Non-Parametric Online Change Point Detection via adapting an Established Method

Z. Todorova
School of Mathematics, University of Bristol, Bristol, United Kingdom

May 17, 2022

Abstract

We consider the multiple online change point detection method proposed in Fearnhead and Liu (2007) and propose a series of adaptations to make it non-parametric, with considerations for speed and storage. We will then compare this to the original method and other online change point detection methods on simulated data.

In a simulation study, we implement and compare the proposed variations, and show that estimating the mean and the distribution of change points produces greatest improvement. Using simulated data, we compare with PELT and Binary Segmentation, and show that despite improvements, these methods remain more reliable in tests.

Keywords: Online change-point detection, Non-parametric methods

1 Introduction

With large amounts of data being generated each day, and this amount only set to grow over the coming years (Sagiroglu and Sinanc, 2013), detecting sudden changes known as *change points*, in data streams allows an immediate reaction to better deal with these changes. Change point detection methods can be used in medical monitoring (Bosc *et al.*, 2003), speech recognition (Rybach *et al.*, 2009), climate change and environmental tracking (Ducré-Robitaille *et al.*, 2003), and also in financial time series and other domains (Frisén, 2009).

Many methods look for a change in a single parameter, such as mean or variance. In addition, in literature there is a strong focus on *offline* change point detection methods, where the method looks to find change points in a complete series of data points. Often, change point detection (CPD) methods are published and built on by other researchers, for example Optimal Partitioning (Jackson *et al.*, 2005), to FPOP (Maidstone *et al.*, 2017), to R-FPOP (Fearnhead and Rigaill, 2019), which add pruning and recursion steps respectively. This allows more rapid improvement in the field as techniques are shared and compared, at the cost of new methods being developed. In this report, we look at the change point detection method proposed by Fearnhead and Liu (2007) and propose some improvements. We then implement the modifications proposed and compare them to each other and some of the existing methods described in Section 2.

In Section 2 we will give some key definitions and concepts, as well as an overview of several other change point detection methods. Section 2.5 explains the Fearnhead Liu 2007 method in detail as well as proposes several adaptations that may improve its performance. There is a discussion on methods for evaluating change point methods in section 4. A simulation study comparing the performance of some of the proposed adaptations with each other and with other leading methods is presented in Section 5, along with a comparison to some of the methods. Section 6 summarises and concludes our results.

2 Background

In this section, we begin by formalising the concept of a change point, before looking at a selection of well established change point methods and comparing them, including the Fearnhead and Liu (2007) method that is considered further in Section 3.

Formally, for an ordered sequence of observations $(y_i)_{i=1}^n$ with non-identical distributions F, G where

$$y_i \sim \begin{cases} F & \text{for } i = 1, ..., \tau \\ G & \text{for } i = \tau + 1, ..., n \end{cases}$$

there is a change point at τ. In this example there is exactly one change point, we can generalise to k change points as follows. For $s \leq t \leq n$, a segment of the series from time s to time t can be written as $y_{s:t} = (y_s, ..., y_t)$. If we assume y contains k change-points this corresponds to the data being split into $k + 1$ distinct segments. We denote the ordered set of the change points in y as $\tau = \tau_0, \tau_1, \tau_2, ..., \tau_k$, where τ_j is the j^{th} change point in the series, we set $\tau_0 = 0$ and $\tau_{k+1} = n$. The j^{th} segment consists of points $y_{\tau_{j-1}+1}, ..., y_{\tau_j}$.

This definition is very open about the distributions F and G, however most existing change point methods will deal with only one change in parameter at each change point, for instance mean, variance, or rate, though some advanced methods may be able to look at change in distribution as well (Zhou *et al.*, 2017).

We can differentiate and categorise change point detection methods by looking at whether they are online or offline, univariate or multivariate, model-based or non-parametric. Of these, the feature that has greatest impact on the use of the method is whether it is offline or online. In *offline* change point detection, a method will look back on existing and historically completed data, aka there is a time series of known length N, and one can look back at the entire time series and decide whether there are any change points, and if so where they are. These often use hypothesis testing methods (Hinkley and Hinkley, 1970). An *online* method will attempt to detect change points in an incoming data stream as each new time point comes in as close as possible to the true time of the change point. In online methods the sooner the change is detected, the heavier the trade-off between Type I error, the probability of detecting a change point when there is no change point and Type II error, the probability of not detecting a true change point.

A univariate method will look at only one variable, whereas a multivariate method will look at more, for example looking at means across multiple simultaneous and possibly interrelated processes where change points may occur at similar times, or looking at multiple parameters from the same process (Bardwell, 2018). Parametric methods assume the distribution of the data is known, whereas non-parametric methods do not make these assumptions and can attempt to find a distribution by estimating parameters. Non-parametric models can be prone to over-fitting, however are better and handling new and unfamiliar data sets. (Aminikhanghahi and Cook, 2017; de Rodriguez, 2020). Model-based methods can be either Frequentist (CUSUM) or Bayesian (Fearnhead Liu).

2.1 Cost Functions and Penalties

In order to find the most likely position for a change point, many methods seek to minimise the global cost function for each segment that the change points divide the data into. The global cost function can be written as:

$$G(m, \tau_1, ..., \tau_m) = \sum_{i=1}^{m+1} C(y_{(\tau_{i-1}+1):\tau_i}) + \beta \tag{1}$$

Where C is the cost function of the segment bounded by the time points $\tau_{i-1} + 1 : \tau_i$. Potential choices of this cost function include quadratic or Squared Error loss

$$C(y_{s:t}) = \sum_{i=s}^{t} (y_i - \bar{y_{s:t}})^2$$

which is good for cleaned data sets and detecting changes in the mean, however it can struggle to distinguish between changes and outliers Fearnhead and Rigaill (2019). Bi-weight loss was proposed as a loss function more able to deal with both changes in mean and outliers, as it caps the loss and means that extreme values do not have an overwhelming effect on loss.

$$C(y_{s:t}) = \begin{cases} (y_i - \bar{y_{s:t}})^2 & \text{if } |y_i - \bar{y_{s:t}}| < K \\ K^2 & \text{otherwise,} \end{cases}$$

Another option is absolute loss

$$C(y_{s:t}) = \sum_{i=s}^{t} |y_i = \bar{y_{s:t}}|$$

. Another common segment loss function is twice the negative log likelihood, however this is only suitable for cases in which we have a likelihood based model before and after the change point.

In (1) β is a penalty function used to prevent over fitting; as if $\beta = 0$ the minimal solution in (1) is the one where each data point is its own segment, the penalty function ensures that this is not the best solution. Penalty functions include the Bayesian Information Criterion (BIC), also known as Schwarz's criterion, (Schwarz, 1978) (β = p log n) where p is the number of parameters introduced to the model by adding a single change point, and alternatively Akaike's Information Criterion (AIC, Akaike (1974)) (β = 2p). The penalty function is chosen according to the desired result and assumptions made. The BIC will hone in on the true model in a set of possible candidate models and may tend to under-fitting because of its focus on penalising additional parameters, whereas the AIC will try and select the model that is the best fit for data that may not have a model of perfect fit, so can over-fit to an excessively complex model (Stoica and Selen, 2004). As well as other parametric methods, there are also heuristic methods which aim to find a good solution as opposed to the best possible solution such as Lavielle (2005), where the quadratic risk is estimated until adding a change point does not significantly reduce this quantity.

2.2 CUSUM

CUSUM Page (1954) monitors for subtle changes in the mean of the probability distribution by tracking the Cumulative Sum of the deviations from a known initial mean with known deviation over time. For $(y_i)_{i=1}^n \sim \mathcal{N}(\mu, \sigma^2)$, the statistic is calculated as follows:

$$\begin{aligned} S_0 &= 0 \\ S_{n+1} &= max(0, S_n + y_n + \omega_n) \end{aligned}$$

Where ω_i is the weight assigned to each value which is commonly set as the likelihood value of y_i, however the ω_i can also equal $(\mu - k)$ where k is the "allowable value" and is often chosen as halfway to a value that is considered a significant or unacceptable change. So $k = 1/2 \times (\mu - m_1)$ where m_1 is the minimum mean value of an unallowable change (Koshti, 2011).

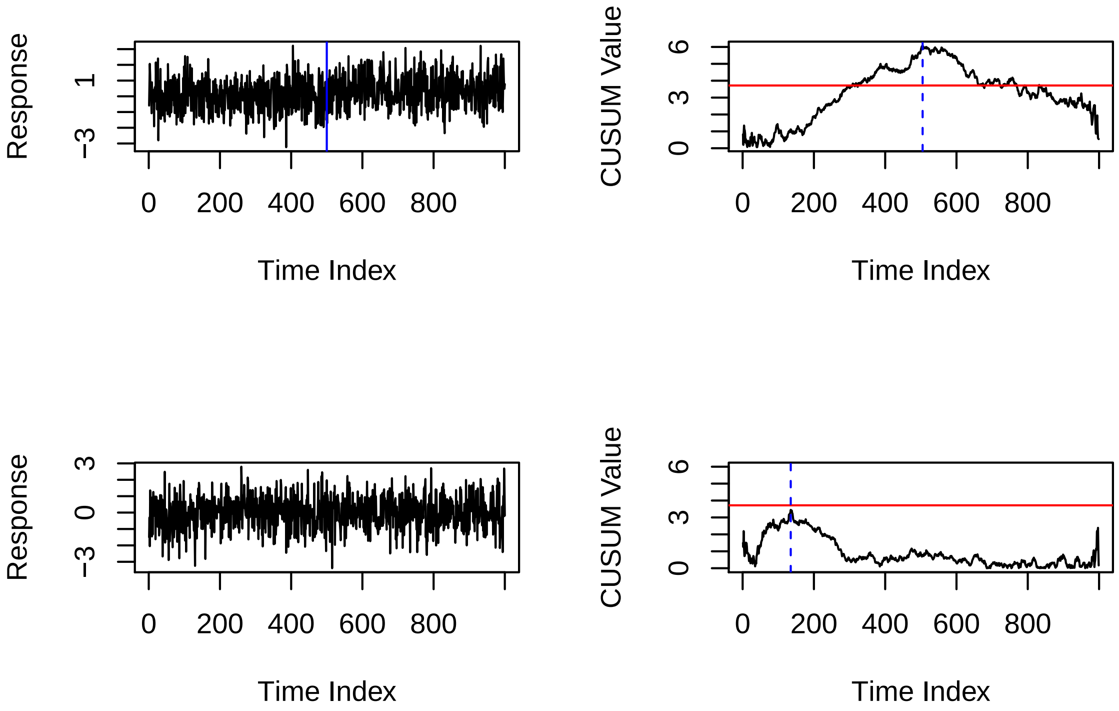

Figure 1: A series undergoing a small change in mean at time $\tau = 500$ (top left) exhibits a peak in the CUSUM statistic values at $\hat{\tau} = 505$. This peak is above the threshold (shown in red) for declaring a change point (top right). Meanwhile, a sequence with no change (bottom left) exhibits no CUSUM values above the same threshold (bottom right), so no change point is estimated here. Figure from unpublished lecture notes for this project.

The above statistic only looks at increases in mean, however decreases can be calculated with an alternative statistic $S_n - \omega_i - y_i$. If either of these test statistics exceeds a threshold h, for example

5σ, we deem that there has been a change. Since CUSUM is particularly good at detecting shifts in mean of less than 1.5σ, μ is often used as the target or ideal value of the series in medical and manufacturing contexts where small fluctuations in mean can have significant impact on outcomes. Its use is restricted to a change in mean setting of a Gaussian distribution with known variance. It can also detect a single change point only if it peaks at least above the threshold h.

2.3 Binary Segmentation

Where CUSUM is able to detect a single change in mean in an online setting, Binary Segmentation (BinSeg) looks for multiple changes across offline data by partitioning the time series around each single change point it finds using a given statistic such as CUSUM, and once again checking for change points in the partitioned data. Initially it will test to see if a change point t exists, using our chosen test statistic. If no change point is found, then we stop. Else, we split the data into two segments: before t, and after t. We then once again look to see if a change point exists using our chosen test statistic. This repeats until no further change points are detected.

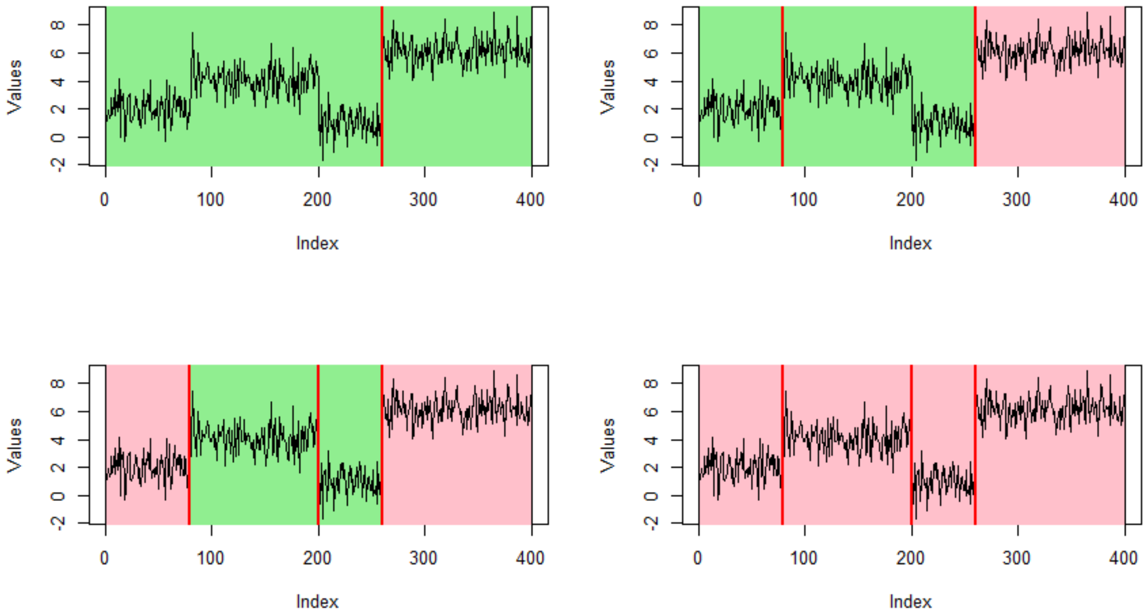

Figure 2: An example of Binary Segmentation applied to a series with three changes in mean. A single pass through the sequence with CUSUM finds it maximised at index i = 260, so a change point is placed there and we segment around this point(top left. We repeat on the data either side of index 260, and find CUSUM is maximised at i=80, but no change point after i=260, so the segment [260, 400] is subsequently ignored (top right). We now consider the segments [0,80) and [80,260], where we find no change points and a significant maximum at i=200 (bottom right). We find no additional change points in the segments [80,200) or [200,260) and conclude (bottom right)

This is a fairly computationally efficient method, resulting in an O(C*nlogn) calculation, with C the computational cost of the cost function, however it can miss smaller change point intervals that it may deem to have a comparably insignificant effect on the test statistic.

2.4 PELT

Killick *et al.* (2012) proposed the Pruned Exact Linear Time (PELT) method which refined the method of Jackson *et al.* (2005) by adding a pruning step that "reduces the computational cost of the method, but does not affect the exactness of the resulting segmentation". It aims to remove the data points that can never be minima of the minimisation performed at each iteration in (1). The method assumes that there is a fixed minimum decrease in cost function for each additional change point, aka a minimum penalty. This is important as it is what allows the method to be pruned and skip computations that will not be necessary to get the exact final list of change points. It is one of the methods that seeks to minimise the global cost function (1), given as $C(.)$ below.

Algorithm 1: The PELT method (Killick *et al.*, 2012)

```
Data: A set of data of the form, $(y_1, y_2, ..., y_n)$ where $y_i \in \mathbb{R}$
      A cost function $C(.)$ dependent on the data
      A penalty value, $\beta$
      A constant $K$, such that for $t < s < T$, $C(y_{(t+1):s}) + C(y_{(s+1):T}) + K \leq C(y_{(t+1):T})$.
Result: A set of estimated change point locations $\hat{\tau_1}, ..., \hat{\tau_m}$
Initialise: n = length of data; $F(0) = -\beta$; $cp(0) = NULL$; $R_1 = \{0\}$;
for $\tau^*$ in 1, ..., n do
    Calculate $F(\tau^*) = min_{\tau \in R_{\tau^*}} \{F(\tau) + C(y_{(\tau+1):\tau^*}) + \beta\}$ ;
    Set $\tau_1 = argmin_{\tau \in R_{\tau^*}} \{F(\tau) + C(y_{(\tau+1):\tau^*}) + \beta\}$ ;
    Set $cp(\tau^*) = [cp(\tau^1), \tau^1]$;
    Set $R_{\tau^*+1} = \{\tau \in R_{\tau^*} \cup \{\tau^*\} : F(\tau) + C(y_{(\tau+1):\tau^*}) + K \leq F(\tau^*)\}$
end
```

Although the original version of this method is offline and takes all the data as an initial parameter, as this method steps through the each of the n data points individually, it would be simple to convert it to an online method by iterating as each new data point is revealed. Though at worst case PELT remains of O(n^2) time, like its parent Optimal Partitioning, under a set of conditions (such as the number of change points increasing linearly with the length of the data) its pruning step means that the true time complexity of PELT is in fact O(n).

2.5 Fearnhead Liu

This project focuses on the exact online inference method that was proposed in Fearnhead and Liu (2007). They proposed two variations on a method, an exact online inference method, which aims to find the exact time of a change point, and an approximate method which aims to control the computational and memory costs while producing adequate approximations of the true change points.

Suppose that we have collected t data points of a series to date, such that t is effectively the "current time". Define C_t to be the time point of the most recent change point strictly prior to time t. If there have been no change points before time t, we take $C_t = 0$. This means $C_t \in \{0, 1, ..., t-1\}$ and $C_1, C_2, ..., C_t$ is a Markov Chain. The method also uses $g(.)$ taken as prior information, where $g(.)$ is the probability mass function for the distance between successive change points. Since we are assuming the times of the change points are a Markov chain, the process is memory-less and everything prior to the most recent change point can be forgotten. This means that a geometric distribution for $g()$ is sensible. Looking at the parameters of g can give us some indication of the volatility and frequency of change points in the time series. From this we can calculate

$$G(n) = \sum_{i=1}^{n} g(i)$$

as the cumulative distribution function for the distance between two successive change points.

We can then use G to calculate the transition probabilities of the Markov chain by conditioning on the previous state $C_t = j$, in the next state C_{t+1} which occurs after the $t + 1^{th}$ data point, we can either have placed no change point at time t meaning that $C_{t+1} = C_t = j$, or we can have placed a change at time t so that $C_{t+1} = t$. If we have a change point, the probability that $C_{t+1} = t$ is proportional to the probability that the change point occurs exactly $t - 1$ time points since the previous change point, which we can write as $G(t - i) - G(t - i - 1)$. If there is no change point, the probability that $C_{t+1} = j$ is proportional to the probability that a change point occurs more than t time points after the previous change point, which we can write as $1 - G(t - i)$. This allows us to express the transition probabilities of the Markov chain as follows:

$$\mathbb{P}(C_{t+1} = j | C_t = i) = \begin{cases} \frac{1-G(t-i)}{1-G(t-i-1)} & \text{if } j = i \\ \frac{G(t-i)-G(t-i-1)}{1-G(t-i-1)} & \text{for } j = t \\ 0 & \text{otherwise.} \end{cases}$$

If we have already observed data points $y_{1:t}$ we can use the law of total probability to get

$$\mathbb{P}(C_{t+1} = j | y_{1:t}) = \sum_{i=0}^{t-1} \mathbb{P}(C_{t+1} = j | C_t = i)\mathbb{P}(C_t = i | y_{1:t})$$

And therefore

$$\mathbb{P}(C_{t+1} = j | y_{1:t+1}) \propto \mathbb{P}(y_{t+1} | C_{t+1} = j, y_{1:t})\mathbb{P}(C_{t+1} = i | y_{1:t})$$

As the probability distribution of y_{t+1} is one of the key factors in deciding whether C_{t+1} is a change point, we can express this probability conditioned on previous values as

$$\omega_{t+1}^{(j)} = \mathbb{P}(y_{t+1} | C_{t+1} = j, y_{1:t})$$

$\omega_{t+1}^{(j)}$ can be thought of as how unlikely a data point is compared to its immediate predecessors, and so a more extreme data point is more likely to be a change point. Finally, we can substitute the transition probabilities above to give us

$$\mathbb{P}(C_{t+1} = j | y_{1:t+1}) = \begin{cases} \omega_{t+1}^{(j)} \frac{1-G(t-i)}{1-G(t-i-1)} \mathbb{P}(C_t = j | y_{1:t}) & \text{if } j < t \\ \omega_{t+1}^{(j)} \frac{G(t-i)-G(t-i-1)}{1-G(t-i-1)} \mathbb{P}(C_t = j | y_{1:t}) & \text{if } j = t. \end{cases}$$

Therefore, we can declare a change point at time $s \geq t$ if

$$\sum_{i=s-k+1}^{s+k} \mathbb{P}(C_{t+1} = j | y_{1:t+1}) > p$$

for p a probability threshold that we can choose.

On each new observation the method can be recursively updated. The incremental weights $\omega_{t+1}^{(j)}$ can be calculated efficiently as these depend on a set of summary statistics than can be stored and updated each time a change point is detected, which means that the computational cost of calculating ω is fixed. Since G does not change between change points, this can be calculated a single time at the start of the algorithm, and then sequentially as the distance from the previous change point grows; this allows a single fixed calculation for G, however increases the storage and if there no change points are detected in the sequence could double the space requirements.

2.6 Comparison of methods

The Fearnhead and Liu (2007) method is somewhat different from PELT and BISEG in that it does not seek to minimise a cost function, but sequentially looks at the probability of the next value given the previous values. As a result, it can struggle to find small changes in mean which are very likely but is much more confident with larger deviations, which is the same for PELT (Wambui *et al.*, 2015). CUSUM is much better than the other methods mentioned at finding of smaller deviations from the mean, however it is not an exact method and so may only find these points well after they have truly happened, as BISEG can use CUSUM as its test statistic, it has the option of also looking for smaller changes in an offline setting.

Figure 3 shows the proportion of time changes of various sizes are detected. All changepoint methods reliably detect abrupt changes in mean of 5σ or greater, as seen at $t = 10$, and are fairly reliable at changes of 4σ. We see that no method picks up changes of a single standard deviation at $t = 40$, though BISEG has the potential for this, minimising across that change must have less impact than the penalty, which demonstrates that a penalty function must be carefully chosen to find the desired types of changes. The differences in performance between FL and the other methods are most felt at moderate changes in mean of 2-3σ ($t = 50, 70$), where BISEG and PELT are quite reliable but FL struggle to produce consistent results. We can also see differences in performance with the rate of false positives, where FL has a persistently noisy bottom line, a low but constant rate of false positives. For smaller changes of 2-3σ, PELT and BISEG have a slight error margin (wider peak), whereas FL sometimes has a very slight lag in detecting the change point($t = 11, t = 21$).

Time complexity is an important consideration as the length of the data grows. PELT prunes Jackson *et al.* (2005)'s $O(n^2)$ time complexity into $O(n)$ under the assumption that the intervals between change points are drawn independently from a probability distribution, without this assumption it is $O(n^2)$. FL also assumes change points are drawn independently from a probability distribution, and has a computational complexity of $O(n)$. CUSUM also has a computational complexity of $O(n)$ as both it and FL make a single update to a test statistic and each of n time steps. BISEG has computational complexity $O(n \log(n))$ due to its divide and conquer approach.

Generally, methods make the assumption that the data has consistent Gaussian noise. Fearnhead Liu and CUSUM assume that this noise has fixed variance and attempt to detect only changes in mean, however BISEG and Optimal Partitioning perform well on both changes in mean and in variance (Rohrbeck, 2013).

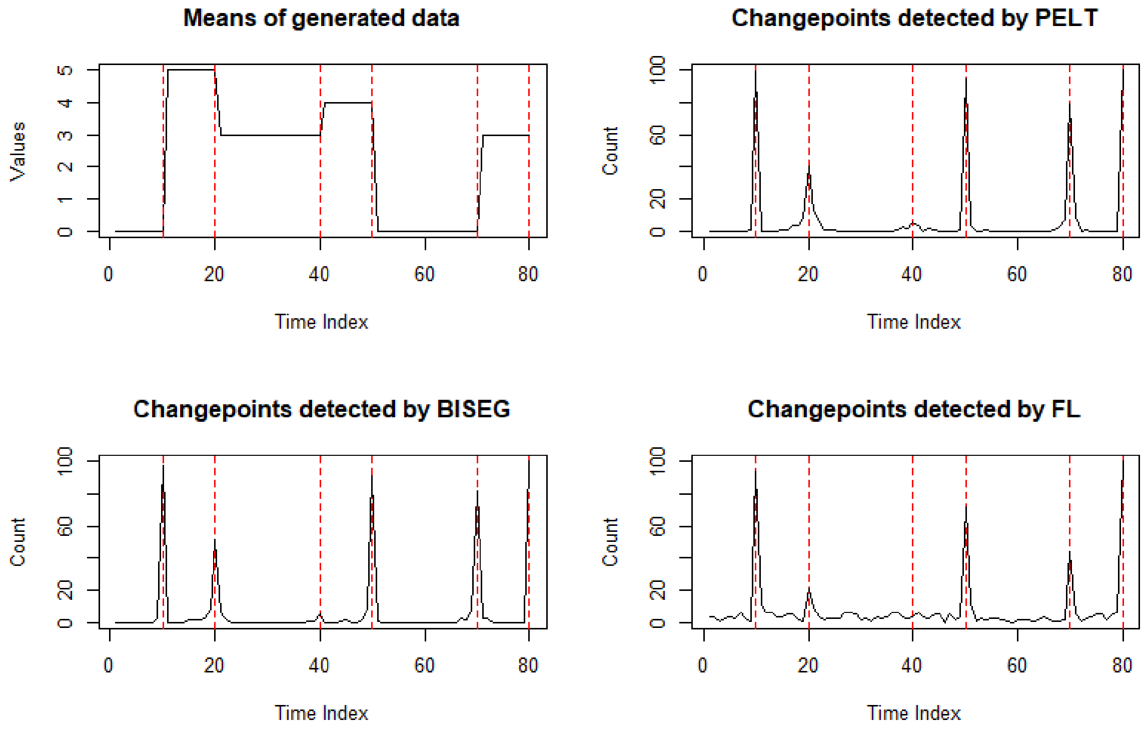

Figure 3: An example of the effectiveness of methods at finding change points of various sizes of intervals. Data generated with random noise of σ= 1 around means (top left), methods were run 100 times. Dashed red lines are change points. We can see that PELT (top right) and BISEG (bottom left) picked up change points approximately the same proportions of the time. All methods reliably spot the change at $t = 10$ of 5σ, with a drop in rate predicted for the change of 2σ at $t = 20$. No method has significant success at detecting a σ change in mean at $t = 40$. All methods have some false positives, however FL(bottom right) has a persistent presence.

Though methods may make an assumption of Gaussian noise, outliers can still be present, and cleaning data by removing outliers is not always possible, particularly in an online setting, therefore a methods response to outliers can impact its accuracy and usefulness. For PELT and BiSeg the choice of loss function dictates the method's response, using a Mean Squared Error will excessively weight outliers, whereas using bi-weight loss (Section 2.1) limits the impact of outliers to a constant and provides better outcomes in the face out outlier prone data Fearnhead and Rigaill (2019). Fearnhead-Liu and CUSUM on the other hand, do not have any way of dealing with outliers due to their unmodified consideration of the probability of a value as part of the decision on whether to call it a change point.

3 Adaptations of Fearnhead Liu

Since there is no publicly available coded version of Fearnhead Liu a major goal of this maths project was to implement and compare the method's performance to other methods. The implementation produced is available at (github.com/ZRova/OnlineFinancialChangepointDetection). It has not been optimised for space and time efficiency in R.

Beyond this, the Fearnhead Liu method requires the variance and distribution of change points be known and constant between change points, in this section we discuss how these can both be estimated in an online way, as well as other areas in which FL could be improved. With these modifications, FL can be made non-parametric, though with a higher error rate as a result of greater estimation of parameters.

3.1 Mean

While the mean is not a required parameter, Fearnhead and Liu (2007)'s use of the previous data point as the expected value of the mean leads to a high false positive rate, and there is potential to decrease this by providing a more robust estimate of the true mean.

Estimation of the mean can be done in a number of ways: in the fixed mean context where data $(y_i)_{i=1}^n \sim \mathcal{N}(\mu, \sigma^2)$ a simple mean of data points since the last change point can be taken, this will converge to the true value of the mean by the law of large numbers. Since Fearnhead Liu is only intended for use in a fixed mean context, this is the method we will use. If a change point is missed, there may be some value in having a Moving Average (MA) estimate of the mean. This involves taking the last x values in the data, and using their mean as the estimate. If there is a false negative, using the moving average will mean that the previous irrelevant mean will no longer be used, however for the rest of the time, estimates of the mean will be less reliable.

An Exponentially Weighted Moving Average (EWMA) could also be used to model a moving mean by giving greater weight to more recent values which may allow FL to step outside its normal comfort zone of changes in constant mean, for example in wind turbine sensors where the mean will be heavily influenced by the wind and not constant (de Rodriguez, 2020).

3.2 Variance

The variance of a series $Y = (y_1, ..., y_n)$ can be calculated directly using $Var(Y) = \frac{1}{n}\sum_{i=1}^n (y_i - \mu)^2$

Another way of estimating variance is using the Median Absolute Deviation (MAD), $MAD = Median(|(y_1 - \mu, ..., y_n - \mu)|)$, this can then be adjusted into an estimate for σ by multiplying by a factor k, which depends on the distribution of the data. In the case of Gaussian noise, $k = 1/\phi^{-1}(3/4) \approx 1.4826$. The MAD is much more robust to outliers than the traditional calculation of σ as it picks the least extreme values of the deviation, as opposed to squaring the most extreme. In the FL method which already struggles to find smaller changes in mean, increasing the variance would increase the rate of false negatives so we have chosen to look at the MAD in this report.

3.3 Distribution of Change Points

Since the method assumes that the change points are distributed by a Markov chain, it is assumed that the distribution of change points is constant for the duration of the data. This means that if the initial distribution is incorrectly provided, or if there is a change in distribution of change points, the method will be working on incorrect assumptions and likely perform worse, increasing its

FPR and/or decreasing its TPR. If we maintain the assumption that change points are distributed geometrically, we can update the Bayesian posterior distribution each time a new change point is detected. This is a Beta distribution with parameters (a, b) where a is the number of updates and b is the mean of the distribution.

Using the method of moments, we estimate the parameters of the distribution as below, where $\bar{x}$ is the mean of the distribution of change points, and $\bar{v}$ is the variance of the distribution, estimated as explained in the sections above.

$$\hat{a} = \bar{x}(\frac{\bar{x}(1-\bar{x})}{\bar{v}} - 1), \hat{b} = (1-\bar{x})(\frac{\bar{x}(1-\bar{x})}{\bar{v}} - 1)$$

For this estimate to be valid, the distribution of the change points needed to be over the interval [0,1] so the data was first normalised using the minimum and maximum interval. In addition, the mean was used as the point around which to scale the resulting probability distribution by aligning the mean with the centre of the cumulative distribution. It should be noted that it will take several change points to be detected before there is enough information about this distribution for it to be estimated, so this addition is best used on longer series.

Alternatively, a MCMC analysis can be done to find $g(.)$, by calculating the density function from previous change points. In order to keep this current through periods of low volatility (infrequent change points) and high volatility (frequent change points) recent values could be weighted more heavily. Using RJMCMC has been shown to be effective in multiple change point detection by Tan *et al.* (2021). In this project, we chose to use the posterior distribution and update it each time as its recursive properties and single means it is generally faster than MCMC.

3.4 Other Adapations

Figure 3 demonstrates FL's struggle to detect changes of less than 2σ and even 3σ, a potential way of overcoming it is by using the CUSUM test statistic using the estimated values. For larger jumps, the two should agree, however for the smaller jumps that only CUSUM can detect, the exactness of the unmodified FL method may be compromised as CUSUM is not able to find subtle changes immediately. As the CUSUM method takes O(1) time to update at each step, this adaptation will not increase the computation complexity. However, using CUSUM may increase the already high rate of false positives, but has the potential to increase the rate of true positives in data streams with a mixture of both small and large changes.

Additionally, estimating the variance can bring in a significant source of error if a change point is missed and the variance grows to encompass all data. To overcome this, it is worth considering a modification to the test statistic to penalise excessively large estimations of variance where a lower one might prove a better fit. In order to not prune the variance when the time point is simply close to the estimated mean, a cumulative metric similar to CUSUM could be kept track of to ensure that the variance remains low for a period of time. This would also expand the method to be able to detect decreases in variance specifically, as increases often appear as unlikely values and can be detected without modification.

As estimating variables can take a some time to converge on true values, a minimum interval could be imposed between change points in order to allow the estimates some time to converge and stabilise. A value of approximately 20 was found to produce best results (Uppal *et al.*, 2021), however even using the last 5 data points will reduce the effect of extreme values and could decrease the rate of false positives.

4 Evaluation Metrics

Some change point methods may seek to find only the approximate location of change points, or may simply be slightly off. When evaluating the methods, it is therefore common to allow a margin of acceptable points on either side of the true change point. If a method detects two change points within the given margin of a true change point, one is recorded as a true positive and the others are left as false positives (Killick *et al.*, 2012).

Variation of Information between sets $X, Y \subset A$ satisfies

$$VI(X;Y) = H(X) + H(Y) - 2I(X,Y)$$

where $I(X) = -log(P(x))$ is the information of X given probability P, and $I(X,Y) = I(X) + I(Y)$ is the mutual information between discrete random variables X and Y with respect to the uniform probability measure on A. $H(X) = E[I(X)]$ is the entropy of X. We want to minimise the Variation of Information between two sets, however in cases where there is great disparity between the size of the sets as common in series containing change points, the variation can be relatively minimal, and it can be difficult to understand what is causing it.

The Jaccard Index (or overlap) for two sets A, A' in [1,N] is $J(A, A') = |\frac{A \cap A'}{A \cup A'}|$ We can also look at the Covering Metric discussed by Arbeláez *et al.* (2011), which compares two partitions G, G'

$$C(G, G') = \frac{1}{N} \sum_{A \in G} |A| \cdot \max_{A' \in G'} J(A, A')$$

The Covering Metric finds the maximum overlap between each segment in G and each segment in G' and normalises to represent the largest proportion of G that is un-separated by segments of G'. This means that it is much more heavily affected by additional segments which will decrease the maximum value by decreasing the maximum union.

Accuracy is calculated as the number of correctly classified data points over the total number of data points. While accuracy provides a general idea of the algorithm's performance, it cannot tell you about whether the errors are due to false positives or false negatives as it gives both types of error equal weight. In order to get more information about specific error types you can use recall, R (also known as True Positivity Rate, TPR), defined as the proportion of correctly detected change points out of all true change points, and precision, P, the proportion of correctly detected change points out of all detected change points.

Precision and recall can be combined into an F_β score,

$$F_\beta = \frac{(1+\beta^2)PR}{\beta^2 P + R}$$

when $\beta = 1$, this corresponds to the F_1 score and weights the two equally; β can be used to adjust the relative weights of precision and recall (Van Rijsbergen, 1979). F-scores have a maximum value of 1, and the larger the score, the better the performance of the method. The F1 score is more considerate of the number of correctly detected change points that the cover metric.

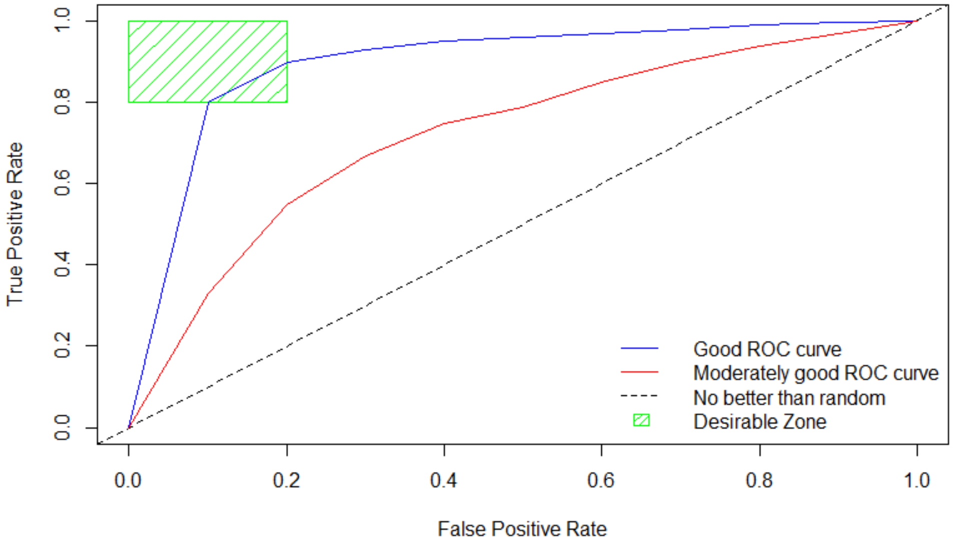

Figure 4: An example ROC curve, with a good method (blue line), moderately effective method (red line), and a no better than random method (black dashed). The good method is closest to the top left corner, and makes it into the "desirable zone" at a FPR<0.2, TPR>0.8

Reciever Operating Characteristic (ROC) curves plot the recall(TPR) on the y-axis against the False Positive Rate (FPR) on the x axis. Multiple TPR and FPR are calculated for the same change point detection method for a range of penalties or thresholds. This allows direct comparison of performance between algorithms at various thresholds, and also allows the most suitable threshold to be picked. A good ROC curve will get as close to the upper left corner ((0,1) co-ordinate) as possible, where there is a high rate of true positives and a low rate of false positives. One way of measuring this is by looking at the area under the ROC curve, the AUC where an AUC closer to 1 is preferred. A perfectly random classifier would plot the $x = y$ line on a ROC graph and yield and AUC of 1/2, so any classifier below this line can be inverted for better performance and a larger AUC.

5 Simulation Study

Using the study framework inspired by Van den Burg and Williams (2020), we can look at each of these adaptations of FL and compare them to PELT and BISEG on artificially generated data with fixed variance and random changes in mean at randomly distributed change points.

The data used for these tests was generated by first generating a series of random interval values, and a set of random mean values, then merging these into a time series with changes in mean and Gaussian noise.

In Figure 5 we can see the impacts that slight adaptations had on the Fearnhead Liu method with no adaptations (red line). All methods plotted perform fairly well and are significantly above the

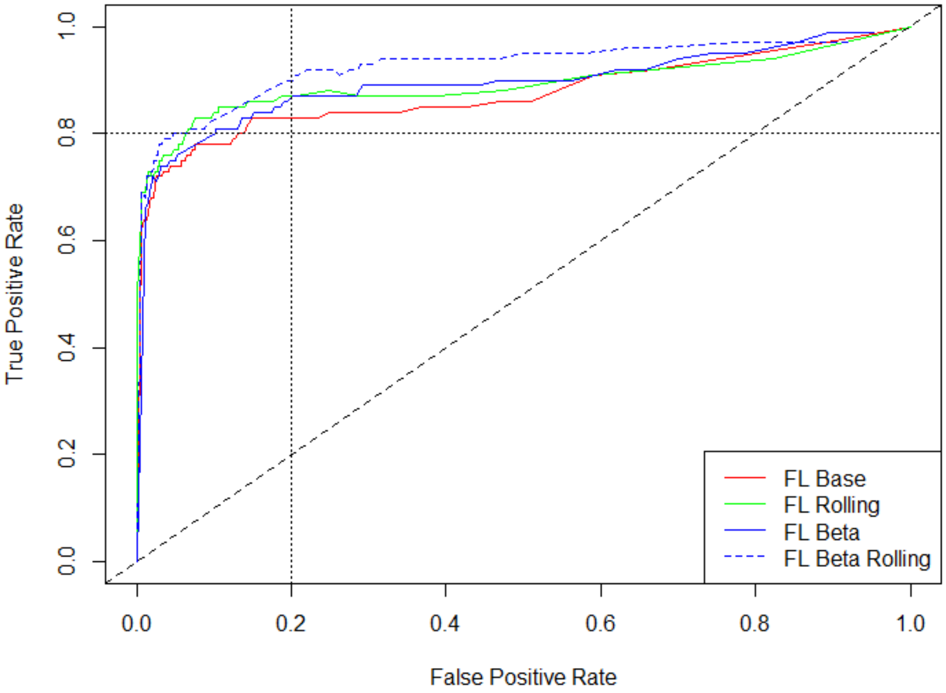

Figure 5: ROC curve of red: Fearnhead Liu, green: FL with rolling mean, blue: FL with beta estimation of changepoint intervals, blue dashed: FL with rolling mean and beta estimation. We can see that the curve performing the best (blue dashed) is using both rolling mean and beta estimation.

trivial line at X=Y (dashed black), and make it into the desirable zone indicated by the black dashed lines FPR < 0.2, TPR > 0.8 . We can see that the adaptations with the greatest impact on effectiveness were the rolling mean (green line), or using both the rolling mean and beta distribution (dashed blue) to represent the change point distribution. It is interesting that the two methods seem to be very similar at some thresholds and not at others, likely areas where there is not enough data to estimate the parameters of the beta distribution.

The ROC curves were calculated at thresholds from 0 to 1 in increments of 0.005 and demonstrate the importance of a good choice in threshold and the differences in true and false positive rate that this gives us. In order to compare these using the cover metric and F1 score, thresholds were picked that would maximise the mean of these results, between 0.98 and 0.99, which are represented at the very steep initial gradient in the ROC curve in Figure 5. These thresholds were very high as the cover metric heavily penalises false positives that break up large segments.

In Table 1, we see that between variations a better estimation of the mean provided the most significant single improvement to the method, followed by a better estimation of both mean and change point distribution. This makes sense as false positives were one of the base method's initial downsides, and a more reliable estimation of the mean would be the best thing that can be done to reduce them. In the oracle case, we see that there is no significant difference between using an estimation of the mean and beta distribution. This is likely because as the distribution of change points given is already true, the estimation cannot be improved significantly.

	Default		Oracle	
	Cover	F1	Cover	F1
FL Base	0.647	0.633	0.675	0.547
FL Mean	**0.780**	0.69	**0.783**	**0.687**
FL Beta	0.674	0.670	0.699	0.563
FL Beta & Mean	0.674	**0.723**	**0.776**	**0.694**
PELT	0.914	0.870		
BISEG	0.060	0.113	0.902	0.881

Table 1: A table of cover metrics and F1 scores for proposed variations on Fearnhead Liu and comparison to methods PELT and BISEG, with the best performing metric shown in bold. Default uses default parameters of the method, Oracle receives true distributions. We see that Mean and Beta & Mean adaptations of FL perform best, and that PELT and BISEG outperform all FL adaptations

From Table 1 we can see that all Fearnhead Liu methods and variants are outperformed by both PELT and BISEG. From Figure 3 we see that PELT and BISEG are able to reliably detect changes in mean of 2-3σ which is where FL struggles and is likely to be a source of the error. Between the default and oracle case, BISEG sees massive improvement with given how many parameters it should be looking for, however the improvements in Fearnhead Liu seem negligible in the rolling mean adaptation, and the F1 score appears to worsen for all adaptations, potentially due to an overall lower number of changepoints detected.

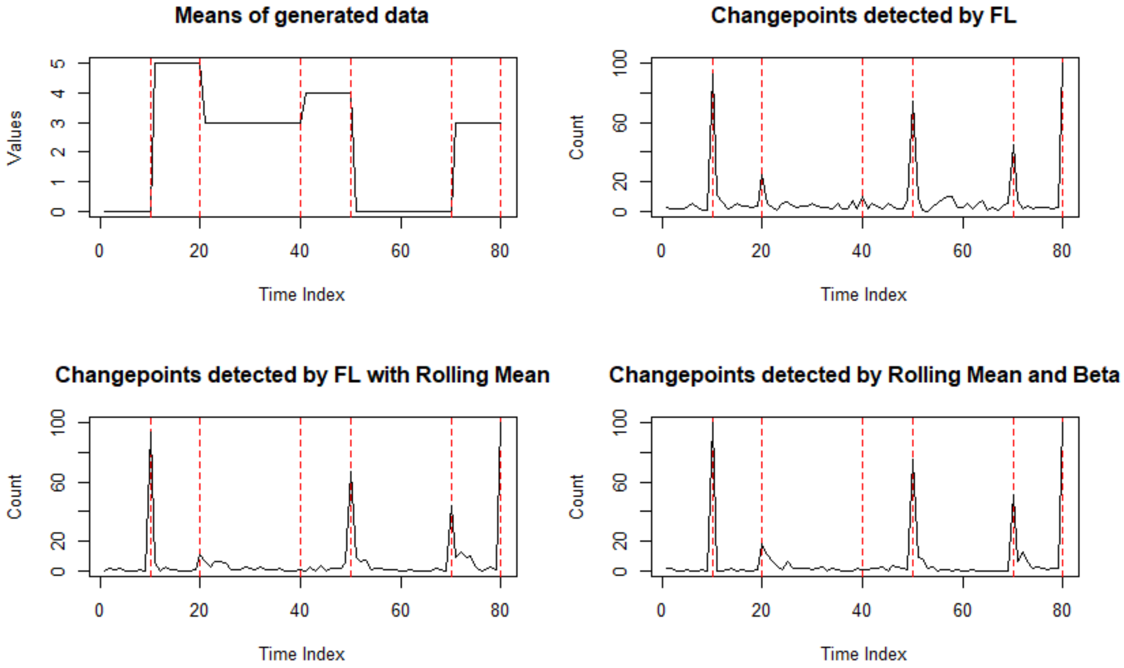

Figure 6: Figure demonstrates rate at which change points are detected over 100 data series using the means in top right graph with random Gaussian noise σ=1. Change points are red dashed lines. We see that the modifications in the bottom two graphs lead to fewer false positives, and while peaks do not seem higher than in unmodified FL (top right), detected change points have a tail where they are picked up with slight delay.

We can see more directly the impacts of the modifications on the method in Figure 6. We see that using a rolling mean in the bottom figures has reduced the rate of false positives, as there is less noise in areas where there is no change point. While the peaks indicating the initial rate of detection do not seem any higher with a rolling mean or rolling mean and beta, the peaks have a "tail" in the following time points where the same change point is able to be detected with delay due to a better estimation of the previous mean. There is initially no significant difference between just the rolling mean and rolling mean and beta variants, as the beta distribution takes several change points in order to have enough information for an estimation of change point distribution, however we see higher peaks relative to the rolling mean alone (bottom left) at $t = 50$ and $t = 70$ in Figure 6 (bottom right). None of the variants have any success detecting the small σ change at $t = 40$.

6 Conclusion

In this report, we have defined a change point and looked at a series of change point methods, including CUSUM, PELT and BISEG. We have then compared all of these methods to the method of Fearnhead and Liu (2007). We have proposed several adaptations of the method in order to estimate the mean, variance, and distribution of change points, and make the method non-parametric.

We then implemented the proposed adaptations, and compared them to find which one had the greatest effectiveness. We found that estimating both the mean and distribution of change points had the greatest impact on predictions, closely followed by only estimating the mean. In addition, we compared the original and improved method to some of the methods discussed earlier, and found that the Fearnhead Liu method did not outperform PELT or BISEG.

This report has worked to improve the Fearnhead Liu change point detection method, and has proposed some additional improvements, inspired by recent contributions to literature to improve existing change point detection methods. From the initial conclusions, while Fearnhead Liu has a much lower run time, its current effectiveness indicates that further improvements in estimation will still under-perform relative to PELT and BISEG without a change in the test statistic. Other unresolved problems include the method's ability to only deal with uni-variate data, and it's strict assumptions about the distributions of change points which limit it's ability to investigate novel data.

Recent literature has continued to improve existing methods with a growing emphasis on computational complexity and run time (Wendelberger *et al.*, 2021; Watkins *et al.*, 2022). With Fearnhead Liu a relatively computationally efficient online method, there is further potential to improve and develop this method.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control* **19**(6), 716–723.

Aminikhanghahi, S. and Cook, D. J. (2017). A survey of methods for time series change point detection. *Knowledge and information systems* **51**(2), 339–367.

Arbeláez, P., Maire, M., Fowlkes, C. and Malik, J. (2011). Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **33**(5), 898–916.

Bardwell, L. (2018). *Efficient search methods for high dimensional time-series*. Lancaster University (United Kingdom).

Bosc, M., Heitz, F., Armspach, J.-P., Namer, I., Gounot, D. and Rumbach, L. (2003). Automatic change detection in multimodal serial mri: application to multiple sclerosis lesion evolution. *NeuroImage* **20**(2), 643–656.

Ducré-Robitaille, J.-F., Vincent, L. A. and Boulet, G. (2003). Comparison of techniques for detection of discontinuities in temperature series. *International Journal of Climatology* **23**(9), 1087–1101.

Fearnhead, P. and Liu, Z. (2007). On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society Series B* **69**(4), 589 – 605.

Fearnhead, P. and Rigaill, G. (2019). Changepoint detection in the presence of outliers. *Journal of the American Statistical Association* **114**(525), 169–183.

Frisén, M. (2009). Optimal sequential surveillance for finance, public health, and other areas. *Sequential Analysis* **28**(3), 310–337.

Hinkley, D. V. and Hinkley, E. A. (1970). Inference about the change-point in a sequence of binomial variables. *Biometrika* **57**, 1–17.

Jackson, B., Scargle, J. D., Barnes, D., Arabhi, S., Alt, A., Gioumousis, P., Gwin, E., Sangtrakulcharoen, P., Tan, L. and Tsai, T. T. (2005). An algorithm for optimal partitioning of data on an interval. *IEEE Signal Processing Letters* **12**(2), 105–108.

Killick, R., Fearnhead, P. and Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association* **107**(500), 1590–1598.

Koshti, V. (2011). Cumulative sum control chart. *International journal of physics and mathematical sciences* **ISSN**, 2277–2111.

Lavielle, M. (2005). Using penalized contrasts for the change-point problem. *Signal processing* **85**(8), 1501–1510.

Maidstone, R., Hocking, T., Rigaill, G. and Fearnhead, P. (2017). On optimal multiple changepoint algorithms for large data. *Statistics and computing* **27**(2), 519–533.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika* **41**(1/2), 100–115.

de Rodriguez, K. K. (2020). *Detecting Change Points in Wind Turbine Sensors Using Exponential Smoothing*. Ph.D. thesis, Utica College.

Rohrbeck, C. (2013). Detection of changes in variance using binary segmentation and optimal partitioning.

Rybach, D., Gollan, C., Schluter, R. and Ney, H. (2009). Audio segmentation for speech recognition using segment features. In: *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 4197–4200.

Sagiroglu, S. and Sinanc, D. (2013). Big data: A review. In: *2013 international conference on collaboration technologies and systems (CTS)*, IEEE, 42–47.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics* , 461–464.

Stoica, P. and Selen, Y. (2004). Model-order selection: a review of information criterion rules. *IEEE Signal Processing Magazine* **21**(4), 36–47.

Tan, C., Hu, J. and Wu, Y. (2021). Detection of multiple change-points in the scale parameter of a gamma distributed sequence based on reversible jump mcmc. *Journal of the Korean Statistical Society* **50**(1), 25–43.

Uppal, R., Nagaraj, S., van Leer, E. and Anderson, D. V. (2021). Non-parametric online change-point detection algorithm. In: *2021 IEEE Statistical Signal Processing Workshop (SSP)*, 396–400.

Van den Burg, G. J. J. and Williams, C. K. I. (2020). An evaluation of change point detection algorithms. *arXiv preprint arXiv:2003.06222* .

Van Rijsbergen, C. (1979). Information retrieval: theory and practice. In: *Proceedings of the Joint IBM/University of Newcastle upon Tyne Seminar on Data Base Systems*, volume 79.

Wambui, G. D., Waititu, G. A. and Wanjoya, A. (2015). The power of the pruned exact linear time (pelt) test in multiple changepoint detection. *American Journal of Theoretical and Applied Statistics* **4**(6), 581.

Watkins, J., Carlson, M., Shan, K., Tezaur, I., Perego, M., Bertagna, L., Kao, C., Hoffman, M. J. and Price, S. F. (2022). Performance portable ice-sheet modeling with mali. *arXiv preprint arXiv:2204.04321* .

Wendelberger, L. J., Gray, J. M., Reich, B. J. and Wilson, A. G. (2021). Monitoring deforestation using multivariate bayesian online changepoint detection with outliers. *arXiv preprint arXiv:2112.12899* .

Zhou, Y., Fu, L. and Zhang, B. (2017). Two non parametric methods for change-point detection in distribution. *Communications in Statistics-Theory and Methods* **46**(6), 2801–2815.