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

April 2019

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 pushing the limit of my intellectual understanding to places I have never thought possible.

I would also like to thank all my colleagues at DRW who have helped me learn about the finance industry over the past two 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 market micro-structure, 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 1930s Walter Shewart recognized the need for monitoring manufacturing processes in real-time in a robust, systematic way. Thus the development and use of statistical control charts was developed for quickly out of control processes. This allowed for the real-time detection of changes in variation that indicated a degradation in quality in the production process. Shewart's method was one of the first formal frameworks to solve the problem of detecting changes in a distribution of observations that arrive one at a time. This problem would come to be known more generally as the *change point detection problem* and would come to apply across various industries. The following are a few motivating examples.

1.1.1 Health Care

Health care is important area for quickly detecting signal changes in monitoring equipment [cite 6,7,8,9 of global survey]. It is easy to see why in the context of taking decisions for medical treatment, quickly detecting changes to a patient's health is absolutely necessary for any system to be of practical use. Furthermore, quickly detecting changes in a patient's health must be balanced with the accuracy of these detections. Detecting a false positive or missing a detection could have life-threatening consequences and thus any real-time method must be robust to these types of errors.

1.1.2 Financial Applications

The application of accurate and timely change point detection is also very appealing to the finance sector where shifts in asset prices can happen suddenly. Change points in the financial literature are sometimes referred to structural breaks, but for this thesis will use the broader term change points. Pepelyshev and Polunchenko [6]

from tarokvsky 2015 book: In addition, we mention the articles [52, 358] and references therein. We also argue that quickest change point detection schemes can be effectively applied to the analysis of financial data. In particular, quickest change point detection problems are naturally associated with rapid detection of the appearance of an arbitrage in a market [421].

1.2 Characteristics of the change point problem

A number of surveys of the literature already exist (Aminikhanghahi and Cook [1]), 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 is the nature of the data that is being observed. If data is generated by a known distribution, such as a univariate Gaussian, then leveraging this information can be very powerful. These models are called *parametric* and come with a greater set of assumptions when applying change point methods. Shewart control charts, CUSUM and autoregressive models are all parametric methods based on XXXXXX. For an extensive review, see the monograph by Chen and Gupta [3]. On the other hand, if the data is being generated by processes that have no fundamental or well-understood process then modelling particular probability distributions becomes intractable or risky. Therefore, models that do not model particular probability distributions must be used. Methods that fall into this category are known as *non-parametric* models and are more flexible for data that do not easily fall into a well known structure. See and Nonparametric methods in change point problems E Brodsky, BS Darkhovsky.

The second consideration is to determine what in what scenario the data is analysed. Some algorithms are off-line or batch algorithms in that they are applied in an ex-post fashion after the dataset has been completely acquired. For a recent review of off-line methods see Truong et al. [8]. The aforementioned Shewart control chart and CUSUM algorithm were both designed for data that is streamed in a near, real-time fashion. In the literature, on-line methods of change point detection are also sometimes called sequential change point detection. For this thesis, we will use the terms interchangeably. For an exhaustive review of sequential change point analysis, see the book by Tartakovsky, Nikiforov, and Basseville [7].

The third consideration that is a consequence of detecting change points in an on-line setting is the presence of outliers. This a particularly important concern for on-line methods that are looking for changes in the underlying distribution of points, not a single anomalous data point

The fourth consideration is determining if there are multiple change points or to assume there is only a single change point to detect. This is a more relevant consideration for off-line change point detection, but 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.

Finally, the last point to address is determining exactly what kinds of statistical changes is an algorithm aiming to detect. Many methods focus solely on detecting changes in the mean of a distribution. Other methods focus solely on detecting changes in the variance of a distribution. Finally, some methods are more general and can detect moment changes past the variance and do not focus on any particular one. Methods like kernel change point detection typically operate in this category.

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

1.3 Related Work

In 2005, Desobry et al. [4] developed an on-line kernel change point detection model that ...... In 2007, Harchoui and Cappe approached the off-line, change point problem with a fixed number of change points by .... This was further extended to an unknown number of change points in 2012 by Arlot et al. [2]

Finally, Garreau and Arlot extended kernel change points in the off-line setting of detecting change points. They tackle the possibility of multiple change points. Fundamentally, their method is the kernel version of the following least squares optimization problem:

The benefits of this 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. His method is theoretically sound and tested but has not been implemented on very many real datasets nor compared with other benchmark methods. In an application setting, the user would use some training set to calibrate the kernel and the penalty parameters, to then be tested with appropriate accuracy measures on some out of sample data set.

Other related work include a 2015 study by Li et al. [5] that propose the use of M-statistics based on the kernel maximum mean discrepancy (MMD) B-test statistic for two-sample testing.

1.4 Problem Formulation

In the 2017 survey done by , the authors give the following definition of a *change point*: A change point represents a transition between different states in a process that generates the time series data.

Generally speaking, the fundamental change point problem is the following:

- $H_0 : P_{X_m} = P_{X_{m+k}}$
- $H_A : P_{X_m} \neq P_{X_{m+k}}$

1.4.1 Performance Measures

Add information about what performance measures will be used to determine the utility of algorithm. E.g. time to detection, F1 measure, etc.

1.5 Our Contributions

Lay out the contributions of this work.

Chapter 2

Chapter 2

2.1 Chapter 2 Section

2.1.1 Chapter 2 Subsection

Subsection Test

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

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

[5] 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.

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

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

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