THESIS TITLE - CHANGE POINT DETECTION

Tyler Manning-Dahan

A thesis
in
The Department
of
Engineering and Computer Science

Presented in Partial Fulfillment of the Requirements
For the Degree of Master of Applied Science
Concordia University
Montréal, Québec, Canada

February 2020

Concordia University
School of Graduate Studies

This is to certify that the thesis prepared

By: **Tyler Manning-Dahan**

Entitled: **Thesis Title - Change Point Detection**

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining commitee:

_____ Chair

_____ Examiner

_____ Examiner

_____ Examiner

_____ Supervisor

Approved _____

Chair of Department or Graduate Program Director

_____ 20 _____ _____

Abstract

Thesis Title - Change Point Detection

Tyler Manning-Dahan

Text of abstract.

Acknowledgments

I would like to thank my supervisor Dr. Jia Yuan Yu for accepting me into his lab and giving me the opportunity to further develop my academic career with this Master's thesis. I have become a much better researcher thanks to him and I am very grateful for his mentorship.

Several fellow researchers have been very helpful in writing this thesis. Thank you especially to Thomas Flynn and Damien Garreau who took a lot time of out their schedule to answer my questions and inform particular directions of this thesis.

I would also like to thank all my colleagues at DRW who have helped me learn about the finance industry over the past three years. I am especially grateful to Yves, Neil and Laura, who have been very patient with me and have me taught me a lot about understanding financial markets, writing clean code and building robust statistical models.

Lastly, I would like to thank my fiancée, Tanya, who initially pushed me to go back to school while I was still young and had the opportunity.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

In the first half of the twentieth century, the use of statistical control charts for detecting real-time changes in variation was pioneered by Walter Shewart. Shewart was interested in reducing the unexpected causes of variation in the manufacturing processes that produced faulty manufacturing equipment [34]. Shewart's method involved charting the process measurements over time and detecting when a statistical process was no longer exhibiting an expected level of variation. Once this detection occurred, the process was stopped and was not restarted until the cause of the variation was fixed. Shewart's control charts were one of the first formal frameworks to solve the problem of detecting changes in a distribution of a sequence of random variables. This problem is now known more generally as the *change point detection problem*. Many industries make use of change-point techniques including healthcare monitoring systems, monitoring computer network traffic, and detecting regime shifts financial markets. The following are a few motivating examples.

1.1.1 Health Care

Health care is an important area for quickly detecting signal changes. Some recent studies include applications to heart rate monitoring [40] [36], epilepsy signal segmentation [25], and multi-modal MRI lesion detection [3] to name a few. Quickly detecting changes to a patient's health is absolutely necessary for any system to be of practical use. However, this quick detection must be balanced with high accuracy as

false positives or missed detections could have life-threatening consequences. Therefore, balancing type I and type II error is a central theme to online change-point detection.

1.1.2 Financial Applications

The application of accurate and timely change point detection is also very popular in the finance sector where shifts in asset prices can suddenly happen. Change point detection is particularly hard in financial applications because of the non-stationary data typically observed in asset price time series. Note, in the financial literature, change-points are also referred to as structural breaks, but for this thesis we will use the broader term change-points.

An on-line, quick detection technique is proposed in [30], where a modified Shiryaev - Roberts procedure is used in a case study to detect a change-point on a single stock's daily returns. They compare their non-parametric method with other classic control chart methods using speed of detection and false alarm rate as measures of performance.

Detecting changes in variance is specifically explored in [18]. The authors propose an off-line change-point algorithm that minimizes a global cost function by using an adaptive regularization function. The algorithm is applied to the absolute returns of the FTSE 100 stock index and the US dollar-Japanese Yen foreign intra-day exchange rate to detect changes in asset price volatility. The change-points identified in the FTSE 100 coincided with key market events such as the stock market crash that occurred on October 14^{th}, 1987 and breaking the 5000 price barrier in August 1997.

See section 1.3.6 of [37] for more applications to options markets and arbitrage opportunities.

1.2 Characteristics of the change point problem

A number of surveys of the literature already exist [1] [28], therefore we will not cover all existing methods but rather touch upon several, important factors to consider when tackling the change point detection problem. Across the body of literature, these factors determine what methods are available to practitioners.

The first factor is selecting between *parametric* and *non-parametric* techniques.

Deciding between these two broad techniques is dependent on the prior knowledge one wants to encode into the problem. For example, if it is known that data is generated by a distribution from the exponential family of distributions, then we can subset the problem from the space of all possible distributions to a smaller space of distributions. Shewart control charts and CUSUM change-point techniques are both parametric techniques based on the Gaussian-family of distributions [29] [6]. In other settings, it is not possible to leverage information about the data and non-parametric techniques must be used instead [4].

The second factor is deciding whether change-points should be detected *offline* or *online*. Some algorithms are off-line—also referred to as batch algorithms or retrospective or a posteriori change-point detection—and they are applied in an ex-post fashion after the dataset has been completely acquired [38]. If change-points must be detected as soon as possible, then waiting for the entire dataset to be acquired is not feasible and methods that operate on data streams must be used. Such methods fall into the category of on-line change-point detection. The aforementioned Shewart control chart and CUSUM algorithm are both designed for data that is streamed in a real-time fashion. In the literature, on-line methods of change point detection are also referred to sequential change point detection [37]. For this thesis, we will use the terms interchangeably.

The third factor is determining if there are multiple change points or to assume there is only a single change point to detect. This is an important factor for off-line change point detection where the decision to detect one or more change-points is often chosen at the outset. Detecting multiple change-points could also be relevant for the on-line case if a situation arises where the window of time series under consideration may contain more than one change point. However, most on-line change-point methods are designed to detect a single change-point at a time.

Finally, the last factor to address is determining exactly what kinds of statistical changes an algorithm should detect. Many methods focus solely on detecting changes in the mean of a distribution [20]. Some methods are more general and can detect changes in the variance or higher order moments and do not focus on any particular one. Methods like kernel change point detection can typically detect any distributional changes, making them attractive in situations where very little is known about the data.

This thesis will concern itself with on-line change point detection, where data is received in a streaming nature. We assume no prior distributional characteristics on the data and operate in a completely non-parametric setting.

1.3 Background

This sections describes how the change-point problem will be formulated in this thesis and, by extension, how all methods will be described using the change-point detection problem notation. Because online change-point detection is closely related to two sample testing, a background on statistical hypothesis testing is presented first.

1.3.1 Hypothesis Testing

Two-sample hypothesis testing concerns itself with the following problem: Suppose data is independently, sampled from two groups of different populations. Sample one is denoted as $X = \{X_1, X_2, ..., X_n\} \sim P$ and sample two is denoted as $Y = \{Y_1, Y_2, ..., Y_n\} \sim Q$. The fundamental question of a two-sample test is determining whether samples are X and Y are significantly different on a statistical level. Therefore, a two-sample hypothesis can be setup where the null hypothesis, H_0, is that the two distributions do not differ based on some statistical test. If the null hypothesis is rejected, then the alternative hypothesis, H_1, is accepted and we conclude the two samples differ in their distributions according to the specific two-sample test used.

$$\begin{cases} H_0 : P = Q & \text{(both samples come from the same distribution)} \\ H_1 : P \neq Q & \text{(samples do not come from the same distribution).} \end{cases} \tag{1}$$

In order to summarize the difference between the two samples, a test statistic, $\hat{t} \in \mathbb{R}$, is computed based on the type of test used. This test statistic is compared to a significance level, $\alpha \in [0, 1]$ that is chosen at the outset. Common choices for α are $0.1, 0.05$ and 0.01. In other words the p-value$= \mathbb{P}(T > \hat{t}|H_0$.

Given the two possible outcomes of a two-sample sample test, it is clear the hypothesis test can fail in two ways. The first is rejecting the null hypothesis when it is correct. This is known as a false-positive or a type-I error and is upper-bounded by the chosen significance level, α. Therefore the probability of falsely rejecting the null

hypothesis is $1 - \alpha = \mathbb{P}(\text{reject } H_0 | H_0 \text{ is true})$. The second possible source of error, is a type-II error or false negative. The probability of committing a type-II error is denoted as β. The quantity $1-\beta$ is referred to as the *power* of a test. It is equivalent to $\mathbb{P}(\text{ reject } H_0 | H_0 \text{ is false })$. Maximizing test power is an important part of designing new algorithms and this measure is typically used to compare different methods. Often there is a trade-off between type-I and type-II error and the practitioner must decide how to balance the two given their domain-specific knowledge of the problem. In some cases, it may be desirable to sacrifice one for the other, such as in medical field where a false positive diagnosis is a more desirable outcome than missing a diagnosis (type-II error) which would result in never giving treatment to a patient.

Many two sample tests exist for tackling the different types of differences two samples may have. For example, the Student t-test is a two sample test for determining if two samples of univariate data come from a population with the same mean. A generalization of the Student t-test for the multivariate case is the Hotelling T^2 test that tests whether the means of two multivariate samples are significantly different. Both of these are parametric tests as they assume the samples are normally distributed.

Non-parametric tests also exist such as the Kolmogorov-Smirnov test (K-S test) for determining whether two samples come from the same distribution. This is done by computing the *supremum* of the difference of the empirical cumulative distribution functions from each sample. That is, KS statistic $= \sup|P(x) - Q(x)|$. Depending on the number of samples and the significance level chosen, the K-S statistic may reject or fail to reject the null hypothesis. The K-S test does not specify what distribution the samples come from, only if they differ according to the K-S statistic. More recently, the kernel two-sample test was introduced in [10] as a very flexible, non-parametric test. It is not limited to one dimensional data, and can be applied to non-numeric data. It is based on the *maximum mean discrepancy* (MMD) criterion and is capable of detecting any kind of change in statistical moment. It is a focus in this thesis and is discussed in further detail in section 1.3.2.

1.3.2 Maximum Mean Discrepancy

The maximum mean discrepancy (MMD) is a type of *integral probability metric* that can be used to compare two probability distributions. It is based on the kernel mean embedding developed in [35]. The kernel mean embedding can be thought of

applying the *kernel trick* to probability measures and mapping them into a higher dimensional feature space. Rather than applying the kernel trick to individual data points and mapping them to an implicit feature space, the kernel mapping, ϕ, (also called feature map) will be applied to a probability distribution in order to represent in a reproducing kernel Hilbert space (RKHS), that is $\phi : \mathcal{X} \to \mathcal{H}$.

Suppose n samples from a set $X = \{x_1, x_2, ..., x_n\}$ and and m samples from a different set $Y = \{y_1, y_2, ..., y_m\}$ are observed from a sample space $\mathcal{X}$. They are distributed as $X \sim P$ and $Y \sim Q$ respectively. The kernel mean embeddings are represented by $\mu_P = \mathbb{E}_{X \sim P}[\phi(X)]$ and $\mu_Q = \mathbb{E}_{Y \sim Q}[\phi(Y)]$ for samples X and Y respectively.

It is not required but typically a *characteristic* kernel function is chosen as it guarantees the mapping into the Hilbert space to be injective. As [27] points out, there is no loss if information in this case. See [9] for this proof. This implies that $||\mu_P - \mu_Q|| = 0$ if and only if $P = Q$. This leads naturally to the definition of the MMD as:

$$\text{MMD}(P, Q) = ||\mathbb{E}_{X \sim P}[\phi(X)] - \mathbb{E}_{Y \sim Q}[\phi(Y)]||_{\mathcal{H}} \tag{2}$$

Where the MMD can be understood as the distance in $\mathcal{H}$ between the mean embeddings of the features. As long as a kernel mean embedding can be defined on the given data structure, the MMD can be used. This is why it can be applied to non-numeric data such as strings, graphs, and other data structures [15]. One advantage of using the MMD for comparing distributions over other classic techniques like Kullback–Leibler (K-L) divergence is the density of the distributions do not have to be estimated as an interim step.

The two-sample kernel test statistic is defined in [10] and uses the MMD as a distance measure for comparing two probability distributions. Where the null hypothesis is that both samples stem from the same distribution, (i.e. $P = Q$) and the alternative hypothesis is that they are not drawn from the same distribution such that $P \neq Q$.

In [10], the unbiased, estimate of the squared MMD is shown to be:

$$\widehat{\mathbf{MMD}}_u^2(\mathcal{F}, X, Y) = \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j=1}^{m} k\left(x_i, x_j\right) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k\left(x_i, y_j\right) + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} k\left(y_i, y_j\right) \tag{3}$$

Common kernel functions used in practice are the radial basis function kernel, $k(x, y) = e^{-\frac{1}{2\sigma^2}||x-y||^2}$ and the laplace kernel, $k(x, y) = e^{-\frac{1}{\sigma^2}||x-y||}$, where $\sigma > 0$. In [11], the authors recommend selecting the bandwidth kernel, σ, based on the *median heuristic*, which results in $\sigma^2 = \text{median}\{||x_i - x_j|| : i, j = 1, ..., n\}$. This heuristic is a good starting point, however, it has been shown to be sub-optimal in high-dimensional and small-sample cases as shown in [26] and [31]. Choosing the proper kernel is often done on a per case basis by maximizing test power, thereby reducing the chance of Type-II error. Overall kernel selection is a difficult problem that is still actively researched.

Originally in [10], it was thought the MMD does not suffer from the curse of dimensionaity when used to compare distributions in higher dimensions. However, it was shown in https://arxiv.org/pdf/1406.2083.pdf that indeed the MMD does struggle in higher dimensions like many other metrics do.

REWORD: We call the function that achieves the supremum, the witness function because it is the function that witnesses the difference in the two distributions. This means that we can interpret the witness function as showing where the estimated densities of p and q are most different.

The witness function

$$f(x) = \mathbb{E}_{x' \sim p}\left[k\left(x, x'\right)\right] - \mathbb{E}_{x' \sim q}\left[k\left(x, x'\right)\right] \tag{4}$$

which can also be estimated from finite samples of data by:

$$\hat{f}(x) = \frac{1}{m} \sum_{i=1}^{m} k\left(x, x_i\right) - \frac{1}{n} \sum_{i=1}^{n} k\left(x, y_i\right) \tag{5}$$

Thus, as [need citation] points at, the witness function tracks where the densities of X and Y are most different. Kernel selection is important because it decides the kind of witness functions that can be learned. For most applications it is simply set to the RBF kernel but ideally it should be selected based on maximizing test power.

As mentioned in hypothesis testing section, every two-sample test follows a similar procedure, therefore using the MMD as a measure of distance, the steps to perform a two-sample test are as follows:

1.Distance MMD2k(P,Q):

Choose a kernel k For characteristic k, k(P,Q)=0 iff P=Q 2.Estimate the distribution distance from data: MMD(X,Y) 3.Choose a rejection threshold calpha:

Under H0, mMMD converges to distribution depending on data, k

1.4 Problem Formulation

1.4.1 Change-point Detection Problem

The basic change-point problem is set up as hypothesis test between two segments of a time series. Let $X_i, ..., X_n$ be a series of independent random variables of dimension $d \geq 1$ be sequentially observed . Then, one of the following hypotheses holds:

$$\begin{cases} H_0 : X_1, X_2, ..., X_n \sim P & \text{(no change-point occured)} \\ H_1 : X_1, X_2, X_{t_0-1} \sim P, X_{t_0}, X_{t_0+1}, \sim Q & \text{(a change-point occured).} \end{cases} \tag{6}$$

Where $i = 1, 2, ..., t_0 - 1$ and $j = t_0, ..., n$ are two distinct segments separated by change-point t_0 that is within the time series window. Because we are operating in a non-parametric setting, the distributions P, Q are assumed to be completely unknown.

If there is no change in the data then we say the change time is equal to infinity and denote this probability as P^∞ and the expectation is E^∞.

Many change-point detection algorithms define a statistic that is computed using each set before and after the possible change-point,t. If the statistic is above a particular threshold then time t is classified as a change-point, $\hat{\tau}$.

In the on-line scenario, the time series under consideration can be thought of as a sliding window with data constantly coming in and out of the window of interest. The size of the window is an important consideration that is typically chosen based on the problem being solved. Too small a window and the sets of data may not yield a statistically significant result. Too large of a window and the problem leans more towards and off-line model, where high volumes must be stored and several change-points may appear in a given window. If the amount of data is not a limitation then throttling the data may not be necessary.

1.4.2 Performance Measures

Because of the unsupervised nature of detecting change-points, it is difficult to evaluate the performance of change-point detection models with real world data. Many papers detail asymptotic or non-asymptotic theoretical guarantees of their proposed change-point methods. These theoretical results are typically compared across different change-point methods for benchmarking a new algorithm.

Two main issues arise when detecting change-points in a stream of data. The first is detecting a change-point when there is no actual statistical change in the observed sequence. These are typically called false positives or *false alarms* the in change-point detection literature. The false alarm rate is defined by a metric known as the *time to false alarm* (TTFA) rate.

$$TTFA = E_{\infty}[T] = E_{\theta}[N] \tag{7}$$

Where it is the expected number of observations that must be recorded before a change-point is incorrectly detected. In other words, it is the average amount of time until a change is detected given a sequence of observations with no change. Therefore, a larger value of TTFA is preferable. From a hypothesis testing perspective, this is equivalent to rejecting H_0 in cite equation in problem statement when it should not be rejected, i.e. type I error.

The second issue is not detecting a change-point when one occurs. This could be caused by detecting a change-point much too late for it to be of any use or simply missing it altogether. For quantifying this error, the worst case detection delay (WCD) metric measures how slow a model will detect a change-point in a worst case scenario. Conversely to TTFA, lower values of WCD are preferable.

$$WCD = supesssup E_t[(T-t)^{+}|F_{t-1}] \tag{8}$$

From a hypothesis testing perspective, this is equivalent to not rejecting H_0 in cite equation in problem statement when it should be rejected, i.e. type II error.

Balancing the TTFA and WCD of an on-line detection algorithm is crucial to for an algorithm to be of any practical use. In 1971 Lorden procedures to reacting..., it was shown asymptotically that the CUSUM algorithm provides an optimal trade-off between TTFA and WCD and ,in moustakides optimal stopping times for detecting.. 1986, it was proved optimal in the non-asymptotic case as well. Note, TTFA and

WCD are also commonly referred to as ARL_0 and ARL_1 respectively where ARL stands for average run length. For clarity, we use the more explicit terms TTFA and WCD.

When detecting changes of a distribution, a user may want to quantify the size of the change in the mean by $|\mathbb{E}[X_\tau] - \mathbb{E}[X_{\tau+1}]|$ or, similarly, the size of the change in the variance by $|\text{Var}[X_\tau] - \text{Var}[X_{\tau+1}]|$.

1.4.3 Other performance measures

If labelled change-points are available for a real world dataset or a synthetic dataset, then the ground truth change-point vector, τ^*, is known. For example the *Hausdorff* metric can be used. It measures the furthest temporal distance between a predicted change-point $\hat{\tau}$ and τ^*. It is defined as:

Other standard classifier metrics can also used for comparing $\hat{\tau}$ and τ^*. This includes the F1-Score that is based on a classifier's precision and recall:

$$F_1(\hat{\tau}, \tau^*) = 2 * \frac{\text{precision*recall}}{\text{precision + recall}} \tag{9}$$

F1-Score is defined as the harmonic mean of precision and recall. Precision is defined as the ratio of true positives (TP) to the number of true positives (TP) and false positives (FP) and recall is defined as the ratio the number of true positives to the number of true positives plus the number of false negatives. F1-Score is best when F1 = 1 (perfect precision and recall) and reaches its worst value at F1 = 0. Depending on the context, any other classifier evaluation tools such as the Receiver Operating Characteristics Curve and the Precision Recall Curve may be used as well.

1.5 Classic Algorithms

Presented below are the fundamental approaches to on-line change-point detection that have been very influential.

1.5.1 Shewart Control Chart

Shewart control charts were originally designed to detect changes in the mean of a process where the values being observed are assumed to be Gaussian distributed.

As the data arrives, the data is batched into samples of size N. The sample mean, $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$, is the then calculated and compared it to a known, true mean μ^*. If the absolute difference is greater than a threshold, then a change-point is declared at the current batch. Therefore, the decision rule is defined as,

$$|\bar{X} - \mu^*| > \kappa \frac{\sigma}{\sqrt{N}} \tag{10}$$

Where κ is a constant that controls how sensitive the algorithm is. Typically, it is set to $\kappa = 3$ as this coincides with the observations within 3 standard deviations of the mean. Under the assumption that the data is distributed normally, 99.7 of the observations are distributed in this region, therefore a change-point is declared if it falls outside this region. The true mean is assumed to known and is defined as $\mu^* = \mathbb{E}[X_i]$. In applications, the true mean can also be replaced by some target specification that a process must adhere to. Similarly, it is assumed the standard deviation, σ, is known in advance but it can also be estimated.

Tuning the hyper-parameters can drastically change the performance of the algorithm. Choosing a lower value for κ makes the control chart detect change-points more often, whereas a higher value results in less detections. The chosen sample size, N, is also critical and its effect on the performance of Shewart control charts was studied in [13].

1.5.2 CUSUM

Similar to the Shewart control chart, the CUSUM algorithm tracks a statistic over time relative to a predetermined threshold. CUSUM is best applied to a process that is already under control. It can be thought of accumulating the information of current and past samples.

The algorithm is defined by a statistic, S_i, that is recursively updated after each sample, X_i, is observed, such that:

$$\begin{cases} S_0 = 0 & \text{(Initialization)} \\ S_i = \max(0, S_{i-1} + Z_i) & \text{for i=1,2,...} \end{cases} \tag{11}$$

Where $Z_i = ln(\frac{f_{\theta_1}(X_i)}{f_{\theta_0}(X_i)})$ and the statistic S_i is compared to a threshold h that is predetermined by the user. If $Z_i \geq h$ then a change-point is declared at time i and

the algorithm is either stopped or restarted. Given that the statistic only flags change-points when greater than a threshold, this algorithm only detects positive changes in the distribution. In [29], it is suggested to combine two CUSUM algorithms to detect positive and negative changes in a distributional parameter.

As a parametric algorithm, it is assumed the distributions, f_0 and f_1, are known at the outset. In most applications, this is quite limiting and unrealistic. Therefore, in cases where parameters θ_0 and θ_1 are unknown, maximum likelihood estimates of the parameters are usually computed.

The filtered-derivative extension of the CUSUM introduced in citebasseville1981edge uses the change of the discrete derivative of a signal over time to detect a change-point.

In [24], a fast initial response (FIR) CUSUM algorithm is proposed where the starting value of initial cumulative sums adapts over time. Instead of resetting S_0 to zero as shown above, it is reset to a non-zero value, typically based on the threshold chosen. This gives the algorithm a head-start in quickly detecting when a process is out of control and is especially useful for processes that don't start in control.

Finally, since CUSUM is typically better at detecting small shifts in signals and the Shewart control chart is faster at detecting larger changes, the two can be combined. The combined Shewart-CUSUM algorithm leverages the strengths of both techniques for better overall performance. See [23], [41], and [39] for more details.

1.5.3 EWMA

First described in [32] as a "geometric average", the exponentially weighted moving average (EWMA) is a type of moving average that applies exponential weighting to samples. First introduced as a forecasting technique in the econometrics field for smoothing noisy functions, the EWMA can also be used for determining out of control processes as shown in [16]. Rather than weight all observations uniformly like the standard CUSUM algorithm or a simple moving average, a decay factor (also called a forgetting factor), λ, is used to control how much weight is distributed over the previous observations. As each new observation arrives, the EWMA statistic is recursively updated and compared to a threshold. If the EWMA statistic exceeds the threshold then the process is deemed out of control or, in other words, a change-point is detected.

The EWMA statistic is calculated as follows at each time step, t:

$$S_t = \lambda X_t + (1 - \lambda) S_{t-1}$$

Where the decay factor, λ, is set to a value between 0 and 1. As λ approaches 1, the EWMA control chart gives more and more weight to the most recent observations similar to a Shewhart control chart that gives weight to the last observation only. Conversely, as λ approaches 0, the weights are distributed further into the past giving the EWMA a longer memory similar to the CUSUM algorithm. Therefore, a EWMA control chart can be interpreted as a trade-off between a Shewhart control chart and a CUSUM control chart.

For detecting deviations away from a mean target value, control limits may be calculated in a similar manner to the Shewart control chart. In [16], control limits for the EWMA are chosen to be $\pm 3 \cdot \sigma_{\text{Shewhart}} \sqrt{\frac{\lambda}{2-\lambda}}$. The standard deviation, σ_{Shewhart} , is calculated similarly to the Shewart control chart problem.

More generally, if a function, $g(x)$ is applied to the observations such that we monitor $\tau = \mathbb{E}[g(X_t)]$ over time. Then the EWMA statistic will be compared to the in-control τ^* at each iteration and is assumed to be known beforehand.

As with the other methods previously mentioned, the standard EWMA is a parametric method as it assumes the time series has some in-control average that is known prior to use. This makes is difficult to apply in situations where the data is non-numeric or where third or fourth moment changes in the distribution occur. It is however very fast due to its recursive structure and does not hold a lot of data in memory making it appealing for live data streams that rely on quick processing.

1.6 Our Contributions

The contributions of this work are several fold. First, given the strong emphasis of theoretical results in the change-point detection community, we test classic on-line change-point algorithms with modern, competitive algorithms on synthetic data and real-world data. Classification error as well as time to detection are compared as on-line methods are tasked with balancing this trade-off and comparisons between algorithms that leave out one of the these two key metrics out is incomplete. The focus of recent algorithms is on kernel methods such as KCUSUM, NEWMA and Mstats-CPD the will be discussed in Chapter 2.

Chapter 2

Kernel Changepoint Detection

2.1 Related Work

One of the first papers to use the term kernel change-point detection was in [7]. The authors present an on-line kernel change point detection model based on single class support vector machines (ν-SVMs). They train a single class support vector on a past set, $\mathbf{x}_{t,1} = x_{t-m_1}, ..., x_{t-1}$ of size m_1 and train another single class support vector on a future set $\mathbf{x}_{t,2} = x_t, ..., x_{t+m_2-1}$ of size m_2. A ratio is then computed between the two sets that acts as the dissimilarity measure in Hilbert space. If the points are sufficiently dissimilar over some predetermined threshold,η, then a change point is assigned to the time spitting the two sets of data. The authors argue that a dissimilarity measure between kernel projection of points in a Hilbert space should estimate the *density supports* rather than estimate the probability distributions of each set of points. While this approach inspired a lot of interesting research that will be discussed below, modern approaches have proven to be more practical.

In 2007, Harchoui and Cappe [12] approached the off-line change point problem with a fixed number of change points by using kernel change point detection. This was further extended to an unknown number of change points in 2012 by Arlot et al. [2]. Finally, Garreau and Arlot extended this in line of research kernel change points in the off-line setting of detecting change points. Fundamentally, their method is the kernel version of the following least squares optimization problem:

$$J(\tau, \mathbf{y}) = \frac{1}{n} \sum_{k=1}^{K} (\tau) \sum (Y_i - \hat{(}Y_k)^2 + \beta \text{pen}(\tau) \tag{12}$$

The benefits of this off-line kernel change point detection is that it operates on any kind of data for which a kernel that properly reproduces a Hilbert space can be applied. For example, it can be applied to image data, histogram data, as well as $d-$dimensional vectors in $\mathbb{R}^d$. Garreau shows their KCP procedure outputs an off-line segmentation near optimal with high probability. Lastly, the authors recommend choosing the kernel based on best possible signal to noise ratio that the distribution gives based on Δ^2/M^2 . Therefore, some prior knowledge or training set is necessary for calibrating the kernel.

In [21], the authors make use of the B-test introduced in [42] and develop an offline and online change-point detection algorithm called the MStats algorithm. The authors also refer to this algorithm as the Scan-B algorithm in a follow-up paper. At each time-step, the online model samples new data from a window of size B_0 and does a B-test with N past samples that are kept as reference samples.

$$Z_{B_0,t} := \frac{1}{N}\sum_{i=1}^{N} \mathrm{MMD}_u^2\left(X_i^{(B_0,t)}, Y^{(B_0,t)}\right)$$

The resulting test statistic is then normalized by $Z_{B_0,t}/\sqrt{Var[Z_{B_0}]}$ where the authors provide a theoretical calculation of $\mathrm{Var}[Z_B]$. If the normalized test statistic exceeds some predefined threshold then a change-point is declared. The B-test is memoryless in the sense that the statistic is calculated each time and only the value of the last calculation has any weight. This is similar to a control chart that calculates a z-score at each iteration. Adjusting the size of the window, B_0, results in the usual trade-off of performance in online change-point detection. A smaller block size will have a smaller computational cost and a smaller detection delay but will result in a worse test power resulting in higher type II error.

From here, theoretical bounds are developed for the average run length and expected detection delay. Experiments are done on real-data sets including a speech dataset and the Human Activity Sensing Consortium (HASC) dataset where the performance was better than the relative density-ratio (RDR) algorithm described in [22].

More recently in [5], a kernel change-point detection method is proposed that uses deep generative models to augment the test power of the kernel two sample test statistic. They point out MMD's lack of test power when using limited samples from the new distribution, Q, which may easily leading to over-fitting with kernels.

Thus they use a generative adversarial neural network (GAN), trained on historical samples of $X \sim P$ with noise injected into X. This surrogate distribution is then used in conjunction with possible change-points to improve the test power of a modified MMD measure that makes use of compositional kernels.

The method is compared to other prominent change-point methods for off-line change detection such as the aforementioned MStats-KCPD, LSTNet, and Gaussian process change-point models. All comparisons are done on synthetic data with piece-wise i.i.d. data. All methods are benchmarked using the AUC metric for classification performance and it is shown the KL-CPD method is competitive or better than state of the art methods. Furthermore, the AUC performance is maintained as the dimensionality of the data is increased, making their kernel learning framework very interesting for future off-line change-point detection. It remains to be seen if this framework can be adopted in an on-line context where time to detection is a key constraint on practicality.

In the on-line setting, several methods use kernel embeddings with a two-sample hypothesis test. This is done in a similar vein to the classic CUSUM and Shewart control charts. They all make use of the maximum mean discrepancy (MMD) test statistic for a two-sample kernel hypothesis test.

A modified, "no-prior-knowledge" exponentially-weighted moving average (NEWMA) is introduced in [17]. Based on the standard exponentially weighted moving average, NEWMA computes two EWMA statistics of different weights. If the difference between the two EWMA statistics exceeds a predefined threshold then a changepoint is declared at that time step. The point of using two EWMA statistics is for one to have a larger forgetting factor. This makes any recent changes in distribution to weigh heavily in one statistic, resulting in a large difference between the two statistics.

Since a standard EWMA is a parametric method, the authors apply a kernel mapping function, Ψ, to the data prior to applying the exponential factors. This provides a memoryless, non-parametric, online changepoint detection method that does not need to constantly store all previously streamed data. Once the statistics are updated at each iteration, the raw data may be discarded.

While kernel mean embeddings could be used for approximating, Ψ, as is the case for standard implementations of MMD, this would require the storage of past examples of data. Because the authors aim to reduce run-time cost and storage cost,

they use a Random Fourier Features (RFF) approach for the mapping Ψ. (Note RFF is sometimes referred to as *random kitchen sinks*) There are several approaches available for optimizing the RFF approach that are well studied in the literature. The authors make use of the standard RFF implementation, the FastFood implementation introduced in [19], and Optical Processing Unit implementation from [33].

Experiments are done with the three implementations of RFF and are compared to the MStats (Scan-B) algorithm using a Gaussian kernel. Experiments with synthetic datasets are run using streaming data that is generated from different Gaussian mixture models. They also use an audio dataset for testing on real data. The variants of NEWMA are very similar, if not better than MStats (Scan-B) in terms of missed detection percentage. In terms of the detection delay and false alarm trade-off, the NEWMA algorithm and its variants appear to be mildly better as well. The largest advantage of the NEWMA variants over the MStats (Scan-B) method is in the execution time. Because MStats (Scan-B)' s execution scales linearly with window size, while NEWMA's execution time does not depend on window size.

Finally, in a recent, preprint paper [8], a kernel CUSUM (KCUSUM) algorithm is proposed, where the classic CUSUM algorithm is adapted using the MMD statistic for on-line detection. The authors use a modified, unbiased MMD statistic that can be computed in linear time. This formulation of the MMD statistic was originally defined in section 6 of [10] as:

$$\mathrm{MMD}_l^2[\mathcal{F}, X, Y] := \frac{1}{m_2} \sum_{i=1}^{m_2} h\left(\left(x_{2i-1}, y_{2i-1}\right), \left(x_{2i}, y_{2i}\right)\right)$$

Where,

$$h\left(\left(x_i, x_j\right), \left(y_i, y_j\right)\right) := k\left(x_i, x_j\right) + k\left(y_i, y_j\right) - k\left(x_i, y_j\right) - k\left(x_j, y_i\right)$$

The algorithm functions as follows, every two observations, the MMD_l is calculated using newly observed data points and data points sampled from some *reference* distribution that is known at the outset. This reference distribution can be thought of as the "in-control" distribution of the data-stream that new observations are compared to. The calculated MMD_l acts as the update term to the cumulative sum statistic, hence the name KCUSUM. If this kernel cumulative sum statistic exceeds some predefined threshold, then a change-point is identified.

Besides its speed of computation, an additional benefit of MMD_l is it is normally distributed under the null distribution unlike the quadratically-calculated MMD. This facilitates analysis of bounds and provides statistical guarantees for worst-case detection delays and time to false alarm rate. While this non-parametric approach can detect any change in the distribution of a sequence, it does struggle with more complicated distributional changes such as variance changes of a single dimension and changes beyond first and second-order moments.

2.2 Our Approach

This section describes our novel method for change-point detection.

Chapter 3

Experimental Results

3.1 Datasets

A common difficulty in change-point detection is evaluating the performance of an algorithm with datasets that aren't overly simplistic and difficult enough to ascertain some real world use.

Unlike fields like image recognition where datasets like MNIST provide a common benchmark, there are no standard datasets that are widely used across the change-point detection literature for evaluating new methods. Most papers propose experiments that are relevant for the specific problem they are trying to solve but lack examples or explanations of when their method would not be applicable. Furthermore, because change-point evolved out of the statistics literature, many papers focus on theoretical results and provide minor experimental results if any.

Given the empirical focus of this thesis, we attempt to put together the most comprehensive, controlled experiments using both synthetic and real-world datasets.

3.1.1 Synthetic Datasets

To the best of our knowledge, no change-point detection paper covers as many variations as presented in this thesis. While synthetic datasets are idealistic in their formulation, they provide a good starting point for comparing different methods because many variables can be controlled for. Often in the real world, the exact location of change-points is not known. Therefore, it is important for the evaluation of a change-point detection algorithm that it performs competitively on synthetic data.

Inspired by recent papers [5] and [8] that attempt to bridge the gap between the statistics and machine-learning literature, the following synthetic datasets are created: change in mean, scaling variance, alternating between two Gaussian mixtures, and alternating between random distributions. It is truly hard to properly generalize all the possible situations a non-parametric algorithm may be used in, but the synthetic cases presented in this thesis are common across domains and cover a range of applications.

For a change in mean, a change-point is inserted in the time series at some random time where the mean is shifted either positively or negatively. There are two variants to this scenario. In the first, the mean change is in all dimensions simultaneously. In the second variation, the mean change is in only one dimension making it harder to detect.

For each experiment above, a Monte-Carlo approach is used to estimate time to false alarm, detection delay, and test power.

3.1.2 Real World Datasets

In addition to synthetic datasets, several real datasets that are publicly available are also used.

Dataset	Type	No. of Dimensions	Length	No. of Changepoints
Mean (all dimensions)	Synthetic	8.872	16.128	1.402
Mean (single dimension)	Synthetic	−2.509	3.442	0.299
3	Synthetic	−2.509	3.442	0.299
4	Synthetic	−2.509	3.442	0.299
5	Synthetic	−2.509	3.442	0.299
6	Synthetic	−2.509	3.442	0.299
7	Real	−0.363	1.826	0.159
8	Real	−0.597	0.598	0.052

Table 1: Datasets Summary

Chapter 4

Conclusion

4.1 Summary of the thesis

Add description of thesis, summarizing the key points from each chapter and what this thesis attempted to add to the literature.

4.2 Discussion and Future Work

One focus of this thesis was using the MMD measure for two-sample testing and, by extension, for change-point detection. It would interesting to explore other statistical distances that could compare two distributions non-parametrically. It would be interesting to re-use the same approaches explored in this thesis, but with the MMD swapped for a different distance measure.

Recently, interesting research has been made on re-purposing binary classifiers as two-sample tests. In [14], random forests are compared to different implementations of MMD. The authors run many tests comparing the test power across the different two-sample tests and their random forest. The results are interesting because while the random forest classifier is not better in every situation, it is better on hard two-sample tests such as the blobs dataset. Therefore, it remains to be seen whether these classifier approaches to two-sample testing can be adapted to on-line change-point detection in an efficient manner.

- lopez paper

Bibliography

[1] Samaneh Aminikhanghahi and Diane J Cook. A survey of methods for time series change point detection. *Knowledge and information systems*, 51(2):339–367, 2017.

[2] Sylvain Arlot, Alain Celisse, and Zaıd Harchaoui. Kernel change-point detection. *arXiv preprint arXiv:1202.3878*, 6, 2012.

[3] Marcel Bosc, Fabrice Heitz, Jean-Paul Armspach, Izzie Namer, Daniel Gounot, and Lucien Rumbach. Automatic change detection in multimodal serial mri: application to multiple sclerosis lesion evolution. *NeuroImage*, 20(2):643–656, 2003.

[4] E Brodsky and Boris S Darkhovsky. *Nonparametric methods in change point problems*, volume 243. Springer Science & Business Media, 2013.

[5] Wei-Cheng Chang, Chun-Liang Li, Yiming Yang, and Barnabás Póczos. Kernel change-point detection with auxiliary deep generative models. *arXiv preprint arXiv:1901.06077*, 2019.

[6] Jie Chen and Arjun K Gupta. *Parametric statistical change point analysis: with applications to genetics, medicine, and finance*. Springer Science & Business Media, 2011.

[7] Frédéric Desobry, Manuel Davy, and Christian Doncarli. An online kernel change detection algorithm. *IEEE Trans. Signal Processing*, 53(8-2):2961–2974, 2005.

[8] Thomas Flynn and Shinjae Yoo. Change detection with the kernel cumulative sum algorithm. *arXiv preprint arXiv:1903.01661*, 2019.

[9] Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *Advances in neural information processing systems*, pages 489–496, 2008.

[10] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[11] Arthur Gretton, Ralf Herbrich, Alexander Smola, Olivier Bousquet, and Bernhard Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6(Dec):2075–2129, 2005.

[12] Zaid Harchaoui and Olivier Cappé. Retrospective mutiple change-point estimation with kernels. In *2007 IEEE/SP 14th Workshop on Statistical Signal Processing*, pages 768–772. IEEE, 2007.

[13] Salah Haridy, Ahmed Maged, Saleh Kaytbay, and Sherif Araby. Effect of sample size on the performance of shewhart control charts. *The International Journal of Advanced Manufacturing Technology*, 90(1-4):1177–1185, 2017.

[14] Simon Hediger, Loris Michel, and Jeffrey Näf. On the use of random forest for two-sample testing. *arXiv preprint arXiv:1903.06287*, 2019.

[15] Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. Kernel methods in machine learning. *The annals of statistics*, pages 1171–1220, 2008.

[16] J Stuart Hunter. The exponentially weighted moving average. *Journal of quality technology*, 18(4):203–210, 1986.

[17] Nicolas Keriven, Damien Garreau, and Iacopo Poli. Newma: a new method for scalable model-free online change-point detection. *arXiv preprint arXiv:1805.08061*, 2018.

[18] Marc Lavielle and Gilles Teyssiere. Adaptive detection of multiple change-points in asset price volatility. In *Long memory in economics*, pages 129–156. Springer, 2007.

[19] Quoc Viet Le, Tamás Sarlós, and Alexander Johannes Smola. Fastfood: Approximate kernel expansions in loglinear time. *arXiv preprint arXiv:1408.3060*, 2014.

[20] Tze-San Lee. Change-point problems: bibliography and review. *Journal of Statistical Theory and Practice*, 4(4):643–662, 2010.

[21] Shuang Li, Yao Xie, Hanjun Dai, and Le Song. M-statistic for kernel change-point detection. In *Advances in Neural Information Processing Systems*, pages 3366–3374, 2015.

[22] Song Liu, Makoto Yamada, Nigel Collier, and Masashi Sugiyama. Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks*, 43:72–83, 2013.

[23] James M Lucas. Combined shewhart-cusum quality control schemes. *Journal of Quality Technology*, 14(2):51–59, 1982.

[24] James M Lucas and Ronald B Crosier. Fast initial response for cusum quality-control schemes: give your cusum a head start. *Technometrics*, 24(3):199–205, 1982.

[25] Rakesh Malladi, Giridhar P Kalamangalam, and Behnaam Aazhang. Online bayesian change point detection algorithms for segmentation of epileptic activity. In *2013 Asilomar Conference on Signals, Systems and Computers*, pages 1833–1837. IEEE, 2013.

[26] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Arthur Gretton, and Bernhard Schölkopf. Kernel mean estimation and stein effect. In *International Conference on Machine Learning*, pages 10–18, 2014.

[27] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

[28] Yue S Niu, Ning Hao, Heping Zhang, et al. Multiple change-point detection: A selective overview. *Statistical Science*, 31(4):611–623, 2016.

[29] Ewan S Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

[30] Andrey Pepelyshev and Aleksey S Polunchenko. Real-time financial surveillance via quickest change-point detection methods. *arXiv preprint arXiv:1509.01570*, 2015.

[31] Aaditya Ramdas, Sashank Jakkam Reddi, Barnabás Póczos, Aarti Singh, and Larry Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[32] SW Roberts. Control chart tests based on geometric moving averages. *Technometrics*, 1(3):239–250, 1959.

[33] Alaa Saade, Francesco Caltagirone, Igor Carron, Laurent Daudet, Angélique Drémeau, Sylvain Gigan, and Florent Krzakala. Random projections through multiple optical scattering: Approximating kernels at the speed of light. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6215–6219. IEEE, 2016.

[34] Walter Andrew Shewhart. *Economic control of quality of manufactured product*. ASQ Quality Press, 1931.

[35] Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pages 13–31. Springer, 2007.

[36] M Staudacher, S Telser, A Amann, H Hinterhuber, and M Ritsch-Marte. A new method for change-point detection developed for on-line analysis of the heart beat variability during sleep. *Physica A: Statistical Mechanics and its Applications*, 349(3-4):582–596, 2005.

[37] Alexander Tartakovsky, Igor Nikiforov, and Michele Basseville. *Sequential analysis: Hypothesis testing and changepoint detection*. Chapman and Hall/CRC, 2014.

[38] Charles Truong, Laurent Oudre, and Nicolas Vayatis. A review of change point detection methods. *arXiv preprint arXiv:1801.00718*, 2018.

[39] JO Westgard, T Groth, T Aronsson, and CH De Verdier. Combined shewhart-cusum control chart for improved quality control in clinical chemistry. *Clinical Chemistry*, 23(10):1881–1887, 1977.

[40] Ping Yang, Guy Dumont, and John Mark Ansermino. Adaptive change detection in heart rate trend monitoring in anesthetized children. *IEEE transactions on biomedical engineering*, 53(11):2211–2219, 2006.

[41] Emmanuel Yashchin. On the analysis and design of cusum-shewhart control schemes. *IBM Journal of Research and Development*, 29(4):377–391, 1985.

[42] Wojciech Zaremba, Arthur Gretton, and Matthew Blaschko. B-test: A non-parametric, low variance kernel two-sample test. In *Advances in neural information processing systems*, pages 755–763, 2013.