Lancaster University

STOR601: Research Topic II

Nonparametric Methods for Online Changepoint Detection

Author:
Paul Sharkey

Supervisor:
Rebecca Killick

May 18, 2014

Contents

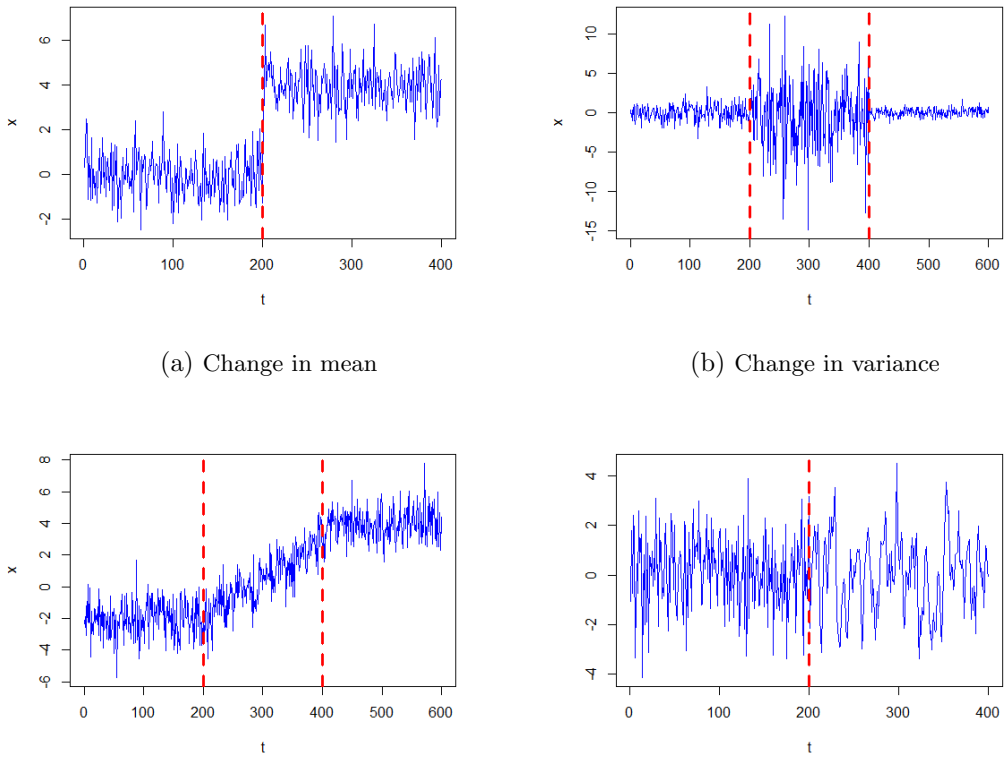
Abstract

Changepoints have been extensively analysed in order to identify structural changes in time series data, typically when the data are of known parametric form. This report presents an exploration of methods to detect changepoints in a nonparametric setting, where no assumptions are made with regard to the distributional structure of the data, yet must still maintain a specified level of performance. In particular, the framework of a two-sample hypothesis testing procedure for sequential testing is developed, in which test statistics based on ranks of observations are adapted for changepoint detection. This framework is then extended to consider multiple changepoints and data streams. The characteristics and performance of the testing procedure are analysed by comparing the impact of test statistics in a range of scenarios. Under this framework, it is found that while parametric techniques tend to outperform nonparametric techniques in a Gaussian setting, nonparametric tests are a suitable alternative. In addition, it is found that tests for arbitrary distributional changes are comparable to tests designed to detect changes in location and scale. Overall, the nonparametric hypothesis testing procedure is found to perform well, and represents a logical course of action when performing changepoint analysis on data of no known distributional form, a common scenario that applies to a wide variety of real-world processes.

1 Introduction

The relevance of so many physical, economic and industrial processes in our daily lives leads to the inevitable question of what happens when these systems change. In particular, interest lies in the time at which these processes undergo a change, usually in order to mitigate the consequences that this change entails. It is essential, therefore, to develop methods for change detection based on a rigorous statistical framework. The point in a time series when the statistical properties of an underlying process change is known as a *changepoint*. In practice, these changes manifest in a shift in mean and variance, though more arbitrary distributional changes are known to occur (Ross and Adams, 2012) (see Figure 1).

Changepoint literature is focused primarily in the *offline* setting, where inference regarding the detection of a change occurs retrospectively, after the data has been received. In contrast, the *online* setting features methods in which analysis is performed sequentially - as every new observation is received, the detection method is implemented in order to locate possible changepoints in previous observations. Changepoint detection methods can be further categorised into the *parametric* class, which incorporates distributional knowledge of the data into the detection scheme, and *nonparametric* class, which makes no such distributional assumptions regarding the data. This report will focus entirely on nonparametric approaches for changepoint detection. An overview of parametric offline techniques can be found in Eckley et al. (2011).



(a) Change in mean

(b) Change in variance

(c) Change in regression

(d) Change in dependence

Figure 1: This figure shows some common occurrences of changepoints in time series.

Changepoint analysis is utilised in a wide variety of applications. In genomics, changepoint methods are used to identify tumour progression and type from regions where DNA copy numbers are amplified or reduced (Zhang and Siegmund, 2007). Such methods are also useful in detecting changes in volatility in financial markets (Allen et al., 2013). The origins of online changepoint analysis are in the field of quality control in industrial processes (see Section 3.2). However, before further discussion of changepoints can take place, the changepoint model must be rigorously defined.

1.1 Changepoint Model

Consider a sequence of time series data $x_{1:n} = (x_1, x_2, \ldots, x_n)$. Each observation at time t, x_t, is assumed to be univariate, though extensions to multivariate data are possible. The data are assumed to have a number of changepoints m, whose positions are denoted $\tau_1, \tau_2, \ldots, \tau_m$. The changepoints are assumed to be ordered so that $\tau_i < \tau_j$ if and only if $i < j$. The m changepoints partition the data into $(m + 1)$ segments. The ith segment consists of data $y_{\tau_{i-1}+1:\tau_i}$. Assuming the data in each segment is independently and identically distributed (i.i.d), the distribution of the

sequence can be written as:

$$X_i \sim \begin{cases} F_0 & \text{if } i \leq \tau_1; \\ F_1 & \text{if } \tau_1 < i \leq \tau_2; \\ F_2 & \text{if } \tau_2 < i \leq \tau_3; \\ \dots & \\ F_n & \text{if } i > \tau_n. \end{cases}$$

Traditional approaches to the problem assume some knowledge of the distributional form F_i before and after the changepoint, with only the parameters being unknown. However, since many real-world processes do not exhibit well-defined behavour, alternative methods are required that do not make such restrictive assumptions (Hawkins and Deng, 2010). In the sequential change detection setting, in particular, incorrect assumptions regarding the distributional form of the data can have large effects on the rate of false positives (Ross et al., 2011). Hence, there is a need to develop distribution-free approaches to changepoint detection that can maintain a specified level of performance.

Single changepoint detection in a time series is analogous to a two-sample test, whereby a formal hypothesis testing procedure is used to identify differences in the statistical properties of the two samples. The majority of the changepoint literature is focused on the application to the fixed-sample setting. Early works include Hawkins (1977) and Pettitt (1979). The application of this testing procedure to the sequential setting was first formulated in Hawkins et al. (2003) and Hawkins and Zamba (2005), which proposed the use of the Student t-test and F test for detecting changes in the location and scale of Gaussian data respectively. In Section 2, the extension of this approach to the nonparametric setting (Hawkins and Deng, 2010) is introduced in order to counter this problem when the distributional form of the data is unknown. For simplicity, the test statistics will be reviewed in the context of the offline setting before being extended to sequential changepoint analysis. The extension made in Ross et al. (2011) to the streaming problem is also discussed, where the procedure must be adjusted to account for memory constraints.

This report is structured as follows. Firstly, the use of a two-sample hypothesis testing procedure for changepoint detection in the offline setting is discussed in Section 2. Section 3 extends this procedure to the online setting, which includes a discussion of multiple changepoint approaches and the computational issues that arise. The results of a simulation study designed to assess the characteristics of this testing procedure are presented in Section 4. Section 5 describes some open problems and alternative approaches in the literature, before concluding the review.

2 Offline setting

The problem of detecting a **single** changepoint in a fixed data setting can be reduced to testing the hypothesis:

$$H_0 : X_i \sim F_0 \ \forall i, \qquad H_A : X_i \sim \begin{cases} F_0 & \text{if } i < \tau; \\ F_1 & \text{if } i \geq \tau. \end{cases}$$

where n is the length of the sequence and $\tau < n$. In the two-sample testing procedure, this requires the calculation of a test statistic $D_{\tau,n}$. However, since the position of τ is almost always unknown, an alternative approach is necessary. A method outlined in Pettitt (1979) is to evaluate $D_{\tau,n}$ for all values of $1 < \tau < n$ and use the maximum value. The test statistic to be used in this procedure is thus defined as:

$$D_n = \left| \max_{\tau} \frac{D_{\tau,n} - \mu_{D_{\tau,n}}}{\sigma_{D_{\tau,n}}} \right|, \qquad 1 < \tau < n.$$

The test statistic is standardised by its mean and standard deviation in order to avoid using a maximum value that may be skewed by values that give a high variance. The absolute value of the statistic is also taken to enable two-sided change detection to take place. In this way, the test procedure will account for both increases and decreases in the mean and variance of the data. The null hypothesis that no change occurred is rejected if $D_n > h_n$ for some threshold h_n. The manner in which h_n is determined is outlined in Section 2.2. Finally, the estimate of the changepoint location $\hat{\tau}$ is given by the value of τ that maximises $D_{\tau,n}$:

$$\hat{\tau} = \arg\max_{\tau} D_{\tau,n}.$$

The following section defines explicit formulas for test statistics that detect changes in mean, variance and more general properties of the data.

2.1 Test statistics

In this section, a range of test statistics are introduced that can be incorporated into the testing procedure defined in the previous section. These test statistics make no restrictive assumptions regarding the distributional structure of the data. In Section 4, the characteristics and performance of these test statistics are compared alongside some parametric test statistics in a number of scenarios.

When little information is available regarding the statistical structure of the data before and after the changepoint, a common approach is to use the Kolgormorov-Smirnov (KS) test for arbitrary changes in distribution (Ross and Adams, 2012). However, if the type of change is known in advance, such as a change in mean and variance, more powerful tests are available that take into account the type of change being analysed (Hawkins and Deng, 2010). Many nonparametric test statistics are based upon the ranks of the observations, where the rank of the ith observation at

time t is defined as:

$$r(x_i) = \sum_{i \neq j}^{t} I(x_i \geq x_j),$$

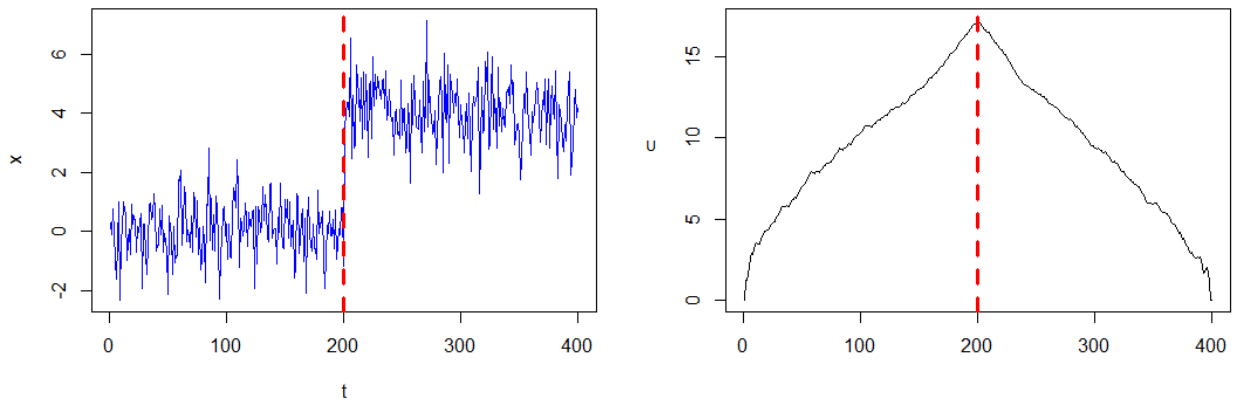
where $I(\cdot)$ is the indicator function.

2.1.1 Mann-Whitney test statistic

The majority of work in the changepoint literature focuses on the shift in the location parameter (Hawkins and Zamba, 2005). Pettitt (1979) proposed a test statistic for detecting changes in mean based on the *Mann-Whitney* two-sample test. Let S and T denote the samples pre-change and post-change respectively. Assuming no tied ranks, then the expected rank of each point is $(n+1)/2$ under the null hypothesis that no change has occurred and that the pooled sample is identically distributed. Therefore, the test statistic is defined as:

$$U_{\tau,n} = \sum_{x_i \in S} (r(x_i) - (n+1)/2),$$

which is a measure of how the ranks of the observations deviate from their expected rank. This quantity is computed for all values $1 < \tau < n$ and the null hypothesis is rejected if the maximum value exceeds some threshold h_n (see Figure 2).



(a) (b)

Figure 2: (a) shows a time series with an obvious location shift at $\tau = 200$, while (b) shows the $U_{\tau,n}$ statistic calculated for all τ. The estimate of the location of the changepoint, that is, the value of τ that maximised $U_{\tau,n}$, coincides with the actual location of the changepoint.

2.1.2 Mood test statistic

The *Mood* statistic introduced in Mood et al. (1954) is used to test for a change in scale between two samples. Like the Mann-Whitney test, the Mood test assesses the extent at which the ranks of the observations deviate from their expected value:

$$M_{\tau,n} = \sum_{x_i \in S} (r(x_i) - (n+1)/2)^2.$$

In this case, however, the squared deviations are used as these are more likely to detect a change in scale. $M_{\tau,n}$ can be used in the context of the hypothesis testing procedure defined in Section 2. It is calculated for all values $1 < \tau < n$ and the maximum value is compared to some threshold h_n. If the null hypothesis is rejected, then a change in scale is said to have occurred at the value of τ that maximises the standardised test statistic.

2.1.3 Kolmogorov-Smirnov and Cramer-von-Mises test statistics

Two omnibus tests are also introduced that are sensitive to more general distributional changes not limited to location and scale shifts. The *Kolmogorov-Smirnov (KS)* and *Cramer-von-Mises (CvM)* tests assess such changes by comparing the empirical distribution functions of the pre-change and post-change samples, which are defined as:

$$\hat{F}_S(x) = \frac{1}{\tau} \sum_{i=1}^{\tau} I(X_i \leq x);$$

$$\hat{F}_T(x) = \frac{1}{n-\tau} \sum_{i=\tau+1}^{n} I(X_i \leq x).$$

The KS test statistic is defined as the maximum difference between these empirical distributions:

$$D_{\tau,n} = \sup_x |\hat{F}_S(x) - \hat{F}_T(x)|.$$

The CvM test uses a measure that is based on the square of the mean distance between the empirical distributions:

$$W_{\tau,n} = \int_{-\infty}^{\infty} |\hat{F}_S - \hat{F}_T|^2 dF_t(x),$$

where $F_t(x)$ is the empirical CDF of the pooled sample. $W_{\tau,n}$ can be computed directly as:

$$W_{\tau,n} = \sum_{i=1}^{n} |\hat{F}_S(X_i) - \hat{F}_T(X_i)|^2.$$

Both test statistics can be used in the hypothesis testing procedure defined in Section 2 (Ross and Adams, 2012). While the procedure for the CvM statistic is straightforward, there are issues with the standardisation of the KS statistic. This is due to the fact that there are no closed-form expressions for the mean and variance of $D_{\tau,n}$, except asymptotically when n is large. Because

asymptotic expressions are usually not compatible with finite-length sequences, especially when small sample behaviour is important, an alternative approach must be considered. Instead of the $D_{\tau,n}$ statistic, the associated p-value $p_{\tau,n}$, defined as the probability of observing a more extreme value than $D_{\tau,n}$, is considered. Defining $q_{\tau,n} = 1 - p_{\tau,n}$ and:

$$q_n = \max_{\tau} q_{\tau,n}$$

then a change is detected if $q_n > h_n$ for some threshold h_n and the formal hypothesis testing procedure can be applied as with the CvM statistic. The procedure for determining the $p_{\tau,n}$ is outlined in Ross and Adams (2012).

2.2 Threshold selection

As with any formal hypothesis testing procedure, a threshold h_n is chosen such that, if the given test statistic exceeds this threshold, then the null hypothesis is rejected. This threshold is typically chosen to bound the rate of false positives α, that is, the probability of detecting a change when no change has occurred. This indicates that h_n should therefore be chosen as the upper α quantile of the distribution of D_n under the null hypothesis.

However, while the asymptotic distributions of some test statistics are tractable, the distribution of D_n generally does not have an analytic, finite-sample form. Because these asymptotic distributions are less accurate when using finite-length sequences of data, numerical simulation is required to estimate the distribution, and in turn, the threshold h_n. The remainder of this report introduces and compares test statistics in the online setting. Thus, the reader is referred to Hawkins and Deng (2010) and Ross and Adams (2012) for more information on the calculations required for the offline setting. As is discussed in Section 3.3, the procedure required for the calculation of thresholds takes an added complexity in the online setting.

3 Online setting

The desire to formulate methods for detecting changepoints in the online setting emerged from the belief that a change should be flagged as soon as possible in order to deal effectively with the consequences of such a change. Early work on this problem featured in the field of quality control (Page, 1954). Online changepoint analysis is used in the form of control charts to monitor output of industrial processes and to identify when the process has gone "out of control". Recent research has focused on extending the changepoint detection procedure outlined in Section 2 to the online setting (Hawkins and Deng, 2010). The need for nonparametric methods in sequential testing is vast as the assumption of a known distributional form is usually violated, particularly as the data is processed in real-time. With high-frequency data streams, these assumptions rarely hold (Ross

et al., 2011).

3.1 Performance measures

The performance of online change detection algorithms is typically measured using two criteria (Basseville et al., 1993). The *Average Run Length*, ARL_0 is defined as the average number of observations before a false positive occurs. The *Mean Detection Delay*, ARL_1 is defined as the mean delay until a change is detected. These quantities can be expressed formally as:

$$\text{ARL}_0 = \mathbb{E}(\hat{\tau}|F = F_0);$$
$$\text{ARL}_1 = \mathbb{E}(\hat{\tau} - \tau|F = F_1).$$

A false positive is said to have occurred if $\hat{\tau} < \tau$. In general, an acceptable value of ARL_0 is chosen before attempting to minimise the detection delay. This is analogous to minimising the Type II error subject to a bound on the Type I error in offline hypothesis testing. There exists an intuitive interplay between these two criteria. For example, if the value of ARL_0 is increased, that is, if the probability of a false alarm decreases, then the detection delay will also increase as a result of collecting more information to ensure that a false positive has not occurred. Researchers use these two criteria as the standard measure of performance in online detection schemes, partly because the ARL_0 is well-defined. Section 5 discusses the issues behind using these performance criteria. Section 4 uses the aforementioned criteria as a measure of the performance of these detection schemes in order to compare competing methods.

3.2 Control charts

Online changepoint analysis stems from the beginnings of statistical process control, in which industrial processes are monitored in order to detect faults. Page (1954) introduced a cumulative sum (CUSUM) control chart designed to signal a sustained change in the process mean μ_1. The CUSUM chart is defined by:

$$S_0 = 0;$$
$$S_i = \max\{0, S_{i-1} + X_i - k\},$$

where the reference value $k = (\mu_1 + \mu_2)/2$. A shift in the mean is signaled if $S_i > h$, where h is a threshold designed to set the rate of false positives (equivalently, the quantity ARL_0) at some acceptable level. If such a shift is signaled, then the estimate of the changepoint is given by the most recent value of j before the shift such that $S_j = 0$. The CUSUM chart procedure is powerful for detecting small shifts. However, it requires knowledge of the in-control and out-of-control means to calculate the reference value as well as for calculating the decision interval h. For a comprehensive

overview of other control chart schemes such as EWMA and Stewhart, see Lucas and Saccucci (1990).

3.3 Test statistics and Threshold selection

The formal hypothesis testing procedure in the two-sample case outlined in Section 2 can be readily extended to the online setting (Hawkins and Deng, 2010; Zhou et al., 2009) in a manner that is related to the control chart scheme. As a new observation x_t is received, the procedure is performed on a new dataset containing the new observation and the preceding observations. In this way, for all t, the dataset $x_{1:t} = (x_1, \ldots, x_t)$ is treated as a fixed-length sequence. The test statistic $D_{\tau,t}$ is calculated for every value of τ, standardised and the maximum value chosen. As before, the test statistic D_t is compared with a threshold h_t and a change is flagged if the threshold is exceeded, much like the control chart procedure outlined in Section 3.2. Since this occurs for all t, this represents a repeated sequence of hypothesis tests. With any multiple hypothesis testing procedure, a Bonferroni correction may be applied. However, this may be too conservative since the test statistics are correlated.

This procedure requires a sequence of time-varying thresholds h_t. It is common practice for changepoint models to have a fixed ARL_0 value, which results in the false alarm probability α being equal at every point. Like in Section 2.2, the upper α quantile of the distribution of D_t is desired, but this has no obvious, analytical form. In addition, there is an added condition on the false alarm probability, which is difficult to compute:

$$\begin{aligned} \mathbb{P}(D_1 > h_1) &= \alpha, \\ \mathbb{P}(D_t > h_t | D_{t-1} \leq h_{t-1}, \ldots, D_1 \leq h_1) &= \alpha, \quad t > 1. \end{aligned}$$

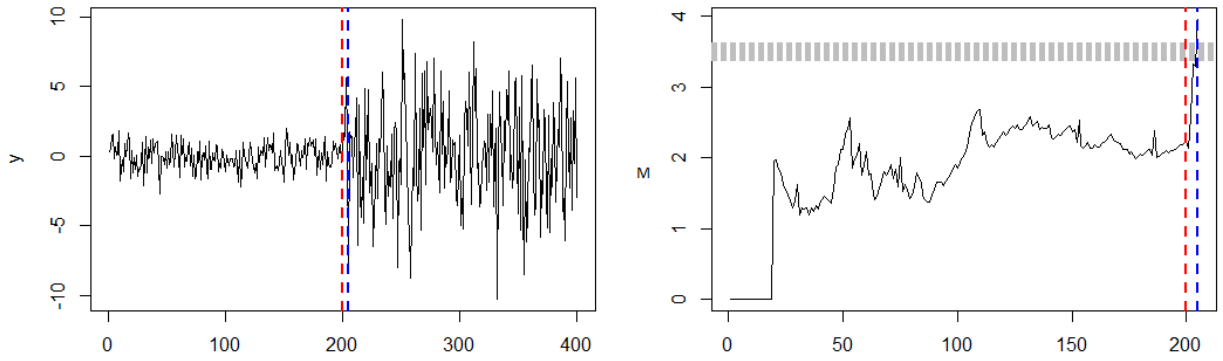
Furthermore, using an asymptotic distribution would skew the ARL_0 of the change detector, since it will not be accurate in the early monitoring period. Because of these issues, Monte Carlo simulation is used, as in the offline setting, to determine the sequence of thresholds. This can be stored externally and will not contribute to additional computational overhead in the procedure. Ross et al. (2011) advocates to generate this sequence as follows:

1. One million streams containing 5000 points are generated, without any changepoints.

2. The changepoint model is evaluated over the simulated streams, and D_n is calculated at each time instance.

3. The distribution of D_n can therefore be approximated, and the required values of the thresholds can be determined.

While it is possible to begin testing from the third observation, the discrete nature of the distribution of D_t is unsuitable for short sequences as it is difficult to have control limits that correspond to a desired value of ARL_0 that would be associated with a low false-alarm. Due to the low power of these hypothesis tests for small sample sizes, Hawkins and Deng (2010) recommended an initialisation period before the testing procedure begins (see Figure 3).



(a) (b)

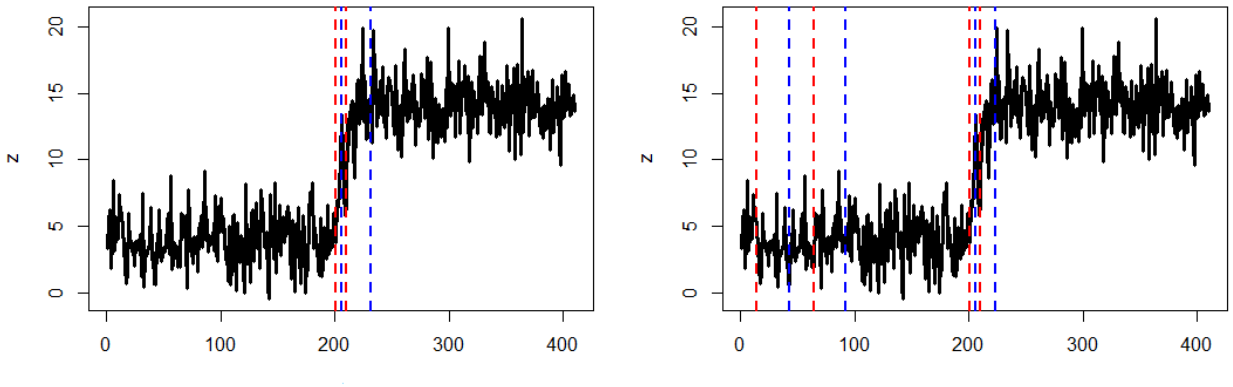
Figure 3: (a) shows a time series with an obvious scale shift at $\tau = 200$, while (b) shows a control chart of the Mood statistic calculated sequentially. The grey line denotes the time-varying threshold h_t based on the value $\text{ARL}_0 = 1000$. The blue dotted line represents the time that the change is flagged, while the red dotted line represents the time of the changepoint. In this example, there is a detection delay of 7 time units. Note that the statistic is calculated only after an initialisation period of 20 observations.

3.4 Multiple changepoint detection

In quality control, the single changepoint detection scheme is sufficient in the monitoring of output of industrial processes. Once a changepoint has been detected, and the process deemed "out of control", the process is halted and the fault fixed, after which the monitoring restarts. However, there are many applications in which a sequence of observations has multiple changepoints. As a result, the change detection mechanism must be able to identify these sequentially. In the context of the hypothesis testing procedure already outlined, this involves discarding all previous observations once a single changepoint $\hat{\tau}_1$ has been estimated. The change detector is then reset to process observations beginning with the $(\hat{\tau}_1 + 1)$th data point. For example, if a changepoint is detected at $\tau = 100$, the first 100 data points are discarded, and the process restarts beginning with the 101st

observation. This procedure is repeated until all the observations have been processed.

While this approach is simple to implement, there are also a number of drawbacks. Firstly, if a sequence of data features many different types of changepoints, then the procedure is not valid for test statistics that focus on a specific type of change (Mann-Whitney, Mood, etc.). As a result, tests for arbitrary distributional changes are more suitable, but these are less powerful for detecting specific changes. Secondly, in data featuring frequent, abrupt changes, the changepoint detection scheme may fail to detect changes efficiently due to the recommended initialisation period for the procedure. If the initialisation period is reduced, more false positives may arise. In contrast, if the period is increased, then the detection delay will also increase (see Figure 4).



(a) (b)

Figure 4: This shows two identical time series with changepoints at $t = 200$ and $t = 210$. (a) has an initialisation period of 30 observations, and as a result, has quite a large delay in the detection of the second changepoint. (b) has an initialisation period of 5 observations and gives two false positives, but a lower detection delay.

3.5 Computational issues

The two-sample hypothesis testing procedure in its current formulation is unsuitable in the streaming problem as the computational overhead of evaluating D_t increases linearly with t. In addition, as the number of observations increase, the number of possible changepoint locations k increases, as does the number of hypothesis tests required. Ross et al (2011) extends the two-sample framework by discretising older observations to satisfy the computational and complexity requirements for the processing of data streams.

An interval is defined in which most observations from the stream lie. This interval is partitioned into m segments $s_1, s_2, \ldots, s_m$ in which a count c_i is maintained of the number of observations that fall in segment s_i. A window of fixed length w is defined, in which the w most recent observations are stored into memory. This discretisation satisfies the condition of constant memory as both the m count variables and the number of observations in the window remains constant over time. The rank of an observation in the window is now computed as the sum of the rank against the other points in the window, and its rank against all previous points in the stream, as approximated by the discretisation. Each point in the jth segment s_j is assigned the value of the mid-point $v_j = (s_j + s_{j-1})/2$. The rank of each point x_t in the window is then:

$$r(x_t) = r_w(x_t) + \sum_{i=1}^{m+1} c_i I(x_t > v_j) - 1,$$

where r_w is the rank of x_t among the observations in the window. While discretisation is successful in extending the changepoint model to the setting of data streams, there are some drawbacks. There is a small loss of accuracy in the computation of ranks. The variability in the test statistics will also be increased, which may negatively impact the results of the hypothesis testing procedure. It is desirable, therefore, to have m as large as possible to strengthen the approximation made by this discretisation method. However, a large m also increases the memory required.

4 Simulation study

A small simulation study is designed in order to assess the performance of the changepoint methods introduced in Section 3. In particular, the Mann-Whitney and Mood statistics are investigated in comparison to parametric detection procedures for detecting location and scale shifts respectively. The parametric tests considered are the Student t-test for changes in mean and the Bartlett test for changes in variance. For both types of change, these given methods are compared with the Kolmogorov-Smirnov and Cramer-von-Mises tests for general distributional changes. The parametric tests are expected to outperform the nonparametric tests simulated from a Gaussian distribution, so the tests are also compared using non-Gaussian data sequences. These analyses are carried out in R using the '`cpm`' package (Ross, 2013). 100 replications are made and the results are summarised over these. The value of ARL_0, the average run length, is fixed at 500 for all tests. Two datasets are designed for the purpose of the study, one simulated purely from Gaussian data and another using a combination of log-normal and Student t-random variables. Both consist of a combination of small and large shifts in location and scale, as well as a mixture of small and large pre-change data sequences (see Tables 1 and 2 and Figure 5).

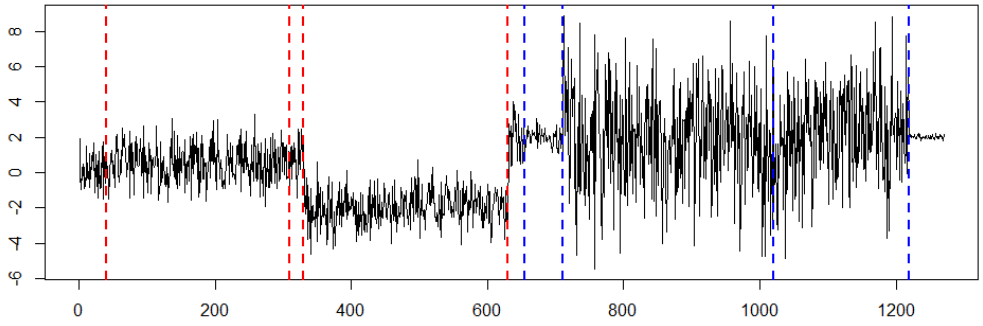
Structure of simulated Gaussian data			
Data sequence	Length	Mean	Variance
1	40	0	1
2	270	0.5	1
3	20	1	1
4	300	-2	1
5	25	2	1
6	55	2	0.5
7	310	2	3
8	200	2	2.5
9	50	2	0.1

Table 1: This shows the distributional structure of the simulated Gaussian data. 8 changepoints separate these nine individual data sequences.

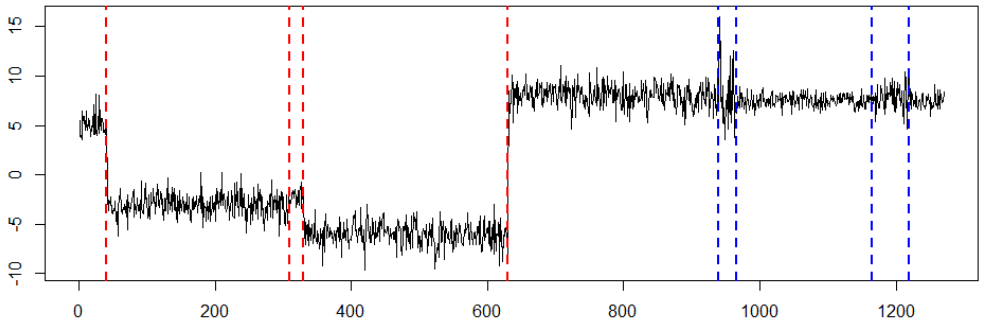
Structure of simulated non-Gaussian data				
Data sequence	Length	Distribution	Mean	Variance
1	40	Log-normal(0.571,0.495)	2	1.111
2	270	Student t(-3,20)	-3	1.111
3	20	Student t(-2,20)	-2	1.111
4	300	Student t(-6,20)	-6	1.111
5	310	Student t(8,20)	8	1.111
6	25	Log-normal(2.015,0.5)	8.499	20.518
7	200	Log-normal(2.015,0.08)	7.525	0.364
8	55	Student t(8,40)	8	1.053
9	50	Log-normal(2.015,0.08)	7.525	0.364

Table 2: This shows the distributional structure of the simulated non-Gaussian data. 8 changepoints separate these nine individual data sequences.

For the purposes of this study, a changepoint is considered detected if the changepoint estimate is within 2 time units of the true changepoint. Tables 3 and 4 summarise the success of each detection scheme in the simulation. Note that tests that are designed to detect location shifts (like the Mann-Whitney test) are not used in the detection of changes in scale, and vice-versa. As expected, the Student t-test mostly outperforms the nonparametric schemes for detecting changes in mean in Gaussian data. However, the nonparametric tests reach a relative success $80 - 95\%$ in all cases. A



(a)



(b)

Figure 5: This shows a single realisation of simulated Gaussian (a) and non-Gaussian (b) data with changepoints in location (denoted in red) and scale (denoted in blue).

point of note is the fact that the Kolmogorov-Smirnov test detects changepoints 2-4 more successfully than the Mann-Whitney test, which is theoretically more powerful for changes in mean. For changepoint 2, this could be due to the ability of the KS test to detect a change in skewness, which is an underlying feature of this changepoint. Similarly, the Mood test performs well in comparison to the Bartlett parametric test, even achieving a success rate of 86% in detecting changepoint 6. The Mood test performs consistently better than the arbitrary change detectors in this case.

Similar to the Gaussian case, the Kolmogorov-Smirnov and Cramer-von-Mises tests have a comparable performance to the Mann-Whitney test in the non-Gaussian setting. Interestingly, however, there are two instances in which the Student t-test detects two changepoints 95% of the time in the

Changepoint detection success (Gaussian data) (%)								
Test	1	2	3	4	5	6	7	8
Mann-Whitney	21	8	95	95	-	-	-	-
Mood	-	-	-	-	35	86	6	98
Student-T	24	10	98	100	-	-	-	-
Bartlett	-	-	-	-	43	57	6	100
Kolmogorov-Smirnov	18	16	97	99	1	51	3	73
Cramer-von-Mises	18	8	99	97	2	40	2	90

Table 3: This shows the percentage success of each test in the detection of the eight simulated changepoints in Gaussian data over 100 replications.

Changepoint detection success (Non-Gaussian data) (%)								
Test	1	2	3	4	5	6	7	8
Mann-Whitney	97	58	95	98	-	-	-	-
Mood	-	-	-	-	80	83	51	39
Student-T	N/A	54	100	N/A	-	-	-	-
Bartlett	-	-	-	-	62	89	45	57
Kolmogorov-Smirnov	100	49	99	99	43	53	42	34
Cramer-von-Mises	99	57	96	98	29	53	32	33

Table 4: This shows the percentage success of each test in the detection of the eight simulated changepoints in non-Gaussian Gaussian data over 100 replications.

region where only one exists. These results are therefore invalid, and highlight the unsuitability of parametric tests when considering data that are not normally distributed. In contrast, there are some cases where the parametric Bartlett test outperforms the nonparametric Mood test. As for the Gaussian data, the Mood statistic outperforms the arbitrary detection schemes in each case.

The detection schemes are also compared with respect to the mean detection delay (ARL_1). Table 5 presents these figures for the Gaussian setting. It is observed that for location shifts, delay is minimised by using the Student t-test as expected. However, this delay appears inflated when detecting the first changepoint. This might be due to the small sample size for the pre-change distribution. The Kolmogorov-Smirnov and Cramer-von-Mises tests are again comparable to the Mann-Whitney test. Due to the reduced success in detecting scale shifts for some changepoints, not enough data is available to give a confident summary of mean detection delay. While delays

Mean detection delay (Gaussian data)								
Test	1	2	3	4	5	6	7	8
Mann-Whitney	49.053	11.125	5.568	4.8	-	-	-	-
Mood	-	-	-	-	23.142	4.931	70.5	10.622
Student-T	25.417	9.7	2.969	3.38	-	-	-	-
Bartlett	-	-	-	-	16.58	9.105	70.6	4.13
Kolmogorov-Smirnov	14.277	10.125	6.762	6.07	20 (N/A)	15.118	24.667 (N/A)	17.164
Cramer-von-Mises	33.77	11.25	5.313	10.78	28 (N/A)	18	13 (N/A)	16.311

Table 5: This shows the mean detection delay for each changepoint corresponding to each test for Gaussian data. This was obtained by calculating the difference between the changepoint estimate and the time of detection and averaging over replications.

Mean detection delay (Non-Gaussian data)								
Test	1	2	3	4	5	6	7	8
Mann-Whitney	5.588	9.069	5.916	5.051	-	-	-	-
Mood	-	-	-	-	4.675	13.145	6.216	28.128
Student-T	N/A	8.6	3.9	N/A	-	-	-	-
Bartlett	-	-	-	-	5.968	6.418	9.489	14.474
Kolmogorov-Smirnov	5.83	7.454	7.596	3.444	11.07	35.924	10.262	25.853
Cramer-von-Mises	4.545	9.316	5.302	4.541	10.345	40.188	13.063	23.485

Table 6: This shows the mean detection delay for each changepoint corresponding to each test for non-Gaussian data. This was obtained by calculating the difference between the changepoint estimate and the time of detection and averaging over replications.

again appear inflated, the Bartett test generally minimises the detection delay, as expected.

The methods are also compared with respect to the mean detection delay in the non-Gaussian data sequence (see Table 6). For reasons discussed previously, the results of the Student t-test are omitted due to concerns regarding its validity when analysing non-Gaussian data. Again, the tests for arbitrary changes are comparable to the Mann-Whitney test, while the Bartlett and Mood tests appear to outperform the KS and CvM tests consistently.

The results of the study can be summarised as follows:

- When small shifts occur in the data, the mean detection delay is lower in the case when

the changepoint is preceded by a large sequence rather than a small sequence of pre-change data. This is clearly observed in changepoints 7 and 8 in the non-Gaussian data sequence. In contrast, for large changes, the size of the pre-change sequence of data seems irrelevant. This is observed in changepoints 3 and 4 of the Gaussian data sequence. An important factor to consider with regard to small pre-change sequences is the effect of the warm-up period on the detection mechanism. The cpm package is inbuilt with a default warm-up period, and it is unclear if this has an adverse effect on detection times, which may explain several instances in the study when the detection delays appear unusually large.

- Parametric tests generally outperform their nonparametric counterparts when detecting changepoints in the Gaussian sequence of data, as expected. However, in many cases, the difference is not excessive, meaning nonparametric tests can be used for such an analysis. In particular, the Mood test performs well compared to the Bartlett test for changes in scale.

- Despite achieving favourable results in some cases in the non-Gaussian setting, the Student t-test demonstrated its unsuitability for non-Gaussian data sequences by over-detecting two changepoints. Ross et al. (2011) discusses how parametric tests fail to meet the target value of ARL_0 in this situation. This stems from the fact that model thresholds in the nonparametric case are distribution-free, while in the parametric models, these thresholds are computed under the assumption of normality. Because the precise control of the false alarm rate is essential in change detection problems, parametric tests based on normality may not be suitable in data sequences that are not normally distributed.

- In Hawkins and Deng (2010) and Ross et al. (2011), a phenomenon is seen where nonparametric tests outperform their parametric counterparts for small shifts in Gaussian data. This is explained by the fact that Gaussian thresholds are slightly higher at the extremes, allowing nonparametric models to react more quickly to small changes. This phenomenon was only observed when detecting changepoint 6 in the Gaussian data sequence. Further investigation is required.

5 Conclusion

5.1 Summary

Because so many real-world processes and applications exhibit behaviour that is not well-defined, it is often difficult to justify the use of parametric models in order to detect changes. Hence, there is a need for a statistically rigorous changepoint model that does not operate on such restrictive assumptions. The two-sample hypothesis testing procedure introduced in this report facilitates this demand. This report details how this procedure can be adapted to detect multiple changepoints

sequentially, as well as to cater for high-frequency data in the presence of memory constraints. The characteristics and specialties of a range of nonparametric test statistics are introduced and compared in terms of their usefulness and performance. Nonparametric tests for arbitrary distributional changes are found to be comparable to the tests designed specifically to detect changes in location and scale. It is found that while test statistics based on the assumption of normality outperform their nonparametric counterparts in a Gaussian setting, the nonparametric testing procedure is a worthy alternative. In contrast, parametric tests are unsuitable in changepoint analysis of non-Gaussian data. Hence, the nonparametric test is the preferred method when analysing data arising from a real-world process, as the assumption of normality is too restrictive, particularly in sequential testing. The report will conclude with a discussion on some open problems in the field and alternative approaches to nonparametric online changepoint detection.

5.2 Open problems and current research

While the hypothesis testing procedure outlined in this report has proven effective for nonparametric changepoint detection in the sequential setting, there are a number of limitations associated with this model. Concerns have been raised regarding the use of ARL_0 and ARL_1 as performance measures. Mei (2008) questioned the validity of these criteria in certain scenarios, based on the fact that finite detection delay may be achieved even with infinite ARL_0, despite the result by Lorden et al. (1971) that indicates a finite detection delay will lead to the probability of ever raising a false alarm being 1. The specification of ARL_0 is also entirely subjective. The user may wish to specify a large value of ARL_0, which will reduce the number of false positives but will increase the detection delay. It is therefore at the discretion of the user whether the number of false positives or the size of delay is the priority of interest. This is, of course, application-dependent.

While control charts exist in the multivariate changepoint setting, the complexities involved mean that no software package has been developed to implement this testing procedure in a simple manner. The 'ecp' package (Matteson and James, 2014) performs nonparametric changepoint analysis on multivariate time series using the E-divisive algorithm, which bisects a time series into clusters at the point of change. However, this method formulation is restricted to the offline setting.

The two-sample hypothesis testing procedure is unsuitable for dependent data. For a detailed overview of nonparametric methods for detecting changepoints in dependent data, see Dehling et al. (2013).

The model is further limited in the respect that it has not been extended to network data, which consists of time series located at fixed nodes on a graph. This is a relatively recent area of re-

search, and while the methods outlined in this report have not been implemented with regard to this problem, techniques from Bayesian nonparametric modelling have been used. Research in this area is focused on the occurrence of a change throughout the network at a particular point in time, whether such a change propagates through the network over time, and whether multiple changepoints can be detected throughout the network over time. For a comprehensive summary of recent advances in this area, see Sharpnack et al. (2012) and Schmidt and Mørup (2013). Bayesian nonparametric modelling has already been used to infer the location of a changepoint and the pre- and post-change distributions in the univariate case (Muliere and Scarsini, 1985; Mira and Petrone, 1996). However, this approach has only been developed in the offline setting.

The vast majority of the sequential changepoint detection literature is focused on parametric models. There exist few alternatives to the hypothesis testing procedure outlined in this report. However, one such alternative is *direct density-ratio estimation* (Kawahara and Sugiyama, 2012). This involves computing the ratio of the pre- and post-change probability densities without the computational overhead of computing the probability densities themselves. The logarithm of this ratio is evaluated as the change detection score and compared with some threshold μ. This algorithm can be extended to the sequential setting using stochastic gradient descent, which facilitates learning of the model parameters efficiently in an online manner. Another such approach is the use of *singular-spectrum analysis* (Moskvina and Zhigljavsky, 2001). This technique is based on the idea that if the distance between the l-dimensional hyperplane spanned by the eigenvectors of the so-called lag-covariance matrix and the M-lagged vectors increases, then a change in the distributional structure of the time series has occurred. However, unlike direct density-ratio estimation, this approach has not been adapted to the sequential setting.

References

Allen, D. E., McAleer, M., Powell, R. J., and Singh, A. K. (2013). Nonparametric multiple change point analysis of the global financial crisis. Technical report, Tinbergen Institute Discussion Paper.

Basseville, M., Nikiforov, I. V., et al. (1993). *Detection of abrupt changes: theory and application*, volume 104. Prentice Hall Englewood Cliffs.

Dehling, H., Rooch, A., and Taqqu, M. S. (2013). Non-parametric change-point tests for long-range dependent data. *Scandinavian Journal of Statistics*, 40(1):153–173.

Eckley, I. A., Fearnhead, P., and Killick, R. (2011). Analysis of changepoint models. In *Bayesian Time Series Models*. Cambridge University Press.

Hawkins, D. M. (1977). Testing a sequence of observations for a shift in location. *Journal of the American Statistical Association*, 72(357):180–186.

Hawkins, D. M. and Deng, Q. (2010). A nonparametric change-point control chart. *Journal of Quality Technology*, 42(2):165–173.

Hawkins, D. M., Qiu, P., and Kang, C. W. (2003). The changepoint model for statistical process control. *Journal of quality technology*, 35(4):355–366.

Hawkins, D. M. and Zamba, K. (2005). Statistical process control for shifts in mean or variance using a changepoint formulation. *Technometrics*, 47(2).

Inclan, C. and Tiao, G. C. (1994). Use of cumulative sums of squares for retrospective detection of changes of variance. *Journal of the American Statistical Association*, 89(427):913–923.

Kawahara, Y. and Sugiyama, M. (2012). Sequential change-point detection based on direct density-ratio estimation. *Statistical Analysis and Data Mining*, 5(2):114–127.

Lorden, G. et al. (1971). Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, 42(6):1897–1908.

Lucas, J. M. and Saccucci, M. S. (1990). Exponentially weighted moving average control schemes: properties and enhancements. *Technometrics*, 32(1):1–12.

Matteson, D. S. and James, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, 109(505):334–345.

Mei, Y. (2008). Is average run length to false alarm always an informative criterion? *Sequential Analysis*, 27(4):354–376.

Mira, A. and Petrone, S. (1996). Bayesian hierarchical nonparametric inference for change-point problems. *Bayesian Statistics*, 5:693–703.

Mood, A. M. et al. (1954). On the asymptotic efficiency of certain nonparametric two-sample tests. *The Annals of Mathematical Statistics*, 25(3):514–522.

Moskvina, V. and Zhigljavsky, A. (2001). Application of the singular spectrum analysis for change-point detection in time series. *Journal of Time Series Analysis, submitted*.

Muliere, P. and Scarsini, M. (1985). Change-point problems: A bayesian nonparametric approach. *Aplikace matematiky*, 30(6):397–402.

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

Pettitt, A. (1979). A non-parametric approach to the change-point problem. *Applied statistics*, 28(2):126–135.

Ross, G. J. (2013). Parametric and nonparametric sequential change detection in r: The cpm package. *Journal of Statistical Software*.

Ross, G. J. and Adams, N. M. (2012). Two nonparametric control charts for detecting arbitrary distribution changes. *Journal of Quality Technology*, 44(2).

Ross, G. J., Tasoulis, D. K., and Adams, N. M. (2011). Nonparametric monitoring of data streams for changes in location and scale. *Technometrics*, 53(4):379–389.

Schmidt, M. N. and Mørup, M. (2013). Non-parametric bayesian modeling of complex networks. an introduction. *IEEE Signal Processing Magazine*, 30(3):110–128.

Sharpnack, J., Rinaldo, A., and Singh, A. (2012). Changepoint detection over graphs with the spectral scan statistic. *ArXiv e-prints*.

Zhang, N. R. and Siegmund, D. O. (2007). A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63(1):22–32.

Zhou, C., Zou, C., Zhang, Y., and Wang, Z. (2009). Nonparametric control chart based on change-point model. *Statistical Papers*, 50(1):13–28.