3.7: 4050 measurements of Nuclear Magnetic Response obtained during the drilling of a well, plotted against Time.

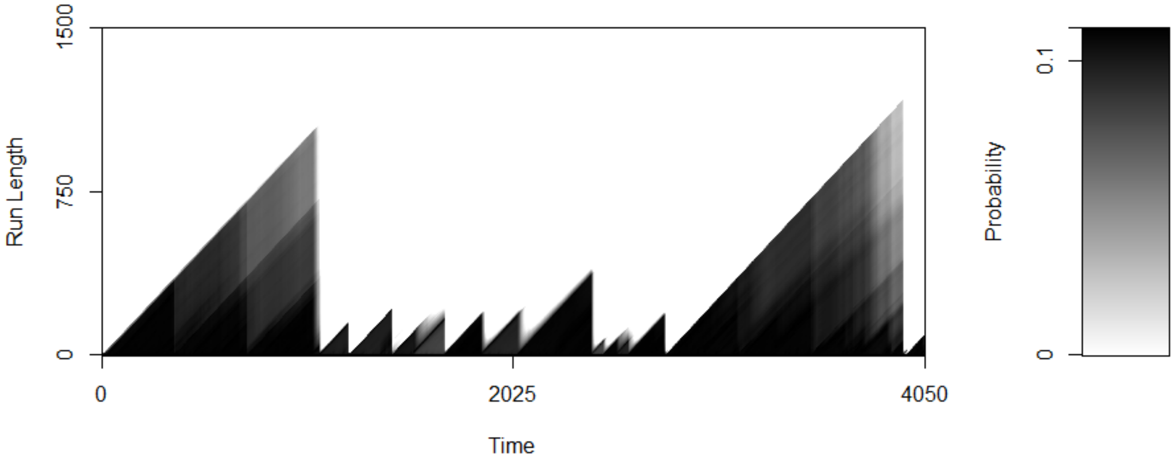

Figure 3.8: Posterior run length distribution at each time step for the Nuclear Magnetic Response data as estimated by the conjugate prior based algorithm. A logarithmic grey scale is used with darker cells indicating higher probability.

entiation in shade between 'probable' run length and 'improbable' run lengths. This is to

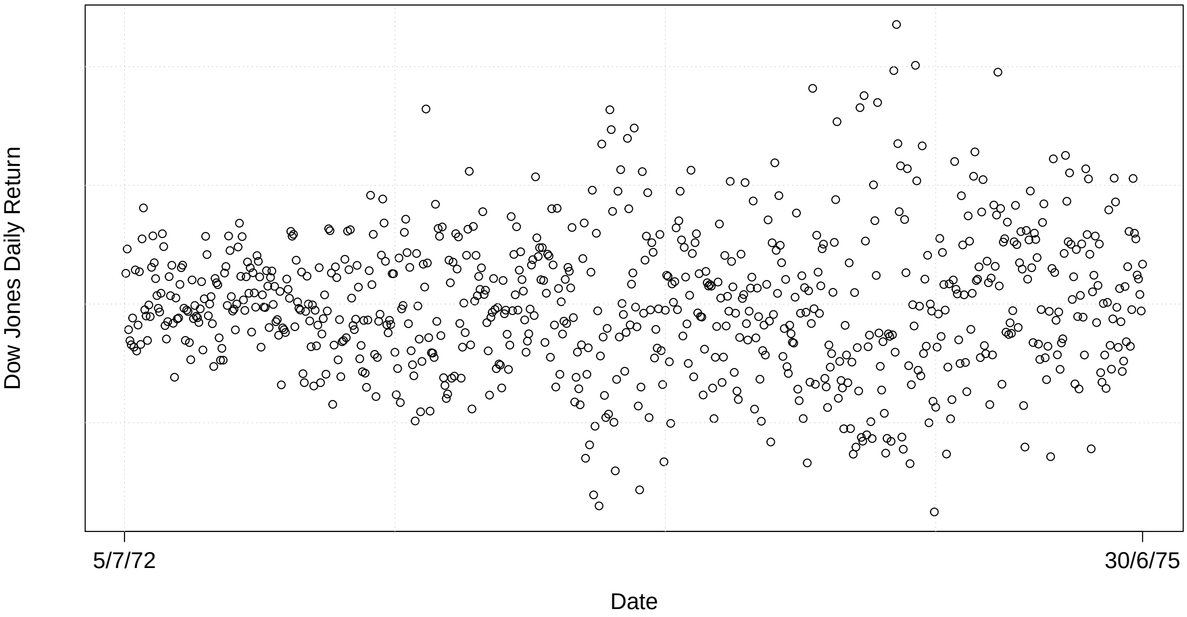

Figure 3.9: Daily returns on the Dow Jones Industrial Average, plotted as circles against Time.

be expected however as in this example, the algorithm is attempting to detect variance changes. It does seem to have detected three distinct changepoints around the times of the events A, B and C. B is the most pronounced changepoint on the heat map - indicated by the greater differentiation in shade shortly after it and the heat map suggests if there was a changepoint around time A, it may have actually happened slightly before the conviction - perhaps the market reacting to a revelation during the trial, or something else entirely. There may also be other more subtle changepoints present, for instance there seems to be another distinct run starting a few months after time C.

Again, when the results in Figure 3.10 are compared with those provided in Adams and MacKay's report the results are very similar. This is a good indication that the algorithm developed in this project is a correct replicate of their's.

The third and final dataset consisted of 184 waiting times, measured in weeks, between coal mine explosions resulting in more than 10 fatalities between the years 1851 and 1962. Figure 3.11 displays a plot of these data.

The conjugate prior based algorithm was applied with $\alpha_0 = \beta_0 = 1$ and $\lambda = 100$. Figure

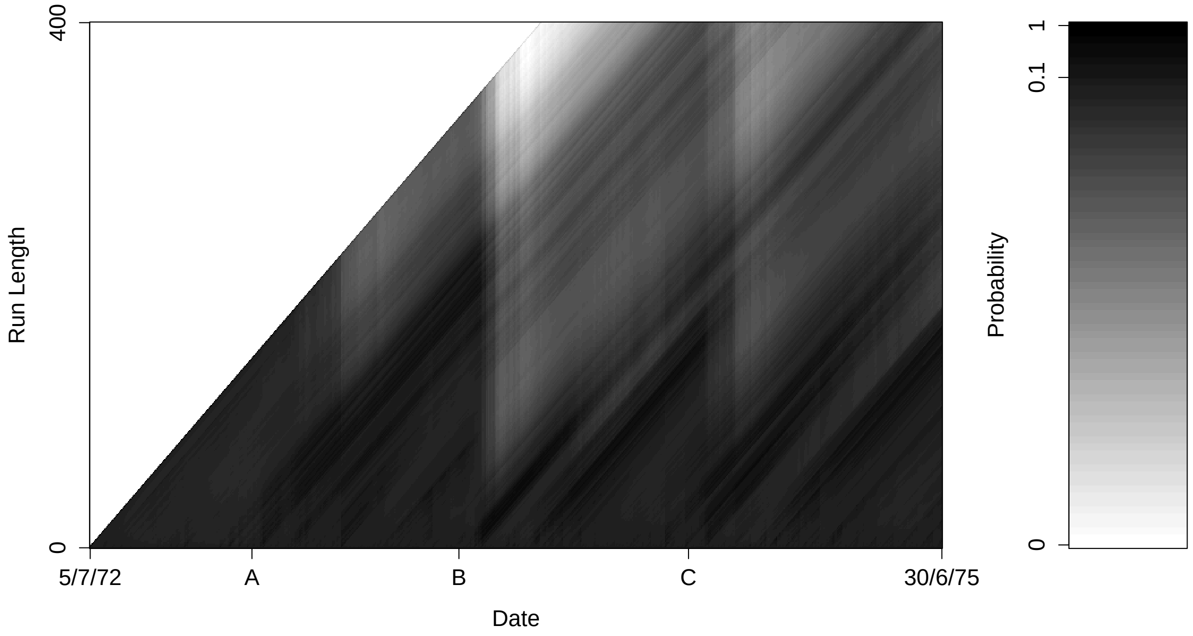

Figure 3.10: Posterior run length distribution at each time step for the daily returns on the Dow Jones Industrial Average as estimated by the conjugate prior based algorithm. A logarithmic colour scale is used with darker cells indicating higher probability. Labels A, B and C refer to the dates of major political events. A: 30th January 1973 - when three of President Nixon's aides were convicted of conspiracy, burglary and wiretapping. B: 19th October 1973 - when OPEC oil embargo began. C: 9th August 1974 - when President Nixon resigned.

3.12 displays the heat map of the posterior run length distribution obtained. There is a clear changepoint at time 127 - corresponding to the waiting time between weeks 2354 and 2526. Adams and MacKay plot their posterior run length distribution in terms of weeks but seem to detect a distinct changepoint at a similar time.

3.2 Introducing Markov Chain Monte Carlo Steps

In order to test the Markov Chain Monte Carlo based changepoint detection algorithm as proposed in Section 2.5 it was applied to the data displayed in Figure 2.1. The aim was to see whether the MCMC based algorithm could replicate the results of the conjugate

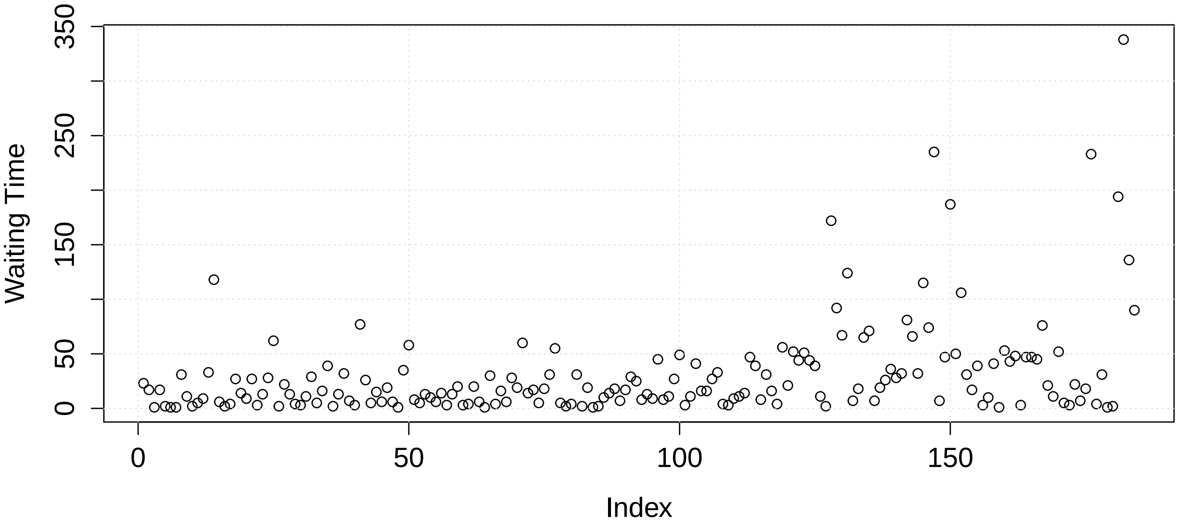

Figure 3.11: Waiting times in weeks between coal mine explosions resulting in more than 10 fatalities between 1851 and 1962.

prior based algorithm. For this reason the values of prior hyperparameters and the hazard function were kept the same ($\mu_0 = 0, \tau_0^2 = 10$, and $\lambda = 18$). The Metropolis-Hastings steps were performed using a normal proposal density, centred on the current state of the chain with variance $\gamma^2 = 4$, to find $M = 10,000$ posterior samples of the mean θ. All M-H chains were initialised with $\theta^{(1)} = 3$. A burn-in of 1,000 iterations was removed, so selected because the chain appeared to have converged a suitably long time before the 1,000th iteration.

The results from applying the conjugate prior based algorithm are shown in Figure 3.2.

Figure 3.13 shows a heat map of the posterior run length distribution as generated by the MCMC based algorithm. The results show how, like the conjugate prior based algorithm, the MCMC based algorithm successfully picks out all the changepoints. The main result from this analysis is that the MCMC based algorithm produces comparable results to those from the conjugate prior based algorithm. Checks were made with several other datasets and similar results suggested that the MCMC based algorithm regularly provides an adequate approximation to Adams & MacKay (2007)'s approach.

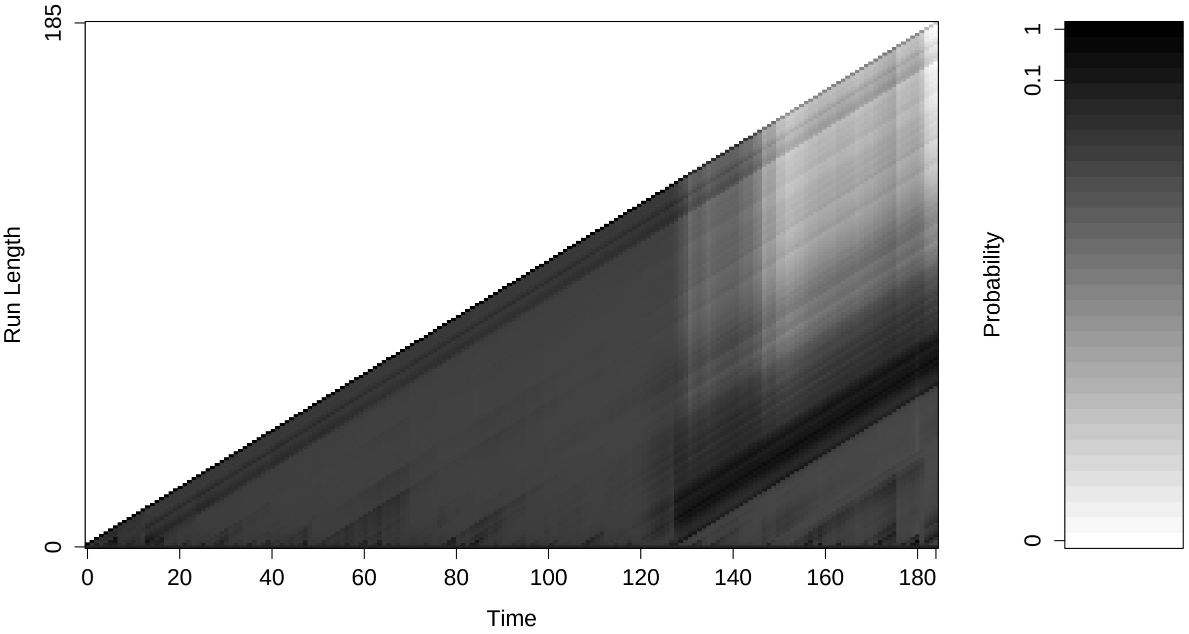

Figure 3.12: Posterior run length distribution at each time step for the coal mine disaster waiting times as estimated by the conjugate prior based algorithm. A logarithmic colour scale is used with darker cells indicating higher probability.

3.3 Harmonic Data

In this section the results of applying the algorithms for harmonic data with both known and unknown angular frequency are given. These algorithms will be applied to the sunspot data in Section 3.4.

3.3.1 Angular Frequency Known

When the angular frequency ω and the variance of the noise term σ^2 are assumed to be known a conjugate prior based approach can be used to analyse data modelled by the simple harmonic model (3) given in Section 2.6. In this section, the conjugate prior based algorithm described in Section 2.6.1 is applied to artificial data generated from the simple harmonic model.

The artificial data was purposely generated from the simple harmonic model to resemble features of the sunspot data and are plotted in Figure 3.14. Each observation y_i is a

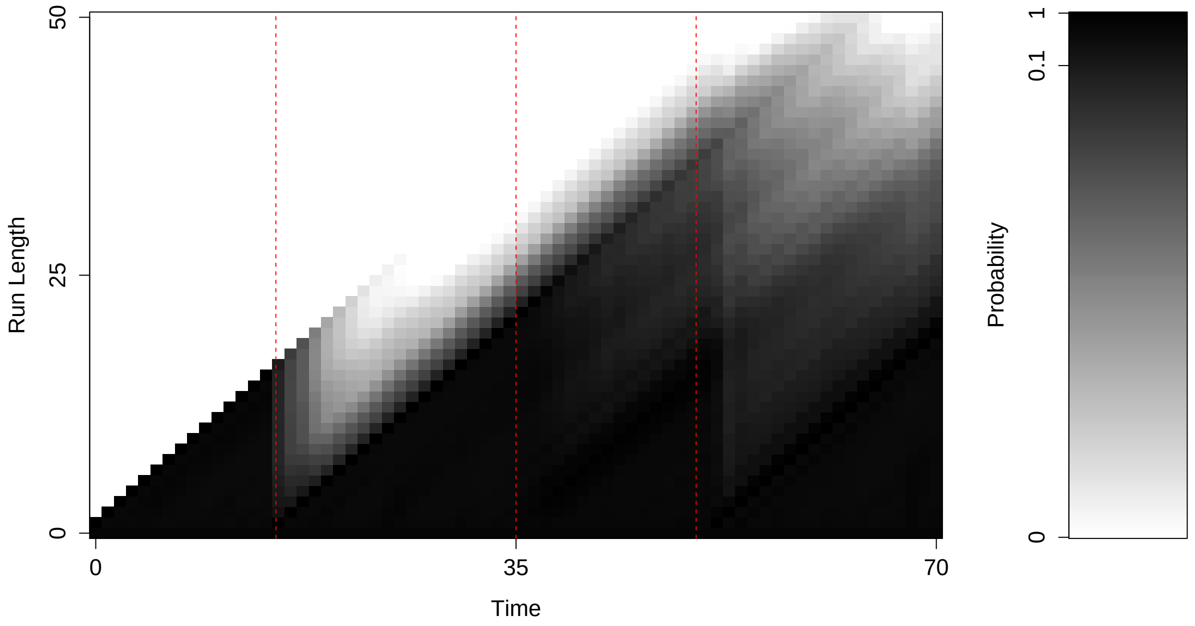

Figure 3.13: Posterior run length distribution at each time step for the changepoint data in Figure 2.1 as estimated by the MCMC prior based algorithm. A logarithmic grey scale is used with darker cells indicating higher probability. Dashed red lines indicate the true changepoints.

realisation of a random variable $Y_i \sim N(C\sin(\omega t_i + \phi), \sigma^2)$ There are 300 observations generated with fixed error variance $\sigma^2 = 1$ and fixed angular frequency $\omega = \frac{2\pi}{22}$, equivalent to a 22 time unit cycle. This angular frequency is chosen to emulate the 22 year cycle of pre-processed sunspot numbers.

The dataset is constructed so that changepoints occur at time 60 and time 90. At time 60 the amplitude parameter C changes from 13 to 6 and the phase parameter ϕ changes from 4 to 1.5. At time 90 the amplitude parameter reverts to 13 and the phase parameter also returns to its original value 4. The parameters A and B therefore take the values $13\cos(4), 6\cos(3/2)$ and $13\cos(4)$, and $13\sin(4), 6\sin(3/2)$ and $13\sin(4)$ respectively.

The conjugate prior based algorithm for data with a harmonic model is implemented with a hazard function parameter $\lambda = 100$ and hyperparameters $\mu_0 = 0$ and $\tau_0^2 = 9$ on the bivariate normal prior described in Section 2.6.1. Figure 3.15 displays the heat map of the posterior run length distribution generated by the algorithm with the true changepoints

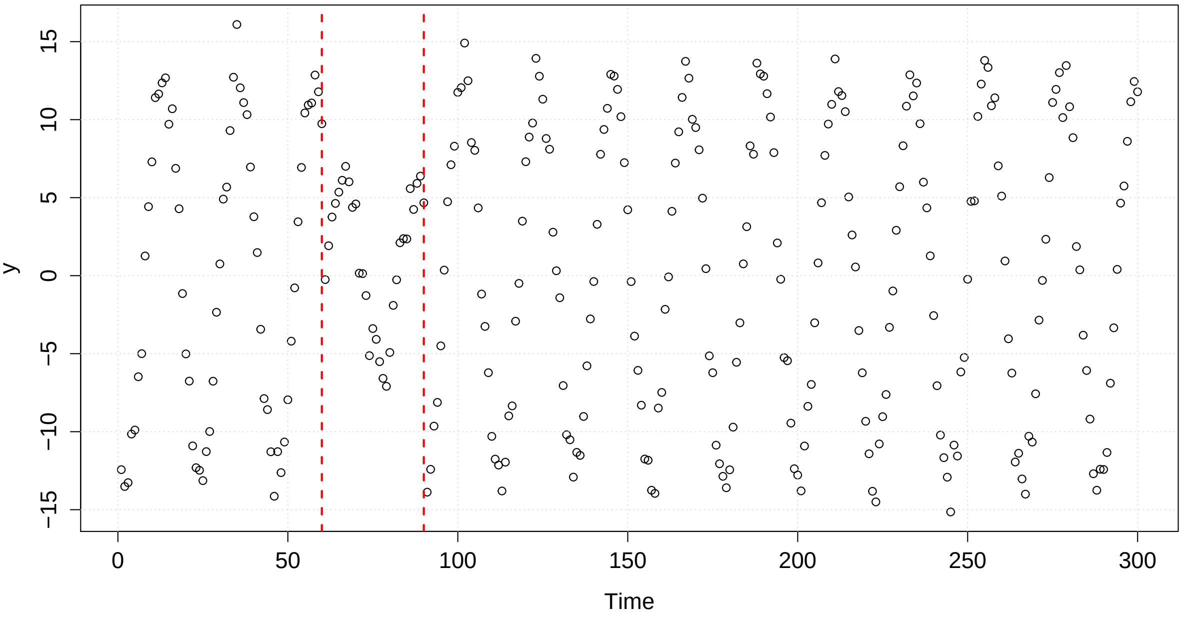

Figure 3.14: 300 observations of harmonic changepoint data with changing amplitude and phase and known, fixed, error variance and angular frequency. Circles represent observations and dashed red lines indicate changepoints.

highlighted once again by vertical red dashed lines.

It is very obvious from the triangular patterns in the heat map that the changepoints are detected with a good degree of confidence. This result provides evidence some that the algorithm described in 2.6.1 is suitable for detecting the changepoints in harmonic changepoint data that can be modelled by the simple harmonic model (3) in the special case when the frequency is known a priori. If the sunspot data belongs to the class of harmonic changepoint data which can be modelled well by this simple harmonic model then it would be hoped that the algorithm could detect changepoints in it. In Section 3.4 the results of applying the algorithm to the sunspot data are provided and discussed.

3.3.2 Angular Frequency Unknown

When it is not assumed that the angular frequency ω is known, a mixed algorithm which uses both MCMC and conjugate prior based steps can be used as described in Section 2.6.2. In this section, the algorithm from Section 2.6.2 is tested by applying it to artificial

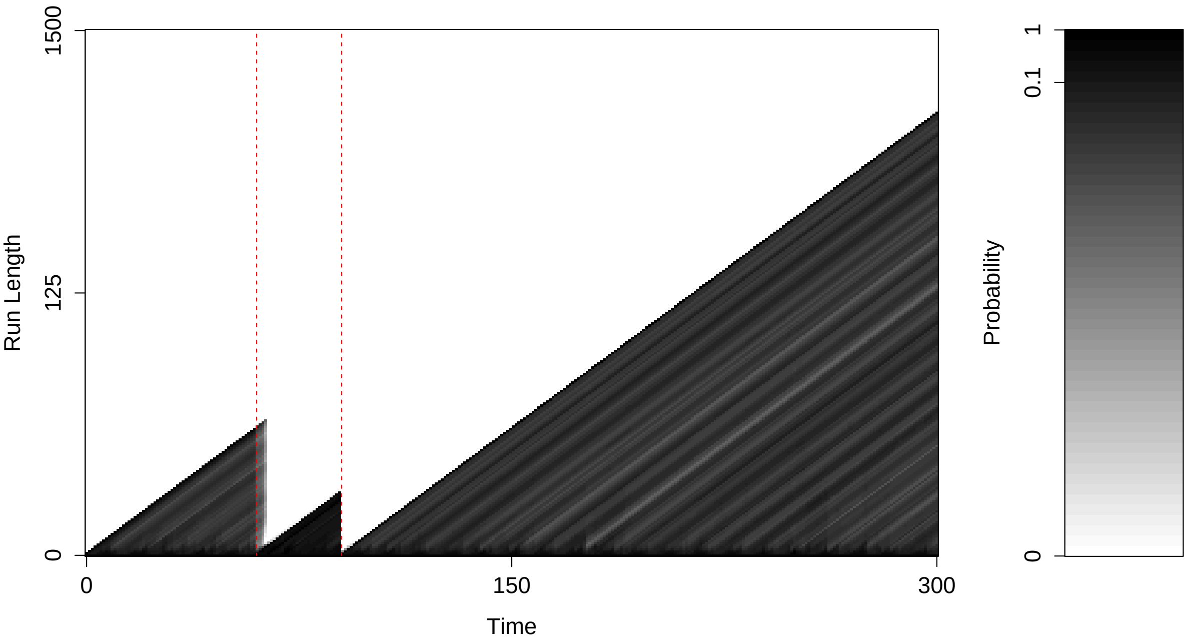

Figure 3.15: Posterior run length distribution at each time step for the changepoint data in Figure 3.14 as estimated by the conjugate prior based algorithm. A logarithmic grey scale is used with darker cells indicating higher probability. Dashed red lines indicate the true changepoints.

data.

The dataset used in this section consists of 40 observations and is plotted in Figure 3.16. The values of C, ϕ and σ^2 are fixed as 13, 4 and 1 respectively. Over the interval [1,20], $\omega = \frac{2\pi}{16}$ and then over the interval [21,40], $\omega = \frac{2\pi}{8}$; so that the only parameter changing at the changepoint is the angular frequency ω - a change which the MCMC steps are needed to detect.

The algorithm as described in Section 2.6.2 was then implemented on this artificial data using the same prior assumptions as in Section 3.3.1, that is, a hazard function parameter of $\lambda = 20$, and prior hyperparameters $\mu_0 = 0$ and $\tau_0^2 = 9$ and the additional assumption of a uniform prior on ω. A normal distribution centred on the current state of the chain with variance $\gamma^2 = 0.01$ was used for the Metropolis-Hastings steps, initialised with $\omega^{(1)} = 0.3$ for all chains. $M = 1000$ iterations were used to calculate each predictive probability, with a burn-in of 100 observations being removed. Ideally, a larger number of iterations would

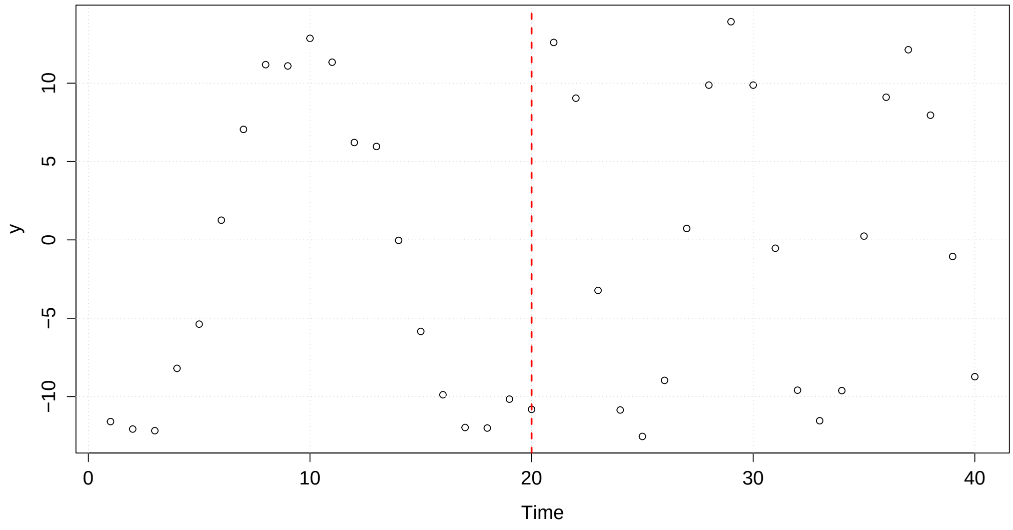

Figure 3.16: 40 observations of harmonic changepoint data with changing angular frequency, and known, fixed, amplitude, phase and variance. Circles represent observations and dashed red lines indicate changepoints.

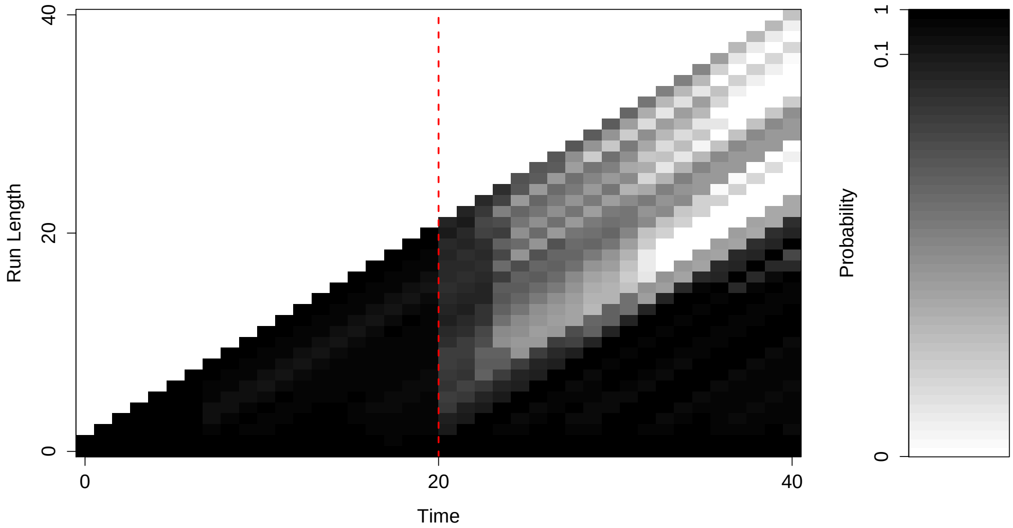

Figure 3.17: Posterior run length distribution at each time step for the changepoint data in Figure 3.16 as estimated by the mixed algorithm. A logarithmic grey scale is used with darker cells indicating higher probability. Dashed red lines indicate the true changepoints.

be used to give a more accurate approximation in the Monte Carlo integration. The values selected were sufficient in this case, however, to illustrate the working algorithm.

Figure 3.17 shows the heat map of the posterior run length distribution generated by running the mixed algorithm on these artificial data. It is very clear from the figure that the changepoint is detected at the correct point, as the position of the darkest cells drops almost immediately after the dashed red line indicating the change.

3.4 Sunspot Data

This final section of results gives the results of applying the two harmonic data algorithms from Section 2.6 to the sunspot data pre-processed as described in Section 2.7.

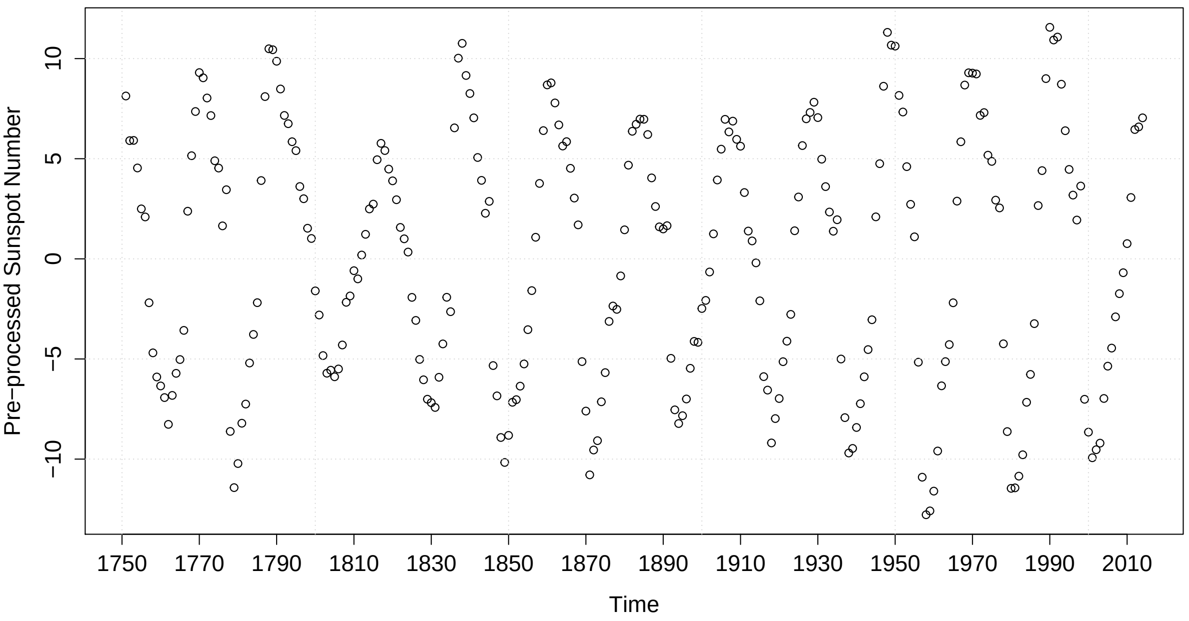

Figure 3.18: 264 pre-processed smoothed yearly average values of Wolf's Sunspot Index, between the years 1750 and 2013. The data are plotted as circles and were pre-processed as described in Section 2.7.

Figure 3.18 shows the pre-processed sunspot data. The Great Solar Anomaly region from 1790-1830 still appears to be a region of inhomogeneity.

Firstly, the conjugate prior based algorithm was applied to the data with the assumed

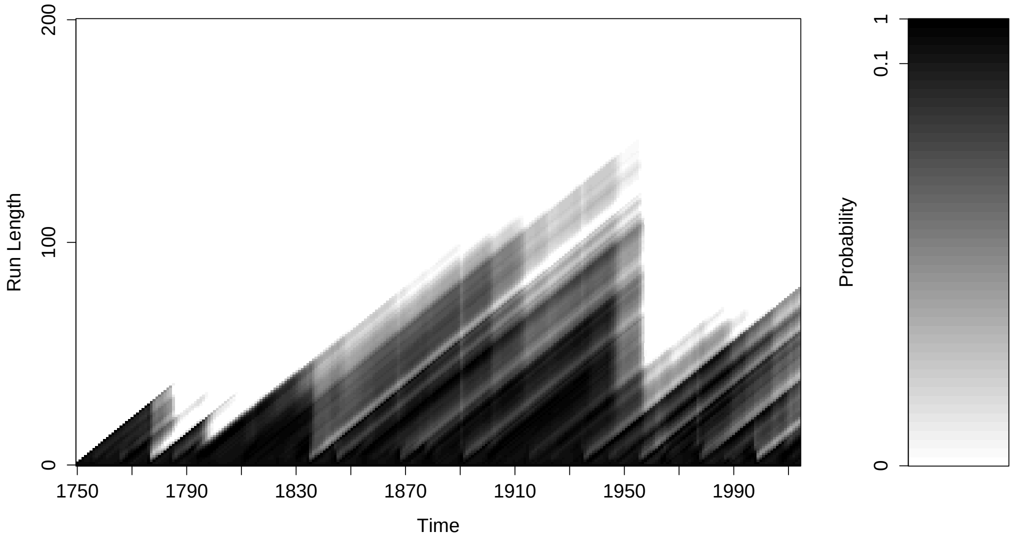

Figure 3.19: Posterior run length distribution at each time step for the pre-processed sunspot data in Figure 3.18 as estimated by the conjugate prior based algorithm. A logarithmic grey scale is used with darker cells indicating higher probability.

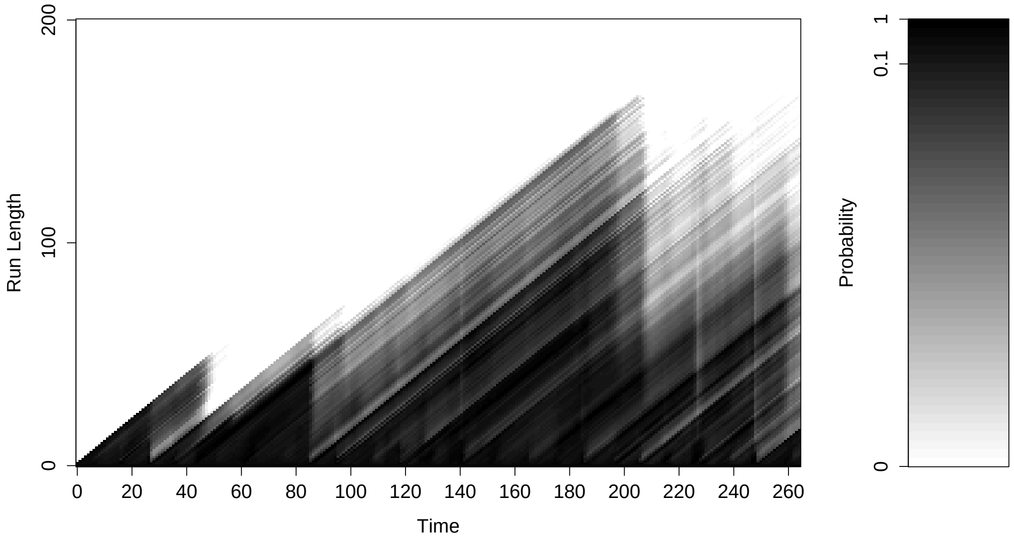

Figure 3.20: Posterior run length distribution at each time step for the pre-processed sunspot data in Figure 3.18 as estimated by the mixed algorithm. A logarithmic grey scale is used with darker cells indicating higher probability.

fixed values $\omega = \frac{2\pi}{22}$ and $\sigma^2 = 1$. As described in Section 2.6.1 a bivariate normal prior was used in the algorithm with parameters $\mu_0 = 0$ and $\tau_0^2 = 9$. The hazard function parameter λ was set at 100. The heat map of the posterior run length distribution is shown in Figure 3.19.

The results presented are not entirely clear in this case, and as such are open to some interpretation. The smallest dark triangles along the bottom of the graph suggest that with every new cycle a changepoint seems possible, with runs then seeming to persist at least until the start of the next cycle. Some runs however seem to persist beyond this and it would seem in particular that there is quite a clear run over the region 1790-1830. This is the Great Solar Anomaly region and suggests that there may be changepoints both around 1790 and around 1830. After this point there seems to be a longer run from 1830-1950 although there is less clarity in the results after this region with seemingly feasible runs starting from 1910 and 1930. It is difficult to say with any certainty where the next changepoint after the 1830 one occurs.

The mixed MCMC and conjugate prior based algorithm was next implemented as described in Section 2.6.2. Again, the assumption was made of a known error variance $\sigma^2 = 1$. The same prior assumptions were also made: $\mu_0 = 0, \tau_0^2 = 9$ and $\lambda = 100$. As each step involves the calculation of various matrix products, inverses and determinants, using matrices which become larger with each iteration of the algorithm, a less than desirable number of iterations was again used for the Monte Carlo integrations. The number of iterations M was chosen as 500 with a burn-in of 100 observations being removed from the start of each chain. Trace plots did not rule out the hypothesis that the chains had converged within this time, but nevertheless it was less than ideal. As with the artificial example, all of the chains were initialised with $\omega^{(1)} = 0.3$ and a normal proposal distribution centred on the current state of the chain with variance $\gamma^2 = 0.01$ was used. A small variance was chosen because Macaulay (1992) suggests that the distribution M-H samples are being drawn from is very sharply peaked.

Figure 3.20 shows the heat map of the posterior run length distribution generated by the mixed algorithm. It is largely similar to the distribution produced by the conjugate prior based algorithm. As Figure 3.20 does not detect any changepoints that are not detected in Figure 3.19 it suggests that either there are no changepoints in the parameter ω or that if there are they are too subtle to be detected by the algorithm with this selection of prior parameters.

4 Discussion

This project has been concerned with the Bayesian and online detection of changepoints. Work has centred on the replication and generalisation of Adams & MacKay (2007)'s conjugate prior based changepoint detection algorithm. Applying the algorithm to various datasets, real and artificial, showed that it can detect changes in a variety of parameters but some changepoints are detected with more or less confidence depending on the features of the surrounding observations.

Adams and MacKay's algorithm can be praised for its efficiency, accuracy and structure which allows code to be altered for different exponential family distributions quickly and easily. A criticism however was that the algorithm was not appropriate for data modelled by a likelihood who lacks a conjugate prior distribution.

This project has further developed Adams and MacKay's algorithm with the introduction of MCMC steps in place of conjugate prior dependent steps. Doing so has broadened the scope of the algorithm, allowing it to be used for a greater range of models and ultimately increasing its value as an online signal processing tool.

Two algorithms were developed to be applied to data from a simple harmonic model (3): one based on Adams and MacKay's algorithm, which was appropriate when the angular frequency and error variance of the model could be assumed to be known, and a second which combined elements of the conjugate prior and MCMC approaches and only required the error variance to be known. Both algorithms were applied to artificial changepoint data and found to accurately detect changepoints.

Finally, these two algorithms were applied to a pre-processed set of Sunspot Index data with the intention of detecting changepoints. The various pre-processing steps were undertaken to make the Sunspot Index data more similar to the data for which the two algorithms had been specifically designed, however, every step that was taken could be justified in the context of the underlying magnetohydrodynamic activity. Both algorithms presented some evidence of changepoints around the years 1790 and 1830, but no clear indication as to the location of any other changepoints, although the low probability of large run lengths over the last 50-100 years suggests that some sort of parameter change must have occurred somewhere.

The aims of this project, as laid out in Section 1 have been met successfully. There are however, a number of areas for improvement, further thought, and further work. Principally some discussion should be given as to why more clear cut results could not be obtained when the algorithms were applied to the Sunspot Index data:

Having demonstrated that both the conjugate prior and MCMC based algorithms work for artificial data suggests that the theory is watertight and both algorithms are effective. Assuming that this is the case, two main explanations as to why the algorithm did not detect changepoints more confidently in the sunspot dataset remain: either the model used does not adequately fit the data, and as a result the algorithm is inappropriate, or the dataset does not contain true changepoints.

In the writer's opinion the mostly likely case is that both of these explanations are in part true. The underlying physics with which one might attempt to explain the full complexities of patterns in solar activity are not at all trivial; consisting of highly non-linear partial differential equations derived from magnetohydrodynamics. As a result it has been rather optimistic to expect a simple model based on a single trigonometric function to accurately capture the features of the dataset.

Individual cycles have been studied extensively by astrophysicists and all have their own characteristics (Clette et al., 2014). As a result when a simple likelihood model is used, it may be the case that the algorithm will see an 'extreme' observation with low predictive probability and incorrectly detect a changepoint. When in fact there is not a true changepoint at that time but a model fit issue. If this is what is happening it may explain the large number of detected changepoints and their positioning at similar points within each cycle.

Another issue which comes under the umbrella of model fitting issues is the consistent assumption that σ^2 was known. This is not the case and its value is likely important. Improvement could be made to the algorithm by placing an Inverse-Gamma conjugate prior on σ^2 to capture the uncertainty.

As well as using a more complex likelihood function, and removing assumptions of known parameter values the algorithm's performance could likely be improved by increasing the number of iterations, M, at the M-H steps. In the final example only 500 were used because of the complexity of the algorithm. In a situation with more computing time available, a larger number of iterations could provide more accurate calculation of predictive probabilities and as a result more reliable inference. Formal tests should also be undertaken to assess the robustness of priors - whether the algorithm still works with different hyperparameter choices - and assess non-convergence in the M-H steps. A test on a Gelman-Rubin statistic could be used, for instance (Gelman & Rubin, 1992).

A further explanation for the issues may be that true changepoints (abrupt changes in generative parameters) simply do not exist. It may be the case that underlying parameters change gradually over a period of time. Clette et al. (2014) suggest that a similar period of low solar activity, the *Maunder Minimum*, occurring in the 1600s came about after a 'progressive decline' in activity rather than an abrupt change. If this period is typical of other inhomogeneous regions it could be the case that the changes leading in to certain periods of low solar activity are simply not abrupt enough to be picked up by the algorithm. This could be why the algorithm experiences some confusion over detecting changepoints in the last 100 years. Clette et al. (2014)'s results should, however, be interpreted with caution as the earlier records of Sunspot Index data (on which their conclusions are based) are known to be somewhat unreliable (Sonnet, 1983).

All in all, while there is some evidence of changepoints around 1790 and 1830, which agrees with scientific theory about the Great Solar Anomaly, and while there is some suggestion that abrupt changes simply do not occur around certain other low activity regions the main plausible explanation for issues in the detection of changepoints remains the algorithm itself. Therefore it would be worthwhile to conduct further study using a larger number of iterations in M-H steps, and perhaps a more complicated likelihood function but certainly without the assumption of any known parameter values.

With more time it could have been worthwhile to explore more flexibility in some of the prior assumptions. For instance, particularly in the real data examples, different assumptions about the hazard function could have been made. As the hazard function is essentially a representation of the prior belief about the distribution of changepoints, further research could have been done to find some specialist opinion on the frequency and patterns of changepoints for each situation and this could have been incorporated somehow. For example, if it were the case that layers of different types of rock in the earth's crust tend to be above a certain thickness, the hazard function could have reflected this by assuming a changepoint is less likely when run length is small. If it had been observed that in other financial markets several changes in volatility tended to happen in quick succession, the hazard function could have been altered to give a higher probability of a changepoint at low run lengths.

In the case of the Sunspot data however it did seem appropriate to make prior assumptions that were as uninformative as possible. While sunspots do occur on other stars, they are in this case referred to as *starspots*, the information available on their numbers and patterns is limited. Technology has only recently (around the mid 1990s) developed to a level where accurate records can be kept. This means that scientific opinion on the pattern of changepoints is mostly based on the very data on which analysis is being performed. Incorporating some of this information into a more informative prior is akin to using the information twice and is quite rightly seen as very bad practice within Bayesian statistics.

Finally, both Adams and MacKay's algorithm and the MCMC based algorithm could be improved by introducing some more formal means of quantifying the confidence with which a changepoint is detected. Throughout Adams and MacKay's report and this project, changepoints are reported based on a somewhat subjective analysis of a visual representation of the posterior run length distribution. Some rule could be established so that the algorithm automatically reports a changepoint when the difference between run lengths representing a changepoint and growth passes a certain threshold, with the threshold being defined based on some pre-selected significance level. This more objective approach would allow for more consistent analysis.

References

R.P. Adams & D.J. MacKay, 2007, *Bayesian Online Changepoint Detection*, (Technical Report) arXiv:0710.3742v1 [stat.ML], University of Cambridge.

D. Barry & J.A. Hartigan, 1993. A Bayesian Analysis of Changepoint Problems. *Journal of the American Statistical Association*, **88**, pp309-319.

C.M. Bishop, 2008, *Pattern Recognition and Machine Learning*, Springer, New York.

R.N. Bracewell, 1953. The Sunspot Number Series. *Nature*, **171**, pp649-650.

S.Chib, 1998. Estimation and Comparison of Multiple Changepoint Models. *Journal of Econmetrics*, **86**, pp221-241.

F. Clette, L. Svalgaard, J.M. Vaquero & E.W. Cliver, 2014. Revisiting the Sunspot Number. arXiv:1407.3231 [astro-ph.SR].

F. Desobry, M.Davy & C. Doncarli, 2005. An Online Kernel Change Detection Algorithm, *IEEE Transactions on Signal Processing*, **53**, pp2961-2974.

P. Fearnhead & P. Clifford, 2003. Online Inference for Hidden Markov Models via Particle Filters. *Journal of the Royal Statistical Society B*, **65**, pp887-899.

R.A. Fisher, 1922. On the Mathematical Foundations of Theoretical Statistics. *Philosophical Translations of the Royal Society A*, **222**, pp309-368.

A. Gelman & D.B. Rubin, 1992. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, **7**, pp457-472.

A. Gelman, J. B. Carlin, H.S. Stern & D.B. Rubin, 1995. *Bayesian Data Analysis*, Chapman, London.

P. Green, 1995. Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika*, **82**, pp711-732.

U. Grenander, 1959. *Probability and Statistics: The Harald Cramer Volume* Wiley.

W.K. Hastings, 1970. Monte Carlo Sampling Methods Using Markov Chains and Their Applications, *Biometrika* **57**, pp97-109

P.D. Hoff, 2009. *A First Course in Bayesian Statistical Methods*, Springer, New York.

R.G. Jarret, 1979. A Note on the Intervals between Coal-mining Disasters. *Biometrika*, **66**, pp191-193.

G. Lorden, 1971. Procedures for Reacting to a Change in Distribution. *The Annals of Mathematical Statistics*, **42**, pp:1897-1908.

V.A. Macaulay, 1992. *Bayesian Inversion with Applications to Physics.* Ph.D. University of Oxford.

K.P. Murphy, 2007. *Conjugate Bayesian Analysis of the Gaussian Distribution*, (Technical Report) University of British Columbia.

E.S. Page, 1954. Continuous Inspection Schemes, *Biometrika* **41** pp100-115

E.S. Page, 1955. A Test for a Change in Parameter Occurring at an Unknown Point, *Biometrika* **42** pp523-527

R Core Team, 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

C.P. Robert, 2001. *The Bayesian Choice*, Springer, New York.

C.P. Sonnet, 1983. The Great Solar Anomaly ca. 1780-1800: an Error in Compiling the Record?, *Journal of Geophysical Research* **88A**, pp3225-3228.

D.A. Stephens, 1994. Bayesian Retrospective Multiple Changepoint Identification. *Applied Statistics*, **43**, pp159-178.

S. Wagner & E. Zorita, 2005. The Influence of Volcanic, Solar and CO_2 forcing on the temperatures in the Dalton Minimum (1790-1830): a model study, *Climate Dynamics.* **25**, pp205-218.

WDC-SILSO, 2014. *Monthly Mean Total Sunspot Number.* [Data ¿ Data Files] Royal Observatory of Belgium, Brussels.