On Aspects of Changepoint Analysis Motivated by Industrial Applications

Thomas D. J. Grundy, B.Sc.(Hons.), M.Res

Submitted for the degree of Doctor of Philosophy at Lancaster University.

November 2021

STOR-i
excellence with impact

Abstract

In numerous industrial applications, organizations wish to monitor time series data to better understand the past and make predictions for the future. To achieve this, appropriate analysis and modelling of time series data needs to be performed. However, it is common for time series to undergo abrupt structural changes, known as changepoints, which can cause major challenges during model fitting. Identifying changepoints is crucial to appropriately understand, analyze and model time series data in a wide range of applications including finance, genomics and countless others. In this thesis, we introduce new methodologies and frameworks for detecting changepoints in two scenarios; when there are multiple series to analyze simultaneously (multivariate setting) and when we receive time points sequentially and aim to detect changes as soon as possible (sequential setting). These methodologies provide novel and innovative ways to detect different types of changepoints in a wide range of data structures. Firstly, we introduce a novel bivariate test statistic for detecting changes in mean and variance simultaneously in multivariate data. Our attention then turns to changes in covariance, specifically, we introduce a cost function and changepoint framework for identifying changes in subspace. Finally, we switch to the

sequential changepoint setting and provide a framework for quickly identifying mean and variance changes in more complex data structures using forecast models. For all the methods we demonstrate their strong empirical performance in both simulated examples and industrial applications.

Acknowledgements

There are numerous people who deserve thanks and whom this PhD would not have been possible without. Most importantly, I would like to thank my supervisor Dr. Rebecca Killick for their constant help, encouragement and expertise throughout this PhD. In particular, their unwavering reassurance that 'simple ideas can be the best ideas' has led me through a number of difficult moments and lapses in confidence. I would also like to thank Dr. Ivan Svetunkov whose forecasting expertise has been greatly appreciated during the later stages of this PhD. Thanks also go to my external examiner Dr. Erik Lindström and my internal examiner Dr. Nicos Pavlidis for an enjoyable and interesting discussion of my PhD during the viva process.

I'm very grateful for the financial support provided by the STOR-i CDT, EPSRC and Royal Mail. Thanks also go to my Royal Mail supervisors over the past few years. In particular, to Dr. Jeremy Bradley and Dr. Gueorgui Mihaylov whose encouragement and kindness during the early stages of my PhD, and my trips down to Royal Mail, is greatly appreciated. In addition, I would like to thank Zhao and many others from the Royal Mail data science team, past and present, who have been incredibly helpful and provided valuable insight into the workings of a data science team.

My PhD experience has, on the whole, been a positive one and a major reason for this has been the bond and friendship within my cohort and across the whole of STOR-i. From STOR-i balls to nightcaps in Hustle, the past few years have been some of the most enjoyable of my (long) university life and made the challenges of a PhD substantially easier. Thank you to Matt, Alan and Mirjam for being great house mates over the past years and for enduring my eagerness to maintain an undergraduate lifestyle.

I would also like to thank my family for their support and keeping me grounded during the PhD process. I have no doubt they have helped with the timely completion of my PhD with their light-hearted, encouraging comments - I am finally going to get a 'real' job. My final thanks go to Maddy, you made my PhD life thoroughly more enjoyable from our sourdough experiments, to camping trips in the Lakes, to numerous board games nights together. You helped make my time during this PhD one of the best I can remember and for that I can't thank you enough.

Declaration

I declare that the work in this thesis has been done by myself and has not been submitted elsewhere for the award of any other degree.

Chapter 3 has been accepted for publication as Grundy, T., Killick, R., and Mihaylov, G. (2020). High-Dimensional Changepoint Detection via a Geometrically Inspired Mapping. *Statistics and Computing*, 30(4):1155–1166.

Chapter 4 was submitted for publication, however, upon review, a connection with independently developed work was highlighted. This is discussed in Chapter 6.

Chapter 5 is currently under review for publication.

Thomas Grundy

Contents

Abstract **I**

Acknowledgements **III**

Declaration **V**

Contents **X**

List of Figures **XIV**

List of Tables **XV**

1 Introduction **1**

2 Overview of Changepoint Detection **6**

2.1 Univariate Changepoint Detection 7

2.1.1 Cost Function/Test Statistic 10

2.1.2 Search Methods 14

2.1.3 Penalization 23

2.1.4 Alternative Methods 29

2.2 Multivariate Changepoints 30
2.2.1 Fully Multivariate Changepoints 32
2.2.2 Sparse Changepoints 41

3 High-Dimensional Changepoint Detection via a Geometrically Inspired Mapping 49

3.1 Introduction 49
3.2 Methodology 52
3.2.1 Problem Setup 52
3.2.2 Geometric Mapping 53
3.2.3 Analyzing Mapped Time Series 57
3.2.4 GeomCP Algorithm 58
3.2.5 Non-Normal and Dependent Data 60
3.3 Simulation Study 62
3.3.1 Size of Changepoints 64
3.3.2 GeomCP Investigation 66
3.3.3 Dense Changepoints 67
3.3.4 Sparsity Investigation 69
3.3.5 Between-series Dependence 70
3.3.6 Computational Speed 72
3.4 Applications 74
3.4.1 Comparative Genomic Hybridization 74
3.4.2 SP500 Stock Prices 77

3.5 Conclusion 79

4 Subspace Changepoint Detection in Multivariate Time Series 80

4.1 Introduction 80

4.1.1 Notation 83

4.2 Methodology 84

4.2.1 Subspace Estimation 85

4.2.2 Test Statistic 87

4.2.3 Threshold Choice 88

4.2.4 Alternative Formulation of Cost Function 90

4.3 Simulation 91

4.3.1 Data Generation 91

4.3.2 Power of Test Statistic 93

4.3.3 Permutation Test 95

4.3.4 Method Comparison 97

4.3.5 Sensitivity to Subspace Dimension 102

4.4 Extension: Multiple Changepoints 104

4.4.1 Simulation 105

4.5 Application: Motion Capture Data 107

4.6 Conclusion 108

5 Identifying Sequential Changes in Mean and Variance Within More Complex Model Structures 110

5.1 Introduction 110

5.2 Sequential Changepoint Detection 115
5.3 Monitoring Forecast Errors 121
5.3.1 Mean Changes 123
5.3.2 Mean and Variance Changes 126
5.4 Common Forecasting Models 130
5.4.1 ARMA Models 130
5.4.2 ETS Models 133
5.5 Simulation 135
5.5.1 Mean Change 139
5.5.2 Variance Change 140
5.5.3 Mean and Trend Changes in Seasonal Data 142
5.5.4 Change in Autoregressive Parameter in Seasonal Data 146
5.6 Application 148
5.6.1 Parcel Delivery Volumes 148
5.6.2 GRP Admissions 149
5.7 Conclusion 152

6 Conclusion **154**

6.1 Key Findings 154
6.2 Open Problems and Future Work 156

A High-Dimensional Changepoint Detection via a Geometrically Inspired Mapping **159**

A.1 Preliminary Lemmas 159

A.2 Proof of Theorem 1 161

A.3 Dense Mean Changepoints 162

A.4 Dense Mean and Variance Changepoints 163

A.5 Sparse Variance Changepoints 165

A.6 Between-series Dependence: Mean Change 166

A.7 Performance under the Null 167

A.8 Dense Change Size Investigation 168

A.9 CROPS diagnostics plots 169

B Subspace Changepoint Detection in Multivariate Time Series 173

B.1 Additional Simulation Results: ROC Curves 173

B.2 Additional Simulation Results: Permutation Test 175

B.3 Application: Justification of Parameter Choices 176

Bibliography 179

List of Figures

1.0.1 Simulated heart rate of an individual before and during exercise . . . 2

2.1.1 Candidate changepoints for the explanation of OP and PELT 17

3.2.1 2-dimensional example of pre-processing translation 56

3.3.1 Distance and angle mappings within GeomCP 66

3.3.2 Locations of detected changepoints in a simulated data set 67

3.3.3 TDR and FDR for GeomCP and E-Divisive for variance changes . . . 68

3.3.4 TDR and FDR for GeomCP, Inspect and E-Divisive for sparse mean changes 70

3.3.5 TDR and FDR for GeomCP and E-Divisive for covaraince changes . . 72

3.3.6 Average run times of GeomCP, E-Divisive and Inspect 75

3.4.1 Log-intensity-ratio measurments of microarray data with identified changepoints 77

3.4.2 Log-returns within the S&P500 with identified changepoints 78

4.3.1 ROC curve for varying change sizes 94

4.3.2 Histogram of changepoint positions for varying change sizes 95

4.3.3 FPR for data containing no changepoints 97

4.3.4 TPR for data containing changepoints 98

4.3.5 Estiamted changpeoint locations for 3 different scenarios 101

4.3.6 ROC curves for misspecified subspace dimensions 103

4.4.1 Histogram of estimated changepoint locations and number of changepoints 106

4.5.1 Motion capture data with identified changepoints 108

5.4.1 Data generated from an ARMA(1,1) model with a change in mean at time point 400; the resulting forecasting errors when using an ARMA forecasting model; and the detector, $D_\mu(m, k)$, where the dashed line shows the associated threshold for detecting a change. The vertical dotted line represents the start of the monitoring period. 132

5.4.2 Data with a variance change at time point 400, with dashed lines showing prediction intervals from an ETS forecasting model. The squared forecast errors are shown along with the Detector with the dashed line being the threshold. The vertical dotted line shows the start of the monitoring period. 136

5.5.1 Detection Probability (DP), False Detection Probability (FDP) and Average Detection Delay (ADD), with error bars showing 2 standard errors either side of the mean, for two scenarios with mean changes of varying sizes. 141

5.5.2 Detection Probability (DP), False Detection Probability (FDP) and Average Detection Delay (ADD), with error bars showing 2 standard errors either side of the mean, for two scenarios with variance changes of varying sizes. 143

5.5.3 Detection Probability (DP), False Detection Probability (FDP) and Average Detection Delay (ADD), with error bars showing 2 standard errors either side of the mean, for two scenarios; a change in mean and a change in trend within seasonal data. 145

5.5.4 Detection Probability (DP), False Detection Probability (FDP) and Average Detection Delay (ADD), with error bars showing 2 standard errors either side of the mean, for two scenarios with a change in autoregressive parameter of different sizes 147

5.6.1 Forecast errors for the number of parcels to be delivered from a specific Royal Mail delivery office in 2020; the detctors $D_\mu(m, k)$ and $D_\xi(m, k)$ with associated thresholds shown by the dashed lines. 150

5.6.2 Data showing the proportion of GRP A&E admissions; the forecast errors from a SARIMA$(1, 0, 0)(1, 0, 0)_{12}$ model; and the detectors $D_\mu(m, k)$ and $D_\xi(m, k)$ with associated thresholds shown by the horizontal dashed lines. The dotted line represents the start of the monitoring period and the vertical dashed line indicates a changepoint identified in the retrospective analysis. 152

A.4.1TDR and FDR for GeomCP and E-Divisive for mean and variance changes 165

A.5.1TDR and FDR for GeomCP and E-Divisive for sparse variance changes 166

A.6.1TDR and FDR for GeomCP and E-Divisve for mean changes in data with a non-diagonal covariance structure 168

A.7.1FPR for GeomCP for data with no changepoints 169

A.9.1CROPS diagnostic plots for comparative genoic hybridization data . . 172

A.9.2CROPS diagnostic plots for S&P500 log-returns data 172

B.1.1ROC curves for varying signal to noise ratios 174

B.1.2ROC curves for scenarios with varying p and q. 175

B.2.1FPR for data containing no changes 176

B.2.2TPR for data containing a changepoint under multiple scenarios. . . . 177

B.3.1Scree plot of eigenvalues and total cost for increasing changepoints for Motion Capture data 178

List of Tables

A.3.1TDR and FDR for GeomCP, Inspect and E-Divisive for mean changes 164

A.8.1TDR and FDR for GeomCP, Inspect and E-Divisive for dense mean changes of varying sizes 170

A.8.2TDR and FDR for GeomCP and E-Divisive for variance changes . . . 171

Chapter 1

Introduction

The volume of data being collected within industry is rapidly increasing. Naturally, organizations want to utilize and learn from this data to make informed decisions across all areas of the business. This drive towards data-driven decision making leads to countless opportunities to develop new statistical methodologies to better understand, analyze and model the ever-increasing volumes of data.

A common data structure encountered in industrial problems are time series. Here data is collected at discrete time intervals, usually equally spaced, and we aim to determine patterns that develop over time. This allows us to better understand the past and make predictions for the future. To appropriately model time series data it is often assumed that the data, or a transformation of the data, is stable over time. This usually means the data follows a predictable pattern and has, for example, a common mean and variance. However, in many industrial problems, time series can often exhibit abrupt structural changes where the properties of the time series change

in some way; we call these changepoints.

In Layman's terms, changepoints are points in time when the expected behavior of a system changes abruptly. A common example is an individuals heart rate and a simulated example can be seen in Figure 1.0.1. From 09:00 to 11:00 the individual is resting and their heart rate is low. Then, the individual starts exercising and we see their heart rate abruptly changes - it becomes much higher and more variable. The time point when the individual starts exercising, and their heart rate changes (here 11:00), is what we call a changepoint.

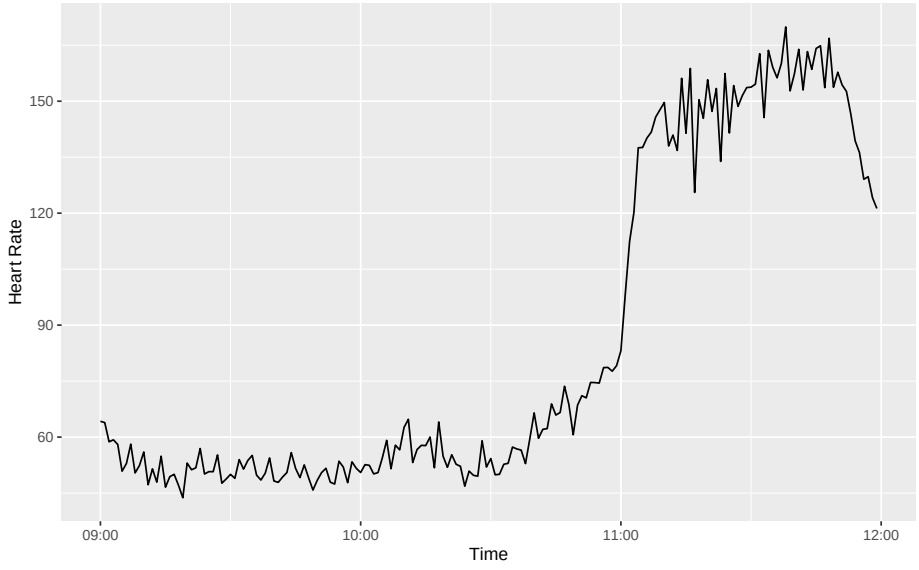


Figure 1.0.1: Simulated heart rate of an individual before and during exercise

Changepoints can cause a number of problems when aiming to model time series. Firstly, if we don't take a changepoint into account then we are likely trying to fit a model to data that is generated from a mixture of models and therefore we will struggle to capture the key properties of the time series adequately. Furthermore, when we make forecasts using time series, we want to be forecasting based on the current state

of the model. Hence, using data prior to a changepoint can add unwanted noise to a model decreasing forecast accuracy. Thus, it is crucial to identify changepoints to allow organizations to improve their understanding and analysis of their data and fit more appropriate models for improved decision making.

There are a number of well-known examples of changepoints that have affected many different industrial areas. Most topically, the outbreak of COVID-19 is a prime example. Many industries, from the retail to the hospitality sectors, saw a major changepoint in March 2020. Intuitively, it makes little sense to use data from before this pandemic to try and make predictions during the pandemic; this is an obvious example of accounting for changepoints in data. Moreover, changepoint detection has been used to assess interventions aimed at reducing the spread of COVID-19 (Dehning et al., 2020). Changepoints aren't always so obvious or widespread and can be more industry-specific. There have been various works on changepoint analysis in a wide range of industrial applications including Epidemiology (Carroll et al., 2019), finance and economics (Bai and Perron, 1998; Modisett and Maboudou-Tchao, 2010), genetics (Bleakley and Vert, 2011), internet security (Peng et al., 2004), oceanography (Killick et al., 2010) and countless other domains

Here, we have defined changepoints as points of abrupt change in a system. However, systems can also vary slowly over time. Consider Figure 1.0.1, we have already identified the abrupt change when the individual starts exercising, however, we can also see more subtle slowly evolving changes. For example, after the individual starts exercising their heart rate is slowly increasing - this could be seen as a slowly evolv-

ing system and we may wish to keep updating the estimate of the mean heart rate, despite there being no clear changepoints. This thesis only concerns abrupt changes, for more details on slowly evolving systems see Ljung and Söderström (1983).

The study of changepoints in a single time series, also known as univariate changepoint analysis, has a long history dating back to Page (1954). Yet, as more data is being collected there has become a need for multiple related time series to be analyzed simultaneously, hereby known as multivariate changepoint analysis. This raises additional challenges as we can have dependence between the time series as well as across time or changepoints only occurring in a subset of the series. On top of this, we need to consider the additional computational cost of analyzing more and more data especially as the number of time points and series grows large. Chapter 2 gives a more in-depth review of the changepoint literature in both the univariate and multivariate settings. Addressing some of these issues when moving from the univariate to the multivariate setting is the focus of Chapters 3 and 4.

In some industrial problems, we want to identify changepoints in a sequential manner as more data points become available. Here we do not have all the data available to us prior to analysis as is often assumed. Thus, the focus switches from accurately locating changepoints to detecting changepoints as soon as possible after they have occurred. This alternative changepoint paradigm is the focus of Chapter 5.

This thesis contains three main contributions to the changepoint literature, two in the multivariate setting and one from the sequential changepoint setting. In all these works, we produce novel methodology and apply this to varying industrial problems

to demonstrate the impact they can have in a wide range of applications.

In Chapter 3, we propose a new method, named GeomCP, for detecting mean and variance changes in high-dimensional time series via a bivariate test statistic. We perform dimension reduction via two novel geometric projections. These projections preserve the changepoint locations and allow us to use univariate changepoint techniques upon them to detect changepoints in mean and variance simultaneously in a computationally efficient manner.

In Chapter 4, we introduce subspace changepoints. It is common for high-dimensional data to exhibit some low-dimensional structure, such as lying in a linear subspace. At some time point, this subspace could abruptly change, resulting in a subspace changepoint. We propose an appropriate cost function and adapt this into a changepoint framework to detect these changes.

In Chapter 5, we move from the multivariate setting to the sequential changepoint setting. We propose a framework for sequentially monitoring forecast models to quickly identify if they become inaccurate. We show how changepoints in the time series being forecast will manifest in forecast errors and propose a sequential changepoint method for detecting these changes. This framework allows us to detect mean and variance changes in more complex data structures and quickly identify if forecasting models need re-evaluating.

Finally, Chapter 6 gives concluding remarks, highlighting the key findings in each Chapter along with open problems and future work.

Chapter 2

Overview of Changepoint Detection

Within this chapter, we give an overview of changepoint detection including historical approaches to the classical problem, recent advancements and open areas of research, of which some are explored in this thesis.

Univariate changepoint analysis is a well-studied research area since its introduction back in Page (1954). The univariate problem forms the basis for the literature and the extensions to more complex problems discussed in the thesis. Hence, Section 2.1 is dedicated to summarizing the univariate changepoint literature. For a general discussion of likelihood and Bayesian approaches to the univariate changepoint problem see Eckley et al. (2011); for a non-parametric overview see Brodsky and Darkhovsky (2013), and for theoretical contributions see Csörgö and Horváth (1997). Truong et al. (2020) provide a more recent review that covers both the univariate and multivariate

settings. Note throughout this thesis we only consider the frequentist paradigm with a very brief discussion of some popular Bayesian methods given in Section 2.1.4.

In recent years, the focus has shifted to the multivariate setting where we detect changes in multiple streams of data simultaneously. This raises new questions and challenges such as the number of streams the change occurs in, also known as the sparsity of the change, along with the computational challenge of analyzing additional data steams. Section 2.2 outlines these challenges in more detail and gives an overview of the current methodology developed to solve these challenges. This leads on to Chapters 3 and 4 where we present work to solve two different problems in the multivariate changepoint framework. Firstly, we present the GeomCP algorithm for the simultaneous detection of mean and variance changes and secondly a method for detecting subspace changepoints we can be seen as a covariance change with additional constraints on the covariance structure.

2.1 Univariate Changepoint Detection

Here we introduce the changepoint setup, splitting this Section into three main properties present in the majority of changepoint methods. We conclude this section with a summary of some methods which lie outside of this framework yet still warrant a mention.

The generic univariate changepoint problem is as follows. Let $y_{1:n} = (y_1, \ldots, y_n)$ with $y_i \in \mathbb{R}$ for $1 \leq i \leq n$ be a sequence of data which contains $m < n$ changepoints

located at time points $0 = \tau_0 < \tau_1 < \ldots < \tau_m < \tau_{m+1} = n$. We define the set of changepoints as $T = \{\tau_1, \ldots, \tau_m\}$ and these m changepoints split the data into $m+1$ segments where

$$y_i \sim \begin{cases} G_1 & \text{for } \tau_0 < i \leq \tau_1 \ , \\ G_2 & \text{for } \tau_1 < i \leq \tau_2 \ , \\ \vdots & \vdots \\ G_{m+1} & \text{for } \tau_m < i \leq \tau_{m+1} \ . \end{cases} \tag{2.1.1}$$

Here $G_1 \ldots G_{m+1}$ are the data generating processes associated with each segment such that $G_k \nsim G_{k+1}$ for $k \in \{1, \ldots, m\}$. Note $T = \varnothing$ represents the case where there are no changepoints in the data and the whole sequence $y_{1:n}$ is stationary and follows the same data generating process, G_1.

The goal of changepoint detection is to identify the best possible segmentation of the data, T. The best segmentation will split the data into stationary segments that follow the same data generating process. Hence, we can define a criterion function, $V(T)$, as

$$V(T) = \mathcal{C}\left(y_{1:n}|T\right) \ , \tag{2.1.2}$$

where $\mathcal{C}(\cdot|T)$ is a measure of the homogeneity of the data given the changepoint locations. Changepoint methods aim to minimize $V(T)$ as much as possible, subject to constraints on m. The formulation of $V(T)$ in (2.1.2) leads to the three main properties that make up the majority of changepoint methods:

1. **Cost Function/Test Statistic:** There are two main ways of defining this property. Firstly, there are model-based methods that use cost functions where

we assume that the total cost of the data is the sum of the cost of the individual segments. We define this as

$$V'(T) = \mathcal{C}(y_{1:n}|T) = \sum_{k=0}^{m} \mathcal{C}(y_{(\tau_k+1):\tau_{k+1}}) \tag{2.1.3}$$

and here we assume independence between the segments. In this setting, the cost function measures the homogeneity of the data within the segment - the lower the cost of a subset of data, the less likely it is to contain a changepoint. This method is usually used when fitting a model to the data; a common cost function is the negative maximum log-likelihood. Note throughout $V'(T)$ represents the minimization where we assume the total cost is the sum of the individual segments and $V(T)$ is the more general minimization of the cost with respect to the changepoints.

Secondly, many methods assume there is At Most One Changepoint (AMOC) in a set of data and they use a test statistic based method to determine whether the data contains a change or not. Hence, they seek to minimize $V(T)$ by finding the location of a single change that minimizes the cost function/test statistic. Note in some methods the aim is to maximize the test statistic, however, this can easily be re-framed as a minimization problem by taking the negative of the test statistic.

2. **Search Method:** Search methods tell us how to locate the changepoints given $\mathcal{C}(\cdot)$. For the model-based methods, we could search all 2^{n-1} possible segmentations to find the one that minimizes (2.1.3). However, this is rarely feasible and, hence, more efficient search methods are required. For the test statistic

based methods, where we assume AMOC, then the location of the change is usually the point that minimizes the test statistic. We then require a method for extending these methods to detect multiple changepoints but note these will only give an approximate solution to (2.1.2).

3. **Penalization:** The number of changepoints, m, is rarely known in practice. Hence, to avoid over-fitting we need to penalize the number of changepoints (or other aspects of the model) in some way. This is often viewed as a model selection problem where we need to choose the model with an appropriate number of changes.

We explore each of these three properties in more detail before giving a short overview of some methods that fall outside of this framework.

2.1.1 Cost Function/Test Statistic

First, we examine the cost functions and test statistics that are common in the univariate changepoint literature. We start by looking at model-based methods and different cost functions that are used within (2.1.3) before looking at different test statistics that are used to identify single changes in test statistic based methods.

Model Based Methods

If we can assume some knowledge of the data generating process then a common choice of the cost function is the negative of the maximum log-likelihood of the data.

Here it is common for the data generating process, G_k, to follow the same distribution and just its parameters can undergo a change. Common examples of these models range from mean and/or variance changes in Gaussian data (Page and E. S., 1955; Krishnaiah and Miao, 1988; Lavielle and Moulines, 2000; Pein et al., 2017) to rate changes in Poisson distributed data (Ko et al., 2015). These models generally assume the data $y_{1:n}$ are i.i.d given the m changepoints.

For a family of distributions with density $g(\cdot|\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ are the parameters that can undergo changes, we define the cost function for a subset of the data $y_{s:e}$ as

$$\mathcal{C}(y_{s:e}) = -2 \max_{\boldsymbol{\theta}} \sum_{i=s}^{e} \log g(y_i|\boldsymbol{\theta}) \ .$$

Despite the likelihood-based cost function covering a wide range of models, knowledge of the underlying data model is required and misspecifications can lead to poor performance of methods that employ this cost function.

Several alternative cost functions fall into the model-based methods category but don't use a classical likelihood function. These include cost functions for piece-wise linear models (Bai, 1994, 1996); data that exhibits temporal dependence (Bai, 2000; Chakar et al., 2017); and non-parametric maximum likelihood (Zou et al., 2014; Haynes et al., 2017b).

Test Statistic Based Methods

An alternative class of changepoint methods utilize a test statistic to determine the presence of a changepoint or not. Here the principle is similar to the model-based

methods, however, the data isn't explicitly modelled using the cost function. Instead, the cost function acts as a test statistic for determining if the data contains a change or not. These test statistics are usually designed for detecting certain types of changes, such as mean changes, however, usually have weaker modelling assumptions compared to model-based methods.

The most common type of change to detect is a mean change and in this setting the most common choice of test statistic is the CUSUM statistic, which we define

$$\mathcal{C}(y_{s:e}|t) = \sqrt{\frac{e-t}{(e-s+1)(t-s+1)}}\sum_{i=s}^{t} y_i - \sqrt{\frac{t-s+1}{(e-s+1)(e-t)}}\sum_{i=t+1}^{e} y_i \,. \quad (2.1.4)$$

The maximum of this test statistic across potential changepoint locations, $s < t < e$, will yield the most likely changepoint position and the size of the test statistic determines how likely it is that the maximum point is a changepoint.

The CUSUM test statistic has been used since at least Hinkley (1971) for mean changes in Gaussian data and builds upon the similar procedure given in Page (1954). The CUSUM statistic has been adapted to other settings such as for changes in variance (Inclan and Tiao, 1994); non-parametric i.i.d setting (Csörgö and Horváth, 1988); and temporally dependent data (Robbins et al., 2011).

Although the CUSUM statistic is a popular choice within changepoint detection it does have some drawbacks. Firstly, it assumes there is AMOC in the data, this can cause issues (specifically masking) when data contains multiple changepoints, something that the model-based cost functions avoid. Masking is discussed further in Section 2.1.2. Secondly, the CUSUM statistic is known to perform poorly at the

boundaries of the data (Csörgö and Horváth, 1997) where it can yield high test statistic values simply due to the sample size on the relevant side of the change. However, this issue is also present when using cost-based methods and a general solution is to set minimum segments length between changepoints (including the boundaries of the data).

Besides the model-based methods and test statistic based methods there are a few other examples of cost functions. Lévy-Leduc and Roueff (2009) and Lung-Yut-Fong et al. (2011) use U-statistic based cost functions which are based on the ranks of the data. These methods don't make modelling assumptions on the data but can have poor power for detecting changepoints. Harchaoui and Cappe (2007) introduced a kernel-based cost function and this idea was developed further in Celisse et al. (2018). Kernel-based cost functions can work well for detecting changes when the data generating processes have more complex and non-linear changes such as the half-moon data example given in Harchaoui and Cappe (2007). Finally, Matteson and James (2014) proposed two non-parametric methods, E-Divisive and ecp3o, that use a divergence measure to identify changepoints. They implement this cost function using both a model (ecp30) and a test statistic (E-Divisive) style approach. These methods have the advantage of detecting distributional changes rather than parameters specific changes, however, these cost functions can be computationally expensive to calculate.

In general, model-based methods work well in scenarios where the data generating process is known and we are looking for a change in a subset of parameters. In partic-

ular, if we are looking for changes in multiple parameters, such as a mean and variance change, then model-based cost functions incorporate this whereas test statistic based cost functions tend to focus on a single specific change type. If we are only interested in a specific type of change and we are less sure about the modelling assumptions of the data then test statistic based methods are typically more robust to model mis-specifications or don't require a specified model at all. Moreover, these methods are generally better suited to data that exhibits temporal dependence, especially if we don't model this dependence in the model-based cost functions. Finally, if we are unsure of the type of change and data generating process then a method such as in Matteson and James (2014) makes much milder assumptions on the change and the data may be the most appropriate.

2.1.2 Search Methods

In Section 2.1.1, we presented a range of different cost functions for measuring how likely a subset of data is to contain a change. Now we examine how to search for multiple changes given these cost functions. Search methods can generally be split into two categories. Firstly, there are optimal methods (otherwise known as exact methods) where we aim to minimize $V'(T)$ exactly given some penalization on the number of changes or a known number of changes. These optimal search methods are usually used alongside the model-based cost functions introduced in Section 2.1.1. Secondly, there are approximate approaches. These generally detect changes one at a time and keep adding changes (dependent on the previously identified changepoints)

until the required number of changes is reached or the penalization deems no more changes should be added. These approximate methods typically partner with the test statistic based cost functions.

Optimal Search Methods

The most naive approach to finding the exact solution to $V'(T)$, would be to search all possible segmentations. However, with m known this would yield $\binom{n-1}{m-1}$ possible segmentations while with m unknown this would yield $\sum_{m=1}^{n-1}\binom{n-1}{m-1}$ segmentations. In both settings, unless m and n are small, searching over all these segmentations is impractical. Hence, more efficient ways of finding the exact solutions are required.

Firstly, for these search methods to work, we need to penalize, otherwise, the majority of methods would keep adding changepoints up to $m = n - 1$. Hence, we seek to find a solution to the optimization problem

$$\min_{T} V'(T) + \beta f(m) \tag{2.1.5}$$

where $\beta f(m)$ is a penalty to guard against over-fitting changepoints. Note that $m = |T|$, hence $\beta f(m)$ is relevant to the optimization. The choice of penalization is explored more thoroughly in Section 2.1.3.

If an upper limit of the number of changepoints, say M, can be set, then Auger and Lawrence (1989) proposed the Segment Neighborhood (SN) approach for solving (2.1.5). This approach utilizes dynamic programming to search the entire segmentation space and return the global minimum. However, this method still has a signifi-

cant computational cost, $\mathcal{O}(Mn^2)$ and if the number of changepoints increases linearly with the number of data points this results in computational complexity of $\mathcal{O}(n^3)$. To help reduce this computational cost, Maidstone et al. (2017) applied inequality based pruning to the Segment Neighborhood method. However, pruning is not guaranteed to reduce the computational complexity especially in situations where the number of time points increases but the number of changepoints stays constant. Hence, in many situations Segment Neighborhood approaches with or without pruning are still computationally intractable for large time series.

To reduce this computational cost, further assumptions are required on the penalty term. If $f(m) = m$, then Jackson et al. (2005) proposed a search method to minimize (2.1.5) called the Optimal Partitioning (OP) method. This method sequentially moves through the data and works by conditioning on the last point of change. More formally, for $t' < t$, let $F(t')$ be the minimization of (2.1.5) up to time point t'. If $F(0) = -\beta$, then

$$F(t) = \min_{t'<t} \left[F(t') + C(y_{(t'+1):t}) + \beta\right] \ . \tag{2.1.6}$$

This recursion says, at time point t consider all previous time points t', which of these is the optimal most recent changepoint i.e. $F(t') + C(y_{(t'+1):t}) + \beta$ is smaller than any other $t' < t$.

Consider Figure 2.1.1, if we are at time t then we need to look back at all previous time points to find t' that minimizes (2.1.6). If we look at time point t'_1, then $C(y_{t'_1+1):t})$ will be large as it contains a changepoint, hence this is unlikely to minimize (2.1.6). Next, consider time point t'_2, here $C(y_{t'_2+1:t})$ will be small as it contains no changepoints.

So t'_2 would be a good candidate to minimize (2.1.6). Finally, consider t'_3, again $C(y_{t'_3+1:t})$ will be small, however, the optimal segmentation $F(t'_3)$ will likely contain the changepoint at t'_2, hence will contain an extra penalty term β in the overall cost. Thus $F(t'_3) + C(y_{t'_3+1:t}) + \beta$ will be likely be larger than $F(t'_2) + C(y_{t'_2+1:t}) + \beta$ as the difference between $C(y_{t'_3+1:t})$ and $C(y_{t'_2+1:t})$ will be smaller than the cost of additional penalty term included in $F(t'_3)$.

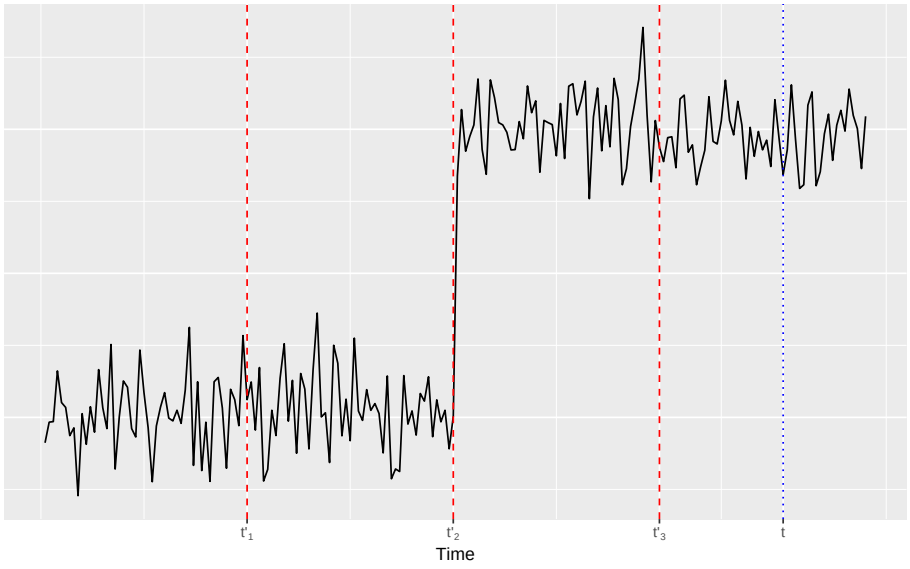


Figure 2.1.1: Candidate changepoints for the explanation of OP and PELT

More recently, efforts have been made to further decrease the computational complexity of optimal search methods via pruning. Using OP, at each time point every previous time point is checked to see if it is the optimal most recent changepoint. Pruning aims to reduce the number of time points that need to be checked in the minimization of (2.1.6) by removing those that can't be the optimal most recent changepoint. Using Figure 2.1.1, the aim of pruning would be to remove time point t'_1 from the list of time points being checked since it is clearly not the optimal most

recent change.

Killick et al. (2012) proposed the Pruned Exact Linear Time (PELT) method which applies inequality based pruning to the OP method and utilizes the observation that introducing a changepoint reduces the cost of the sequence. If for $t_2' > t_1'$,

$$F(t_1') + C(y_{(t_1'+1):t_2'}) + \beta \geq F(t_2'), \tag{2.1.7}$$

then for any $t > t_2'$, t_1' can't be the most recent changepoint as t_2' is a more optimal candidate. Consider $F(t_2') + C(y_{t_2'+1:t}) + \beta$, this will always be smaller than $F(t_1') + C(y_{t_1'+1:t}) + \beta$ due to the inequality in (2.1.7) and the assumption,

$$C(y_{(t_1'+1):t_2'}) + C(y_{(t_2'+1):t}) + \beta \leq C(y_{(t_1'+1):t}) \ .$$

Hence, t_1' can never be the optimal solution to the minimization in (2.1.6) for the current time point t or any future time points. Hence, PELT will prune this time point and it will no longer be a candidate for the optimal most recent changepoint. This results in substantial computational gains over OP and when the number of changes grows linearly with n the computational cost of PELT is $\mathcal{O}(n)$. However, pruning is not guaranteed especially if the number of changes remains fixed as $n \to \infty$, hence the method has a worst-case computational cost of $\mathcal{O}(n^2)$. Despite this PELT has become one of the main methods used for univariate changepoint detection in a wide range of applications.

Functional pruning has also been used to improve the speed of optimal search methods. Maidstone et al. (2017) proposed the Functional Pruning Optimal Partitioning (FPOP) method which can drastically improve the computational speed of the OP

method even further than PELT. Despite the vastly improved computational time achieved with functional pruning, these methods require many functions to be stored as the method runs. This means that only one parameter can feasibly be checked for a change at any given time. Hence, when searching for changes in two or more parameters inequality based pruning would likely be faster.

In summary, for detecting changes in more than one parameter then the PELT method would likely be the most appropriate. If you are only looking for a change in one parameter then using the FPOP method would likely give an improved computational saving. However, if you wish to use a more complex penalty than $f(m) = m$ then SN based methods may be more appropriate. More details on the choice of penalties are given in Section 2.1.3.

Approximate Search Methods

Using optimal search methods can be computationally expensive depending on the problem setup and therefore approximate search methods are commonly used. Approximate search methods usually involve searching for a single changepoint in a subset of data (commonly using test statistic based cost functions) and if a changepoint is found then further individual changepoints are identified dependent on the detected changes.

In Section 2.1.1, we saw how test statistic based cost functions aim to determine if a changepoint is present in a subset of data or not. Approximate search methods are commonly used to extend these methods to detect multiple changes. To determine

if a test statistic is significant or not we can compare it to a threshold (or penalty), β. These thresholds/penalties will be discussed further in Section 2.1.3. If the test statistic exceeds the threshold then we deem a changepoint to have occurred and we place it at the time point which maximizes the test statistic. Different approximate search methods can then be used to identify additional changepoint locations if $m > 1$. Alternatively, the approximate search method can detect up to a pre-defined maximum number of changepoints, M, and then model selection style penalties can be used to choose the most appropriate number of changepoints. More details on this are given in Section 2.1.3.

Binary Segmentation (BS) is 'arguably the most established search method used within the changepoint literature' (Killick et al., 2012). This method dates back to Scott and Knott (1974) and is the basis of the majority of approximate search methods. BS works by using the chosen test statistic to identify a single changepoint (if present). If a changepoint is present the data is split at the position of the change and the two segments on either side of the change are searched for a single changepoint using the given test statistic. This process is repeated until all segments separated by the detected changes are deemed to not contain a change or M has been reached.

BS, usually in combination with the CUSUM test statistic, has been a popular method for many years. This is likely due to its ease of implementation and strong theoretical results (Venkatraman, 1992; Chen et al., 2011; Cho and Fryzlewicz, 2012; Fryzlewicz, 2014). Despite this, BS does have some drawbacks. The main issue is masking

(Padmore, 1993), where the presence of multiple changepoints causes the failure of the test statistic to detect any changes or can cause a changepoint to be placed in an incorrect location. This becomes a serious consideration in situations where there is an abundance of changepoints as the chance of masking increases.

In an attempt to overcome the issue of masking, alternative variations of BS have been suggested. The most well-known example is Wild Binary Segmentation (WBS) created by Fryzlewicz (2014). This method works by randomly drawing I intervals across the data and performing the tests solely on these intervals. The premise behind this approach is that the probability that there is an interval with exactly one changepoint is high and therefore masking won't occur in this interval. While this method does help overcome the issue of masking, many intervals are usually required in practice which can make the method computationally intensive. Moreover, when used in practice this method is known to over-fit the number of changepoints due to testing across many intervals - this issue can be overcome by selecting a more appropriate penalty.

Further alternative extensions to BS include Circular Binary Segmentation (Olshen et al., 2004; Venkatraman and Olshen, 2007) where the data is assumed to go through regimes and usually returns to a common null level. More recently the Narrowest-Over-Threshold approach of Baranowski et al. (2019) was proposed which aims to improve upon WBS by selecting the change from the smallest interval I subject to it exceeding the threshold. This further reduces the effect masking and multiple changepoints within an interval, however, the choice of the intervals I is still an issue.

These recent additions indicate that developing extensions to BS is still an active research area.

Windowed methods are an alternative to BS and can be categorized as approximate search methods. Here a rolling window is passed over the data and the test statistic is calculated within each window to detect the presence of a change. The main windowed procedure in the literature is based on MOSUM statistics (Hušková and Slaby, 2001; Eichinger and Kirch, 2018). The main drawback with these methods is their dependence on a user-specified bandwidth parameter for the size of the window. Cho and Kirch (2021) suggested using different bandwidths for detecting different change sizes, however, this can result in a computationally expensive method if many bandwidths are required along with multiple testing issues for detecting changepoints across the different bandwidths.

The choice of using an approximate search method or an exact search method usually is dependent upon the choice of the cost function. Typically, model-based cost functions will use an exact search method unless the series is extremely long in which case the faster approximate methods may be more appropriate. However, FPOP has been shown to have a computational time that is similar, if not faster, than many approximate methods especially when there are many changes in the data. When using a test statistic based cost function then the approximate methods are generally the only way to extend these for detecting multiple changepoints.

2.1.3 Penalization

So far we have considered choosing an appropriate cost function/test statistic and how to search for changepoints using either optimal or approximate search methods. Within the optimal search methods, we noted that a penalty term, $\beta f(m)$ needs to be included with minimization of $V'(T)$ to control the number of changepoints. Moreover, for the approximate search methods either a threshold β is needed to deem a changepoint significant or a model selection style penalization is required to determine the appropriate number of changes. Here we explore the choice of penalty in more detail.

The main aim of penalization is to control the false-positive rate of detecting changes when none are present in the data. Within optimal search methods the choice of $\beta f(m)$ in (2.1.5) is usually based on information criteria. For approximate techniques, the choice of the threshold, β, for determining if a change is significant is commonly based on the limiting distribution of the test statistic either theoretically or evaluated through simulating many repetitions of null data. Here we explore these three methods for choosing an appropriate penalty in more detail.

Information Criterion Based Penalization

We saw in Section 2.1.2, that for some search methods the choice of penalty has to be set as $f(m) = m$, such as in Optimal Partitioning (Jackson et al., 2005) and PELT (Killick et al., 2012). This assumption is stricter than necessary but is a common choice. In reality, as long as $\beta f(m) = m\beta + f(y)$ then Optimal Partitioning and

subsequent pruning can be used.

One common choice of penalization is to set $f(m) = m$ and use the Bayesian Information Criterion (BIC), also known as Schwarz Information Criterion (SIC), introduced by Schwarz (1978). In the setting where a single parameter is allowed to change then, the BIC penalty is $\beta = 2 \log n$. Note that we have at least two new parameters upon adding a changepoint; the estimated change location and a new parameter estimate. Dependent on the number of parameters allowed to change and that are estimated, the BIC penalty is adjusted accordingly. In the Gaussian mean change setting, Yao (1988) developed a theoretical justification for using the BIC penalty.

Despite its popularity, the BIC penalty can have difficulties in the large sample, many changepoints and non-Gaussian settings, hence Zhang and Siegmund (2007) developed a Modified Bayesian Information Criterion (MBIC) to overcome these issues. This penalty is defined as

$$\beta f(m) = (2m - 1) \log n + \sum_{k=1}^{m+1} \log \left(\tau_k / n - \tau_{k-1} / n \right) ,$$

and can be written in the form $\beta f(m) = m\beta + f(y)$, hence can be used within Optimal Partitioning. This penalty has gained increasing popularity as it is less sensitive to model misspecification which is common when using model-based cost functions.

Alternative, information criteria such as the Akaike Information Criterion (AIC) proposed by Akaike (1974) are rarely used in practice due to poor performance (Haynes et al., 2017a; Jones and Dey, 1995; Reeves et al., 2007) and as it is not asymptotically consistent (Yao, 1988).

Other forms of penalties have been suggested for model-based approaches. Lebarbier (2005) suggested a penalty of the form

$$\beta f(m) = \frac{m}{n}\sigma^2 \left(c_1 \log \frac{n}{m} + c_2 \right) ,$$

for the Gaussian mean change setting. However, this penalty requires the tuning of c_1 and c_2 and is limited to the Gaussian mean change setting. Tibshirani and Wang (2008) have examined penalties based on the Least Absolute Shrinkage Selection Operator (LASSO) and Harchaoui and Lévy-Leduc (2010) note the similarity between the LASSO and detecting changes in mean in the Gaussian setting. However, recently it has been noted that such penalties typically struggle to detect the true number of changes (Ng et al., 2018).

Information Criterion can also be used within the approximate search methods. If we set a maximum number of changepoints then we can produce a segmentation for each number of changepoints $m = 1, \ldots, M$ sequentially. Then, choosing the appropriate number of changepoints becomes a model selection problem and we can use Information Criterion. In this setting, Fryzlewicz (2014) introduced the Strengthened Schwarz Information Criterion (sSIC) for choosing the appropriate number of changepoints after using Wild Binary Segmentation for detecting mean changes. This criterion has also been used in Baranowski et al. (2019) and works by defining a maximum likelihood estimator of the residual variance,

$$\hat{\sigma}^2_m = n^{-1} \sum_{k=1}^{m} \sum_{i=\tau_{k-1}+1}^{\tau_k} \left(y_i - \bar{y}_k \right)^2 ,$$

where $\bar{y}_k$ is the mean estimate in segment k. The sSIC is then defined as

$$\text{sSIC}(m) = \frac{n}{2} \log \hat{\sigma}_m^2 + m \log^\alpha n \ .$$

Note if $\alpha = 1$ this corresponds to the standard BIC penalty. The downside to this penalty is that the arbitrary choice of α and this penalty can suffer similarly to the related BIC penalty with frequent changes.

A similar idea for optimal search methods was developed by Haynes et al. (2017a). They proposed the CROPS algorithm which produces a solution path for all choices of penalties across a continuous range. This allows for multiple segmentations to be evaluated and therefore a suitable penalty to be chosen by considering a scree plot of the total cost $V'(T)$. The aim is to find the elbow of the scree plot such that introducing a change does not drastically decrease the total cost of the segmentation. An example of using the CROPS algorithm can be seen in Appendix A.9.

Limiting Distribution Based Penalization

For detecting single changepoints it is common to use the limiting distribution of the test statistic under the null hypothesis of no change to set an appropriate threshold penalty. The threshold is then chosen to control the false alarm rate under the scenario of no change. Here the difficult task is to determine the limiting distribution of the test statistic.

For an overview of calculating the limit distributions for some common AMOC changepoint methods see Csörgö and Horváth (1997). Generally, the convergence of these

test statistics to their limit distributions is slow especially if the data is non-Gaussian. Furthermore, even when the limiting distribution has been established it often does not have an analytical form and hence sampling from the distribution is required. In practice, thresholds based on limiting distributions are often overly conservative in finite sample settings and simulation-based penalization is used instead.

Multiple changepoints also cause a problem when using limiting distributions. When testing for a single change we can use the $(1-\alpha)$-percentile of the limiting distribution as the threshold (where α is the user-defined false alarm rate). However, if we are searching for more changes this raises the issue of multiple testing. If we keep testing for multiple changepoints it becomes more likely that we incorrectly identify a changepoint where one does not exist. This makes controlling the overall false alarm rate a challenging task as if we increase the threshold to account for multiple changepoints and control the overall false alarm rate then we are more likely to miss true changepoints.

Simulation Based Penalization

To overcome the overly conservative thresholds based upon limiting distributions, we can instead use a simulated threshold. To do this, we simulate from many repetitions of the null model, run the test statistic on these null data examples and this will give us samples from the distribution of the test statistic under the no-change scenario. We use these samples to set a threshold that controls the false positive rate at the desired level, however, the issue here is the required knowledge of the null model.

Moreover, these penalties have the same issues with multiple changepoints as seen with the thresholds based on limiting distributions. Despite this, simulated thresholds are commonly used in the implementation of methods even if a theoretical threshold based on the limiting distribution of the test statistic has been found, see Gallagher et al. (2013), Wang and Samworth (2018) and more.

Instead of simulating from the null model, we could instead utilize the fact that under the scenario of no change the ordering of the data is not important, assuming independence between time points. Hence, we could permute the data points multiple times and by running the test statistic on these permuted data sets we would get approximate samples from the distribution of the test statistic under the null model. Note if a change was present in the data then the points from either side of the change would become mixed therefore removing the change from the data and hence the test statistic would not be as high as with the original data with the changepoint present. This idea of using a permutation test was used in Matteson and James (2014) where it was used in conjunction with BS.

The choice of penalty is one of the most challenging aspects of changepoint analysis. For model-based cost functions, the use of information criteria is very common with the MBIC and BIC being the most popular. Despite their popularity, theoretical guarantees (on the number of changes and their location) are less common with model-based cost functions and generally strong assumptions are placed on the data generating process and change type. Moreover only recently have these theoretical guarantees been extended to pruning methods such as PELT (Tickle et al., 2020). On

the other hand, the theoretical properties of the test statistic methods and subsequent approximate approaches have been well studied. Here the issue lies in the practicality and despite theoretical guarantees being presented for many of the methods, they usually come with the caveat that using a data generated or simulation-based penalty would work better in practice.

In examining different cost functions, search methods and penalties, we have presented a wide range of methods for solving univariate changepoint problems. In Section 2.1.4, we briefly highlight a few methods that do not readily fit into this framework or come from a frequentist viewpoint.

2.1.4 Alternative Methods

A major contribution to the univariate changepoint literature is the SMUCE method of Frick et al. (2014). This method utilizes a multi-scale test statistic to search over the entire space of discrete step functions in exponential-family generated data. This method along with extensions including H-SMUCE (Pein et al., 2017) have the advantage of providing natural confidence intervals for the changepoint estimates, something that is rare in the frequentist literature. Despite this, these methods require the choice of several tuning parameters which can have drastic effects on the analysis and the power of these methods is sub-par in comparison to other state-of-the-art methods.

Throughout this thesis we focus on the frequentist setting, however, there has been some popularity within Bayesian methods for changepoint detection. Fearnhead

(2006) introduced the perfect simulation procedure which updates its posterior similarly to PELT. Another popular concept is to use hidden Markov models, where the hidden states represent the different segments. This idea has seen numerous works including Chib (1998), Luong et al. (2013), Ko et al. (2015). These models can also be referred to as jump models (Bemporad et al., 2018; Nystrup et al., 2020, 2021). These models can be particularly useful when there are a set number of regimes and at changepoints, the model switches to one regime or another. Moreover, Bayesian methods can naturally provide credible intervals for the changepoint locations something which is still ongoing research in the frequentist setting.

2.2 Multivariate Changepoints

Historically, changepoint detection focused on the univariate setting, however, recently the multivariate setting has received substantially more attention. If we let the number of data streams/variates be p, we have data $y_{1:n,1:p} = (y_{1,1:p}, y_{2,1:p}, \ldots, y_{n,1:p})$ where $y_{i,1:p} \in \mathbb{R}^p$ for $1 \leq i \leq n$. Similarly, to (2.1.1), we now have

$$y_{i,j} \sim \begin{cases} G_{1,j} & \text{for } \tau_0 < i \leq \tau_1 \ , \\ G_{2,j} & \text{for } \tau_1 < i \leq \tau_2 \ , \\ \vdots & \vdots \\ G_{m+1,j} & \text{for } \tau_m < i \leq \tau_{m+1} \ . \end{cases} \tag{2.2.1}$$

Again, $G_{1,j}, \ldots, G_{m+1,j}$ are the data generating processes associated with each segment for a given variate j. Note that the $G_{k,j}$ need not be the same distributional

form for all j. In the multivariate setting, not all the data generating processes across the different variates need to change for a changepoint to be present. Hence, we define $\mathcal{S}_k \subseteq \{1, \ldots, p\}$ as the non-empty set of variates that undergo a change at τ_k. This means that at a changepoint, $G_{k,j} \nsim G_{k+1,j}$ if and only if $j \in \mathcal{S}_k$.

Changepoints only occurring in a subset of the variates is one of the key challenges associated with multivariate changepoint detection. We refer to the number of variates that undergo a change as the sparsity of the changepoint and we use this as a way to divide this Section. Firstly, we will consider the *fully multivariate changepoint* setting where we assume $\mathcal{S}_k = \{1, \ldots, p\}$ i.e, the changepoint occurs in all variates. Here we consider extensions of univariate methods to the multivariate paradigm along with a thorough review of covariance changepoints where the dependence structure between the data generating processes can change. This leads to Chapter 4 where a special type of covariance change, a subspace changepoint, is considered. Secondly, we consider the *sparse changepoint* setting where just a subset of the variates may change. Here dimension reduction or aggregation techniques are commonly used to account for the sparsity of the changepoints. We use our own dimension reduction technique in Chapter 3 albeit to solve an open problem in the fully multivariate setting. Note that we divide fully multivariate and sparse changepoints based upon the cost function used and not the search method. For example, many methods that use AMOC test statistics for detecting sparse changes, extend to multiple changepoints by assuming the change occurred in all series and hence still use a BS style approach despite not all the variates changing.

The main differences between univariate and multivariate methods revolve around the cost function or test statistic. Many of the search methods and penalization methods intuitively carry forward from the univariate to the multivariate setting. For example, with test statistic based methods, BS is still a popular way to detect multiple changepoints and using the limit distribution of these test statistics is a common way to obtain a threshold. Hence, this Section puts increased focus on the different cost functions used in the multivariate setting and for a thorough critique of different search methods and penalization see Sections 2.1.2 and 2.1.3.

2.2.1 Fully Multivariate Changepoints

The fully multivariate changepoint setting was first seen in Srivastava and Worsley (1986) who proposed a multivariate likelihood ratio test for detecting mean changes in multivariate Gaussian data. Since then, the extension of different univariate cost functions to the multivariate domain has occurred, along with the introduction of changes in the covariance structure between variates. We first examine extensions of model-based cost functions to the fully multivariate domain before examining how test statistic style cost functions (mainly based on the CUSUM statistic) have been extended to detect fully multivariate mean changes. Then we discuss some recent methods for detecting covariance changepoints and highlight the open research area discussed in Chapter 4. Finally, we mention a few alternative non-parametric and Bayesian methods for completeness.

Model Based Cost Functions

The extension of model-based cost functions to the fully multivariate domain is intuitively a simple task. Instead of modelling the data from a univariate data generating process and using the negative log-likelihood, we can simply model the data from a multivariate data generating process. Our cost function can remain as the negative log-likelihood of the multivariate data and changes can be detected accordingly using an appropriate univariate search method (e.g. SN, OP or PELT). As we have additional parameters to estimate, the commonly used information criterion for penalization has to be adjusted slightly so that it scales with n, p and the number of changepoints.

Although this extension of model-based cost functions is intuitively simple, it has many issues. Firstly, we saw in Section 2.1.1 that these approaches, when paired with an optimal search method, can have a large computational cost; this issue persists in the multivariate domain. Lavielle and Teyssière (2006) and Maboudou and Hawkins (2009) propose methods that are based on the SN approach and in the multivariate domain this has computational cost $\mathcal{O}(Mpn^2)$ so even for a small number of variates this can quickly become intractable to compute. Moreover, in the multivariate domain, the minimum segment length between changepoints can grow much larger for certain change types i.e, if a covariance matrix needs to be estimated this makes the minimum segment length at least p when using a traditional covariance estimator. Hence, if $p > n/2$ then we can not fit any changepoints to the data. Hallac et al. (2019) propose a method to overcome this issue, they use a regularization parameter

to aid in the estimation of the covariance matrix when p is large and look for changes in the mean and covariance of Gaussian data. This method can be used with optimal search methods (which is computationally inefficient) or they recommend using an approximate approach by estimating changepoints using a BS style search method. This is one of the few methods that can detect multiple changes while remaining computationally efficient, however, the method does not come with any theoretical guarantees regarding the detection of changepoints or their locations.

Model-based approaches excel when there is potential for multiple parameters to change at once and the data generating process is known. However, due to the large computational cost of model-based methods in the multivariate domain, there is a lack of suitable methods for detecting changes in multiple parameters in a potentially high-dimensional setting. Hence, in Chapter 3, we propose a method that uses a dimension reduction technique to detect changes in mean and covariance simultaneously in a computationally efficient manner.

Test Statistic Based Methods

As model-based cost functions have a high computational cost, test statistic based approaches have become increasingly popular in the multivariate domain, in particular multivariate extensions of the CUSUM statistic (Zamba and Hawkins, 2006, 2009; Wang and Reynolds, 2013; Tartakovsky et al., 2014; Enikeeva and Harchaoui, 2019). These methods work by applying the CUSUM statistic defined in (2.1.4) to each variate and then aggregating these CUSUM statistics in some way.

Let $S_{t,j}$ be the CUSUM statistic from (2.1.4) for variate j at time point t, so

$$S_{t,j} = \sqrt{\frac{e-t}{(e-s+1)(t-s+1)}} \sum_{i=s}^{t} y_{i,j} - \sqrt{\frac{t-s+1}{(e-s+1)(e-t)}} \sum_{i=t+1}^{e} y_{i,j} \ , \qquad (2.2.2)$$

based on an interval $s \leq t \leq e$. An obvious test statistic for detecting a fully multivariate change would involve the l_2-norm of the test statistics, $S_{t,1:p}$, at each potential changepoint location t. Here we examine test statistics where the CUSUM statistic from each variate is used in the final test statistic, note there are different aggregations designed for sparse changepoints but these are discussed in Section 2.2.2. Zhang et al. (2010) introduced a test statistic that uses the sum of squared CUSUM statistics across all the variates, with

$$\mathcal{C}(y_{1:n,1:p}) = \max_{1 \leq t \leq n} \sum_{j=1}^{p} \left(S_{t,j}\right)^2 \ . \qquad (2.2.3)$$

They derived the test criterion for a theoretical threshold under the i.i.d Gaussian setting and implement this in a circular BS framework (Olshen et al., 2004). Hence, this is suited to data that returns to a known base level sometime after a change has occurred. Furthermore, Horváth and Hušková (2012) proposed a similar test statistic

$$\mathcal{C}(y_{1:n,1:p}) = \max_{1 \leq t \leq n} \frac{1}{\sqrt{p}} \frac{t(n-t)}{n^2} \sum_{j=1}^{p} \left[(S_{t,j}) - 1\right]$$

and the limit distribution derived for independent variates. Whilst these methods have well studied theoretical properties and perform well in the Gaussian change in mean setting, the issues from the univariate setting remain. Namely, the convergence of these test statistics to their limiting distributions is slow especially for non-Gaussian and dependent data, thus setting a threshold for the test statistics is

challenging. Moreover, these methods are typically used within a BS or WBS setup to detect multiple changepoints. This raises the multiple testing issue when using limiting distribution based thresholds. On the plus side, these methods are more computationally efficient and easier to implement than the model-based cost function approaches. Moreover, issues regarding thresholds can be overcome by using simulated or data-driven thresholds.

Covariance Changepoints

Covariance changepoints arise when the covariance between the data generating processes $G_{i,1:p}$ changes at the unknown changepoints τ_k. Lavielle and Teyssière (2006) introduce a model-based cost function for detecting covariance changepoints but this has issues regarding the estimation of the covariance within the cost function, particularly in the case where p is large. Hence, Aue et al. (2009a) propose a CUSUM statistic based method to detect covariance changes. This method detects mean changes in the stacked covariance vector. Define vech$(\cdot)$ to be the operator that stacks the columns below the diagonal of a symmetric matrix as a vector with $p(p+1)/2$ entries. Using this they define S_t as

$$S_t = \frac{1}{\sqrt{n}} \left(\sum_{i=1}^{t} \text{vech}\left(y_{i,1:p} y_{i,1:p}^T\right) - \frac{t}{n} \sum_{i=1}^{n} \text{vech}\left(y_{i,1:p} y_{i,1:p}^T\right) \right) ,$$

for $t = 1, \ldots, n$ where y^T is the transpose of the vector y. Two test statistics are proposed for detecting changes,

$$\mathcal{C}_1(y_{1:n,1:p}) = \max_{1 \leq t \leq n} S_t^T \hat{\Sigma}_n^{-1} S_t \ , \tag{2.2.4}$$

$$\mathcal{C}_2(y_{1:n,1:p}) = \frac{1}{n} \sum_{t=1}^{n} S_t^T \hat{\Sigma}_n^{-1} S_t \ , \tag{2.2.5}$$

where $\hat{\Sigma}_n^{-1}$ is a long-run covariance estimator of the vectors, $\text{vech}(y_{i,1:p} y_{i,1:p}^T)$. Aue et al. (2009a) show that using (2.2.5) has more power for detecting changepoints and the limit distribution of this test statistic is well established with reasonable convergence. However, this test statistic does not identify the location of the changepoint since the cost function is made up of evidence for a changepoint at every time point, hence (2.2.4) is needed to identify the changepoint location. This method is non-parametric and has been shown to work well in a number of economic models including Multivariate ARMA and linear processes as well as many GARCH models. However, the estimation of the long-run covariance is non-trivial and these test statistics are sensitive to its estimation. Furthermore, as this method stacks the covariance matrix estimates this method does not scale well with increasing p, hence it is only recommended for a small number of variates.

To extend this method to high-dimensional problems, Dette et al. (2020) have recently proposed an alternative version of S_t. This method is technically a sparse changepoint method, however, we include it here for completeness of the covering covariance changepoints. They propose only including the variates in the stacked covariance vector where the difference between pre and post-change values is large. A similar idea is shown in (2.2.8) in Section 2.2.2. This reduction in the number of variates allows the

method to work well in higher-dimensional data provided the thresholding substantially reduces the number of entries in the stacked covariance matrix. Moreover, they propose a bootstrap approach for threshold selection to get around the difficulties of estimating the limiting distribution of (2.2.4). However, the issue of estimating the long-run covariance matrix remains.

To address the covariance changepoint problem in high-dimensional data where p can be larger than n, Avanesov and Buzun (2018) propose an approach based on the comparison of precision matrices. They use a rolling window approach, where they search for changes by comparing estimates of high-dimensional precision matrices. Similarly, to Cho and Kirch (2021) they propose using different sized windows to obtain a multi-scale approach. For shorter windows, the estimation of the high-dimensional precision matrices may be poor and hence larger changes are needed to be detected. Whereas for larger windows, we gain a better estimation of the precision matrices at the cost of the minimum segment length between changes being larger. To determine which changes across the different windows are significant, a bootstrap calibration approach is used to set appropriate thresholds. Although this method can be used in a high-dimensional setting, the method lacks power for detecting smaller changes and usually requires large windows to detect changes. Thus the method is inappropriate for data sets with many changes that could be reasonably close together.

Recently, Wang et al. (2021) proposed an approach based on the leading singular

vector of the operator norm of the covariance CUSUM statistic defined as

$$S_t = \sqrt{\frac{e-t}{(e-s)(t-s)}} \sum_{i=s+1}^{t} y_{i,1:p} y_{i,1:p}^T - \sqrt{\frac{t-s}{(e-s)(e-t)}} \sum_{i=t+1}^{e} y_{i,1:p} y_{i,1:p}^T \, .$$

They provide a thorough theoretical investigation of this test statistic and provide guarantees on the size of changes that can be detected and on the localization of the changepoints for a range of thresholds. However, less information is given about the implementation of this method such as how to generate an appropriate threshold. Moreover, the method requires a large minimum segment length, $2p \log n$, to accommodate the theoretical results. Hence, despite its strong theoretical results, implementation of the method is challenging and it cannot be applied to series with several changes.

A common property of the covariance methods mentioned so far is they compare the difference of the covariance matrices to detect changes. This means their test statistics require an estimate of the underlying covariance matrix which is impractical when unknown changepoints are present. Hence, Ryan and Killick (2021) propose a test statistic that is based on the multivariate ratio matrix to alleviate this issue. This method is shown to have superior performance to other covariance methods for detecting and locating changepoints, however it can be computationally intensive for larger data sets due to the need to calculate the inverse of a matrix at each time point. Moreover, the method has a minimum segment length of p, thus the method is intractable for high-dimensional data sets or those with short segments.

The methods above aim to detect changes in a general covariance structure. However, in many applications, there is some underlying structure to this covariance matrix that

can be exploited to yield better detection rates. One such structure is to assume the data lie in a lower-dimensional subspace which enforces restrictions on the eigenvalues of the covariance matrix. Detecting changepoints in covariance matrices with this structure hasn't been explored in the offline multivariate changepoint setting and hence in Chapter 4 we propose a method for detecting changes in subspace. We show that when this restriction on the eigenvalues of the covariance matrix holds, we gain better power for detecting changes than using the general covariance changepoint methods mentioned above.

Alternative Methods

In Section 2.1.1, we introduced the non-parametric E-Divisive approach of Matteson and James (2014). This can also be used in a multivariate setting to detect changes in multivariate distributions and has become popular in the fully multivariate setting due to its strong performance at detecting mean and other types of changes. Another recent non-parametric approach is the MultiRank procedure of Cabrieto et al. (2017). They propose a test statistic designed to detect changes in correlation structure and interestingly use an SN approach to detect multiple changes rather than BS. This test statistic can struggle under more challenging data generating models and the use of SN to detect multiple changes raises issues regarding the computational efficiency. Finally, we mention that there are several popular Bayesian approaches for detecting fully multivariate changepoints such as Peluso et al. (2019) and many methods based on the popular reversible jump MCMC (Steward et al., 2016; Bolton and Heard,

2018).

2.2.2 Sparse Changepoints

Since multivariate changepoint analysis has started receiving more attention, a key question is how to deal with the sparsity of changepoints. In the fully multivariate setting, it is intuitive how to extend univariate methods to the multivariate setting, however, if we also assume that we do not know how many or which variates will change, the problem becomes substantially harder. For model-based cost function methods, the additional requirement of checking each combination of variates for a change, in addition to searching for an optimal segmentation across time, means the problem quickly becomes computationally infeasible. However, some methods have been presented in this area and we discuss these in this Section.

For test statistic based methods, the most common type of change to detect is a mean change and this extends naturally to sparsity - which of the variates have undergone a mean change and which have not. Here it is common to perform dimension reduction on the data, or some test statistic, to obtain better power for detecting changes. Eliminating the variates which do not change, reduces the noise in the test statistic making the change easier to detect in the variates containing a change. However, these methods do not give inference on which variates change nor have any theoretical results on this. Their aim is simply computational efficiency, and they embed their AMOC statistics within a search strategy that assumes a changepoint has occurred in all the series. The second part of this Section is a summary of the main test statistic

based methods for detecting sparse changes.

Model Based Methods

The sparsity of a changepoint can be extremely challenging especially if we are unsure how many of our variates will change. Maboudou-Tchao and Hawkins (2013) propose overcoming this issue by performing a fully multivariate approach for a change in a multivariate Gaussian distribution and then performing a hypothesis test on each variate to determine if it contained a change or not. This method is able to give some indication to which variates change but has some drawbacks. Firstly, if the change is very sparse then it will be difficult for the method to detect any changes in the first place as the noise from the constant variates will mask the changepoint. Furthermore, there is a multiple testing issue when considering the hypothesis test of each change in each variate. This could lead to either variates that change being missed if the multiple testing is dealt with appropriately, or many false alarms if each hypothesis test has a fixed false alarm rate which doesn't take into account multiple testing. Performing a hypothesis test on each variate is similar to using a univariate changepoint method on each variate and hence if there was a small change in all the variates the fully multivariate changepoint detection could detect a change in the first stage but then the hypothesis tests may not detect a change in any of the variates. Finally, performing a hypothesis test on each variate can be time-consuming especially as n and p grow large.

This issue of computational complexity is common with model-based cost functions

and optimal search methods. Pickering (2016) formulated a cost function that allows for any number of variates to change at each changepoint and they presented a pruned dynamic programming approach, similar to PELT, to solve this. The method, named Subset Multivariate Optimal Partitioning (SMOP), is able to solve the minimization of the global cost function exactly, however, has a computational cost of $\mathcal{O}(pn^{2p})$ making it infeasible unless both the number of time points and variates is small. Hence, it has become common to no longer use optimal search methods and instead uses an approximate search method for minimizing the global cost function (Bardwell et al., 2019; Fisch et al., 2019). In particular, Tickle et al. (2021) use a penalized cost approach but solve this approximately by assuming AMOC and using WBS to detect multiple changepoints. As this is a likelihood-based approach, it has the advantage of being able to detect different types of changes, however, a major limitation is the assumption of independence across time and variates. Violations of these independence assumptions can be dealt with by modifying the penalty, yet choosing how to do this is non-trivial.

Model-based cost functions typically have the advantage that they recover not only the changepoint locations but the variates that changed. Yet, combining these cost functions with optimal search methods appears to be an impractical way of dealing with sparse changes as their computational complexity is too high for even moderately sized data sets. Using model-based cost functions with approximate search methods improves the computational time of these methods and has lead to some promising work in this area. However, these methods then gain the challenges associated with

approximate search methods, such as masking. Striking a balance between these approaches is still an open area of research.

Test Statistic Based Methods

The main challenge with sparse changepoints is the lack of knowledge regarding the set S_k and which or how many of the variates will change. Siegmund et al. (2011) simplify this assumption by assuming a known fraction of the variates will change. They then use a statistic similar to (2.2.3) from Zhang et al. (2010) yet only apply it to the known fraction of the largest test statistics across the variates. This method has the advantage of being able to eliminate the noisy test statistics that don't exhibit a change but the assumption that there are a known fraction of the variates that change is limiting.

To overcome this assumption, Groen et al. (2013), proposes using two test statistics, one for the setting where lots of the variants undergo a change and one for when very few variates change. The first of these test statistics is based on the mean of the CUSUM statistics defined in (2.2.2),

$$\mathcal{C}(y_{1:n,1:p}) = \max_{1\leq t\leq n} \frac{1}{p} \sum_{j=1}^{p} S_{t,j} \ . \tag{2.2.6}$$

This could be seen as a fully multivariate test statistic and is similar to that presented in Horváth and Hušková (2012). The second is based on the maximum of the CUSUM statistics across the variates at each time point,

$$\mathcal{C}(y_{1:n,1:p}) = \max_{1\leq t\leq n} \max_{1\leq j\leq p} S_{t,j} \ . \tag{2.2.7}$$

By selecting the largest CUSUM statistic from all variates this test statistic is best suited for changes that occur in very few (ideally one) variate(s). Considering both of these statistics at once allows for the detection of changes with varying sizes of the set $\mathcal{S}_k$. However, using these test statistics gives no indication of which variates change, simply that a change has occurred. A post-processing step such as that in Maboudou-Tchao and Hawkins (2013) could be used to determine which variates change but this is cumbersome compared to model-based methods where the variates that change are recovered for free.

The maximum of the CUSUM statistics, as defined in (2.2.7), was also considered in Jirak (2015) in the high dimensional setting. They showed the asymptotic distribution of the test statistic as both $n, p \to \infty$, thus allowing for an appropriate choice of penalty in the high-dimensional setting. However, this method is only suited to very sparse changes and the variates that change are unknown.

To strike a balance between using all or just one of the CUSUM statistics in the test statistic, Cho and Fryzlewicz (2015) proposed a threshold approach. Here the CUSUM statistic contributes to the test statistic only if it exceeds a certain threshold, say γ. This leads to a test statistic of the form

$$\mathcal{C}(y_{1:n,1:p}) = \max_{1 \le t \le n} \sum_{j=1}^{d} S_{t,j} \mathbb{1}\left[S_{t,j} > \gamma\right] , \tag{2.2.8}$$

where $\mathbb{1}[\cdot]$ is the indicator function. This test statistic is able to recover which variates change by storing which variates have a test statistic greater than γ but this is not theoretically justified so in practice you can't make inference on those selected. Choosing this threshold γ is not an easy task and is one of the downfalls of using a threshold

approach such as this one. Moreover, the authors note that a post-processing step is required to trim down the changepoints as the method is prone to over-fitting changes.

To help with the choice of γ in the above approach, Cho (2016) proposed using a Double CUSUM statistic. Here the $S_{t,k}$ are ordered at each time point and a CUSUM statistic is used on the variates at each time location. Identifying the most likely changepoint across the ordered $S_{t,k}$ aids the choice of threshold γ. We can simply choose γ such that those after the changepoint where the $S_{t,k}$ are small are not included. This method provided a great way of choosing the threshold γ but in turn requires us to choose another penalty for the CUSUM across the $S_{t,1:p}$. The authors suggest using a bootstrap approach to pick these thresholds eliminating the need for convergence to a limiting distribution. Again however, there is no theoretical guarantees for the selection so we can't perform inference on the selected variates.

Another method that aims to aggregate only certain variates from the CUSUM statistics is the Inspect method of Wang and Samworth (2018). Here the authors aim to compute the oracle projection direction of the CUSUM statistics to maximize the change size. The oracle projection direction is the normalized difference of the mean vectors on either side of the change however estimating this projection direction is difficult. To do this, the authors aim to find the k-sparse leading left singular vector of the CUSUM statistics, however, this is an NP-hard problem and hence instead a convex relaxation of the problem is solved. This method is designed for sparse changepoints and struggles in settings where many variates change. Furthermore,

the method is prone to over-fitting changepoints which is likely due to the tuning of the required hyper-parameters in the estimation of the projection direction and the implementation of the WBS algorithm, which is known to over-fit changepoints.

We have seen some different ways of overcoming the issue of sparsity within test statistic based methods, mainly based on the CUSUM statistic. The main issue is the unknown number of variates that undergo a change and whether to use a method that looks at all of the CUSUM statistics as in (2.2.6) or just the maximum as in (2.2.7). In recent work, Enikeeva and Harchaoui (2019) present a thorough theoretical consideration of the issue of sparsity. They find that for changes that occur in greater than $\sqrt{p}$ of the variates (dense change), then using methods that consider all of the CUSUM statistics will have more power than those that consider only a subset of the variates. On the other hand, if the number of changing variates is less than $\sqrt{p}$ (sparse change), then using a subset of the test statistics such as those in (2.2.7) and (2.2.8) will yield more power for detecting changes. Using these findings, the authors present two test statistics, one based on the l_2-norm of the CUSUM statistics, similar to that in Horváth and Hušková (2012), and one based on a scan statistic that aims to find the best subset of CUSUM statistics to use. They then propose combining these two test statistics to create an overall test statistic that is capable of detecting changes in the dense and sparse changepoint settings. If the change is in the sparse regime, then it will be detected by the scan statistic and here we could recover the variates that change easily. However, in the dense regime, as all variates are used in the test statistic, it is not possible to determine which variates changed.

The choice of model-based methods or test statistic based methods for sparse changepoints comes down to the problem at hand. If you require knowledge of the variates that change then model-based methods accommodate this more easily, and allow inference on the identified variates. However, these methods can be computationally expensive especially if you wish to solve the global cost function exactly. This computational expense can be decreased by using an approximate search method, however, additional issues such as independence between the variates and time points along with the required knowledge of the data generating process can be inhibiting. If you are searching for mean changes and only need the changepoint locations then test statistic based methods may be more appropriate. Moreover, these methods tend to be more robust to dependence structures across the variates and time, meaning model misspecification is less of an issue. With all these methods the choice of search method and penalty presents similar issues to the univariate case as discussed in Section 2.1.2 and 2.1.3.

Chapter 3

High-Dimensional Changepoint Detection via a Geometrically Inspired Mapping

3.1 Introduction

Time series data often have abrupt structural changes occurring at certain time points, known as changepoints. To appropriately analyze, model or forecast time series data that contain changes we need to be able to accurately detect where changepoints occur. High-dimensional changepoint analysis aims to accurately and efficiently detect the location of changepoints as both the number of dimensions and time points increase. High-dimensional changepoint analysis is an ever-growing research area and has multiple applications including finance and economics (Modisett and Maboudou-

Tchao, 2010); longitudinal studies (Terrera et al., 2011) and genetics (Bleakley and Vert, 2011).

Changepoint analysis in the univariate setting is a well-studied area of research with early work by Page (1954) and overviews can be found in Eckley et al. (2011) and Brodsky and Darkhovsky (2013). The multivariate extension has received less attention, see Truong et al. (2020) for a recent review. One major challenge with high-dimensional changepoint analysis is the computational burden of an increasing number of dimensions. To partially reduce this computational burden, a common assumption is that changepoints are assumed to occur in all series simultaneously (Maboudou-Tchao and Hawkins, 2013); a sparse set of series (Wang and Samworth, 2018); or a dense set of series (Zhang et al., 2010). Within these settings, a common approach is to first project the time series to a single dimension and then use a univariate changepoint method on the projected time series. For example, Zhang et al. (2010), Horváth and Hušková (2012) and Enikeeva and Harchaoui (2019) consider an l_2-aggregation of the CUSUM statistic while Jirak (2015) considers an l_∞-aggregation that works well for sparse changepoints. A recent advancement was the Inspect method proposed by Wang and Samworth (2018) who aim to find an optimal projection direction of the CUSUM statistic to maximize a change in mean.

Current projection methods are generally limited to detecting changes in a single parameter, usually the mean. Therefore, these methods cannot be used in many practical scenarios where multiple features of the time series change. An alternative, nonparametric approach was taken in Matteson and James (2014) where U-Statistics

were used to segment the time series. As this is a non-parametric method it can detect different types of changes in distribution but becomes computationally infeasible as the number of time points increases. The methods above almost exclusively use a Binary Segmentation approach (Scott and Knott, 1974; Vostrikova, 1981), or derivations thereof (Fryzlewicz, 2014), to detect multiple changepoints. This can lead to poor detection rates as conditional identification of changes can lead to missing or poor placement of changepoints due to factors such as masking. This occurs when a large change is masked by two smaller changes on either side acting in opposite directions; this idea is explained further in Fryzlewicz (2014).

A key novelty in this paper is to map a given high-dimensional time series onto two dimensions instead of one. Inspired by a geometric representation of data, we map each high-dimensional time vector to its distance and angle from a fixed pre-defined reference vector based upon the standard scalar product. These mappings show shift and shape changes in the original data corresponding to mean and variance changes. Given the geometric inspiration, we denote the method GeomCP throughout.

In Section 3.2, we set up the high-dimensional changepoint problem before defining the geometric mappings used in GeomCP. Also, we discuss an alternative approach to Binary Segmentation that can be applied to the univariate mapped series. An extensive simulation study is performed in Section 3.3, which compares GeomCP to competing available multivariate changepoint methods, Inspect (Wang and Samworth, 2018) and E-Divisive (Matteson and James, 2014). Section 3.4 presents two applications from genetics and finance. Section 3.5 gives concluding remarks.

3.2 Methodology

In this section, we set up the high-dimensional changepoint problem for our scenario. We define our new method, GeomCP, and discuss how changes in high-dimensional time series manifest themselves in the mapped time series. We then suggest an appropriate univariate changepoint detection method for detecting changes in the mapped time series - although practically others could be used.

Before proceeding, we define some notation used throughout the paper. We define the $\mathbb{1}_p$ vector as a p-dimensional vector where each entry is 1 and the number of dimensions, p, is inferred from context. For a vector, $\boldsymbol{y} = (y_1, \ldots, y_p)^T$, we define the l_q-norm as $\|\boldsymbol{y}\|_q := \left(\sum_{j=1}^{p} |y_j|^q \right)^{\frac{1}{q}}$ for $q \in [1, \infty)$. We define $\langle \cdot, \cdot \rangle$ as the standard scalar product such that for vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ we have $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_{j=1}^{p} x_j y_j$. Finally, the terms variables, series and dimensions shall be used interchangeably to indicate the multivariate nature of the problem.

3.2.1 Problem Setup

We study the time series model where $\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n}$ are independent, p-dimensional time vectors that follow a multivariate Normal distribution where,

$$\boldsymbol{Y_i} \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2 \boldsymbol{I}_p), \quad 1 \leq i \leq n \, .$$

We assume there are an unknown number of changepoints, m, which occur at locations $\tau_{1:m} = (\tau_1, \ldots, \tau_m)$. These changepoints split the data into $m + 1$ segments, indexed k, that contain piecewise constant mean and variance vectors, $\boldsymbol{\mu}_k$ and $\boldsymbol{\sigma}_k^2$. Note, we

assume a diagonal covariance matrix so the covariance matrix can be described by the variance vector and the identity matrix. We define $\tau_0 = 0$ and $\tau_{m+1} = n$ and assume the changepoints are ordered so, $\tau_0 = 0 < \tau_1 < \ldots < \tau_m < \tau_{m+1} = n$.

The following section introduces the geometric intuition and mappings used within GeomCP. These mappings reduce the dimension of the problem to make the problem computationally feasible as n and p grow large.

3.2.2 Geometric Mapping

When analyzing multivariate time series from a geometric viewpoint, we seek to exploit relevant geometric structures defined in the multi-dimensional space. Here we aim to detect changepoints in the mean and variance vectors of multivariate Normal random variables; therefore, we wish to utilize geometric properties that capture these changes.

A change in the mean vector of our data generating process will cause a location shift of the data points in the multi-dimensional space. Consider a distance between each data point and some fixed reference point, if the data points are shifted in the multi-dimensional space then their distance to the reference point would be expected to change. Hence, we can detect when the mean vector of the data generating process changes by observing a change in the distances. For a change in distance not to occur after an underlying mean change, the new mean vector must remain exactly on the same $(p-1)$-sphere (centered in the reference point) that the old mean vector lay on. Given that the computation of the mean vector is a linear operator on the

multivariate time series, the requirement to lie on the same sphere (a quadric in $\mathbb{R}^p$) is highly non-generic from a geometric prospective. As a result, these scenarios are rare especially in high-dimensions.

A change in the covariance of our data generating process will cause a change in the shape of the data points. More specifically in our setup, a change in the variance would cause the shape of the data points to expand or contract. Consider the angle between each data vector and a reference vector, as the shape of the data points expands (contracts) the angles will become more (less) varied. Hence, we can detect changes in the variance of the data generating process by detecting changes in the angles.

By using distances and angles, we can map a p-dimensional time series to two dimensions. To calculate these mappings, we need a pre-specified reference vector to calculate a distance and angle from. Naturally, one may think to use the mean of the data points. However, this requires a rolling window to estimate the mean of data points prior to the point being mapped. Not only does this introduce tuning parameters, such as the size of the rolling window, but will result in spikes in the distance and angle measures at changepoints. To detect changepoints, we would need a threshold for these spikes and calculating such a threshold is a non-trivial task, hence, we seek an alternative.

We propose setting the reference vector to be a fixed vector, $\boldsymbol{y_0}$. We then translate all the points based upon this fixed reference vector,

$$y'_{i,j} = y_{i,j} - (\min_i y_{i,j} - y_{0,j}) \ , \ i \in [1, \ldots, n] \ , \ j \in [1, \ldots, p] \ . \tag{3.2.1}$$

This translation works by finding the smallest value in each dimension, j, across all time points, i, and taking away the reference vector from this. All the points are then translated in the same way so the new vector of minimum values is equal to the reference vector. Figure 3.2.1 shows this for a set of points in 2-dimensions with the reference vector set as $\boldsymbol{y_0} = \mathbb{1}$.

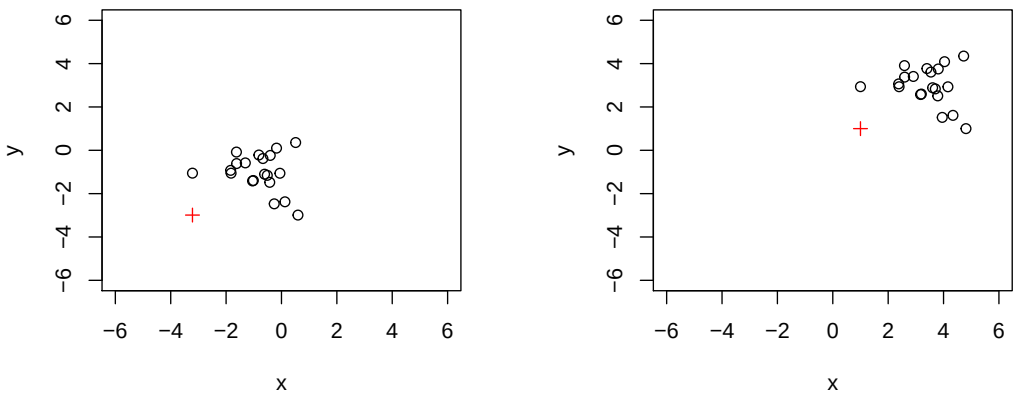
We choose to set $\boldsymbol{y_0} = \mathbb{1}$ as this bounds the angle measure between 0 and $\pi/4$ meaning we do not get vectors close to the origin facing in opposite directions causing non-standard behavior within a segment. Moreover, having a non-zero element in every entry of $\boldsymbol{y_0}$ ensures changes in the individual series will manifest in the angle measure. Note due to the translation in (3.2.1), the choice of $\boldsymbol{y_0}$ does not affect the distance measure. Throughout we assume the reference vector is set as $\boldsymbol{y_0} = \mathbb{1}$.

For data points in the same segment, we would expect their distances and angles to the reference vector to have the same distribution. When a mean (variance) change occurs in the data, this leads to a shift (spread) in the data, hence, the distances (angles) will change. Therefore, by detecting changes in the distances and angles, using an appropriate univariate changepoint method, we recover changepoints in the p-dimensional series.

We define our distance and angle measures based upon the standard scalar product. To obtain our distance measure, d_i, we perform a mapping, $\delta : \mathbb{R}^p \to \mathbb{R}^1_{>0}$,

$$d_i = \delta(\boldsymbol{y_i}) = \sqrt{\langle (\boldsymbol{y_i'} - \mathbb{1}), (\boldsymbol{y_i'} - \mathbb{1}) \rangle}\,, \tag{3.2.2}$$

which is equivalent to $\|\boldsymbol{y_i'} - \mathbb{1}_p\|_2$.



(a) Pre-translation (b) Post-translation

Figure 3.2.1: 2-dimensional example of the translation of the data points based upon a reference vector y_0. In the pre-translation, the black circles indicate the points y_i and the red cross is smallest value in each dimension. In the post-translation, the black circles are the translated points y_i' and the red cross is the reference vector y_0.

To obtain our angle measure, a_i, we perform a mapping $\alpha : \mathbb{R}^p \to [0, \frac{\pi}{4}]$,

$$a_i = \alpha(\boldsymbol{y_i}) = \cos^{-1}\left(\frac{\langle \boldsymbol{y_i'}, \mathbb{1}\rangle}{\sqrt{\langle \boldsymbol{y_i'}, \boldsymbol{y_i'}\rangle}\sqrt{\langle \mathbb{1}, \mathbb{1}\rangle}}\right), \tag{3.2.3}$$

which is the principal angle between $\boldsymbol{y_i'}$ and $\mathbb{1}$.

By using the standard scalar product we are incorporating information from each series in the distance and angle measures. As such, we would expect GeomCP to perform well in scenarios where a dense set of the series change at each changepoint. This idea will be explored further and verified in Section 3.3.

3.2.3 Analyzing Mapped Time Series

Understanding the distributional form of the distance and angle mappings will aid in the choice of univariate changepoint methods. Under our problem setup, Theorem 3.2.1 shows that the distance measure, asymptotically in p, follows a Normal distribution.

Theorem 3.2.1. *Suppose we have independent random variables, $Y_i \sim N(\mu_i, \sigma_i^2)$. Let $X = \sqrt{\sum_{i=1}^{p} Y_i^2}$, then as $p \to \infty$,*

$$\frac{X - \sqrt{\sum_{i=1}^{p}(\mu_i^2 + \sigma_i^2)}}{\sqrt{\frac{2\sum_{i=1}^{p}(\mu_i\sigma_i)^2 + \sum_{i=1}^{p}\sigma_i^4 + 2\rho\sqrt{2\sum_{i=1}^{p}\sum_{j=1}^{p}\mu_i^2\sigma_i^2\sigma_j^4}}{2\sum_{i=1}^{p}(\mu_i^2+\sigma_i^2)}}} \xrightarrow{\mathcal{D}} N(0,1) \ ,$$

where ρ is an unknown correlation parameter (see proof).

Proof. See Appendix A.2. □

Theorem 3.2.1 shows that, asymptotically in p, the distance between each time vector and a pre-specified fixed vector follows a Normal distribution. Hence, for piecewise constant time vectors, the resulting distance measure will follow a piecewise constant Normal distribution. It is common in the literature to assume that angles also follow a Normal distribution, as in Fearnhead et al. (2018). We found by simulation, for large enough p, the angle measure defined in (3.2.3) is well approximated by a Normal distribution with piecewise constant mean and variance.

Whilst any theoretically valid univariate method could be used to detect changepoints in the mapped series, we use the PELT algorithm of Killick et al. (2012) as this is

an exact and computationally efficient search. For $n \to \infty$, PELT is consistent in detecting the number and location of changes in mean and variance (Tickle et al., 2020; Fisch et al., 2018), hence, using Theorem 3.2.1, we gain consistency of our distance measure as $p \to \infty$ also. When the Normal approximation of the distance and angle measures holds, we use the Normal likelihood as our test statistic within PELT and allow for changes in mean and variance. If p is small, we may not want to make the Normal assumption. In this case, we recommend using a non-parametric test statistic, such as the empirical distribution from Zou et al. (2014) (where consistency has also been shown) as embedded within PELT in Haynes et al. (2017b).

3.2.4 GeomCP Algorithm

Algorithm 1 details the pseudo-code for GeomCP. As changepoints can manifest in both the distance and angle measure, we post-process the two sets of changepoints to obtain the final set of changes. We introduce a threshold, ξ, and say that a changepoint in the distance measure, $\hat{\tau}^{(d)}$, and a changepoint in the angle measure, $\hat{\tau}^{(a)}$, are deemed the same if $\left|\hat{\tau}^{(d)} - \hat{\tau}^{(a)}\right| \leq \xi$. If we determine two changepoints to be the same we set the changepoint location to be the one given by the angle measure as Section 3.3.2 demonstrates, this results in more accurate changepoint locations. The choice of ξ should be set based upon the minimum distance expected between changepoints. Alternatively, ξ could be set to zero and then an alternative post-processing step would be required to determine whether similar changepoint estimates correspond to the same change.

Algorithm 1 GeomCP

Input: $\boldsymbol{Y} \in \mathbb{R}^{n \times p}$, threshold $= \xi$, *Univariate Cpt Method*.

Step 1: Centralize data by $y'_{i,j} = y_{i,j} - \left(\min_i y_{i,j} - 1\right)$.

Step 2: Perform distance mapping: $\boldsymbol{y_i} \xrightarrow{\delta} d_i, \ \forall i$.

Step 3: Perform *Univariate Cpt Method* on $\boldsymbol{d}$ to recover cpts, $\hat{\boldsymbol{\tau}}^{(d)}$.

Step 4: Perform angle mapping: $\boldsymbol{y_i} \xrightarrow{\alpha} a_i, \ \forall i$.

Step 5: Perform *Univariate Cpt Method* on $\boldsymbol{a}$ to recover cpts, $\hat{\boldsymbol{\tau}}^{(a)}$.

Step 6: $\forall k$, **if** $\min \left| \hat{\boldsymbol{\tau}}^{(a)} - \hat{\tau}_k^{(d)} \right| < \xi$ **then** remove $\hat{\tau}_k^{(d)}$ from $\hat{\boldsymbol{\tau}}^{(d)}$.

Return: $\hat{\boldsymbol{\tau}} = \text{sort}(\hat{\boldsymbol{\tau}}^{(a)}, \hat{\boldsymbol{\tau}}^{(d)})$

One of the major downfalls of many multivariate changepoint methods is they are computationally infeasible for large n and p. Within GeomCP, the computational cost to calculate both the distance and angle measures in (3.2.2) and (3.2.3) is $\mathcal{O}(np)$. If we implement the PELT algorithm for our univariate changepoint detection, this has expected computational cost $\mathcal{O}(n)$ under certain conditions. The main condition requires the number of changepoints to increase linearly with the number of time points, further details are given in Killick et al. (2012). If these conditions are not satisfied, PELT has an at worst computational cost of $\mathcal{O}(n^2)$. Hence, the expected computational cost of GeomCP is $\mathcal{O}(np+n) = \mathcal{O}(np)$ (under the conditions in Killick et al. (2012)) and has at worst computational cost $\mathcal{O}\left(np + n^2\right)) = \mathcal{O}\left(n(p+n)\right)$.

3.2.5 Non-Normal and Dependent Data

The current problem setup assumes multivariate Normal distributed data with a diagonal covariance matrix. These assumptions are made to facilitate our theoretical analysis and result in the Normality of the mapped series. If these assumptions are broken, the geometric intuition described in Section 3.2.2 still holds, but we can say less about the theoretical properties of the mapped series.

Firstly, if we allow for an arbitrary covariance matrix, this describes the shape and spread of the data points. Suppose our data undergoes a change from $X_{\text{pre}} \sim N(\mathbf{0}, \Sigma)$ to $X_{\text{post}} \sim (\mathbf{0}, \boldsymbol{\sigma}\Sigma)$ this will cause the data points to spread out in the directions of the principle components. Hence, we would still expect the angles between the time vectors and the reference vector to change, revealing the change in covariance. We investigate this further in Section 3.3.5. In fact, a Normal distributed data set with a known covariance matrix could be transformed into a Normal distributed data set with a diagonal covariance matrix (satisfying our initial problem setup) by an orthogonal transformation that aligns the axes with the principle components. Such a transformation would preserve the distances and angles by definition but requires knowledge of the true covariance structure.

Alternatively, we could consider other inner products in our distance and angle mappings defined in (3.2.2) and (3.2.3); here the geometric motivation of the method would remain valid. In this case, for an underlying mean change to occur without the distance measure changing, the new mean vector must remain exactly on the more general $(p-1)$-quadric in $\mathbb{R}^p$. This is still a highly non-generic requirement from a

geometric prospective. In particular, we could use scalar products directly derived from the covariance matrix, such as the Mahalanobis Distance (Mahalanobis, 1936). In such cases, the direct relation between angles and the correlation coefficients is well known (Wickens, 1995). However, such inner products require an estimate of the covariance in each segment, which is non-trivial and therefore left as future work.

If we allow the data to be distributed from a non-Normal distribution then we would expect changes to the first and second moment of these distributions to still manifest in the distance and angle mappings. However, being able to understand the distribution of the mapped series would be more challenging. In practice, the empirical cost function could be used within PELT (Haynes et al., 2017b) yet this would lead to less power in the detection of changes in the univariate series.

Finally, if we allowed temporal dependence between the time points this would lead to temporal dependence in the mapped series and an appropriate, cost function for PELT could be used. Understanding how the temporal dependence in the multivariate series manifests in the mapped series is non-trivial and is left as further work.

In the next section, we provide an extensive simulation study exploring the effectiveness of GeomCP at detecting multivariate changes in mean and variance and demonstrate an improved detection rate on current state-of-the-art multivariate changepoint methods. Furthermore, we illustrate the improved computational speed of GeomCP over current methods, especially as n and p grow large.

3.3 Simulation Study

In this section, we provide a comparison of GeomCP; the Inspect method of Wang and Samworth (2018); and the E-Divisive method of Matteson and James (2014) using the statistical software R (R Core Team, 2019). First, we investigate how changes in mean and variance of time series manifest themselves in the distance and angle measures within GeomCP. We then compare GeomCP to Inspect and E-Divisive in a wide range of scenarios including dense changepoints, where the change occurs in all or a large number of dimensions, and sparse changepoints, where the change occurs in a small number of dimensions. Changes in both mean, variance and a combination of the two will be considered.

Inspect is only designed for detecting changes in mean, therefore, it will only be included in such scenarios. In addition, Inspect is designed for detecting sparse changepoints, however, Inspect 'can be applied in non-sparse settings as well' (Wang and Samworth, 2018) so we also include it in the dense change in mean scenarios. Like GeomCP, E-Divisive is designed for dense changepoints, but, we will also include it in the sparse changepoint scenarios to assess performance.

For GeomCP we perform the mappings in (3.2.2) and (3.2.3) before applying the PELT algorithm using the *changepoint* package (Killick and Eckley, 2014). Unless otherwise stated, we use the default settings; namely, the MBIC penalty (Zhang and Siegmund, 2007), Normal distribution and allow for changes in mean and variance. We implement the Inspect method using the *InspectChangepoint* package (Wang and

Samworth, 2016). The thresholds used to identify significant changepoints are calculated before timing the simulations using the data-driven approach suggested in Wang and Samworth (2018). For the remaining user-defined parameters, we use the default settings with $Q = 0$. Setting $Q = 0$ implements a Binary Segmentation approach (Scott and Knott, 1974; Vostrikova, 1981) for identifying multiple changepoints. When using $Q = 1000$, as suggested in Wang and Samworth (2018), a Wild Binary Segmentation (Fryzlewicz, 2014) approach is implemented to detect multiple changes. However, this becomes computational infeasible even at moderate levels of n and p while only resulting in minor improvements in detection rate at the expense of higher false discovery rates. For $p > 1000$, the data-driven calculation of the thresholds was computationally infeasible, hence, the theoretical threshold derived in Wang and Samworth (2018) was originally implemented. However, this led to an excessive number of false positives and, as such, is not included. For the implementation of the E-Divisive method, we use the *ecp* package (James and Matteson, 2014) with $\alpha = 1$; minimum segment size of 30; a significance level of 0.05; and $R = 499$ as suggested by Matteson and James (2014).

Unless indicated otherwise, we simulate data from a Normal distribution with changes in mean and variance given in each scenario. Additionally, the number of changepoints is set as $m = \lceil \frac{n}{200} \rceil$ and we distribute the changepoints uniformly at random throughout the time series with the condition that they are at least 30 time points apart. Where computationally feasible, we perform 500 repetitions of each scenario and dis-

play the true detection rate (TDR) and false detection rate (FDR) along with their confidence intervals given by two standard errors. For scenarios with $n \geq 1000$, E-Divisive was only run on 30 replications due to the high computational cost. Changepoint estimates are deemed correct if they are the closest to, and within 10 time points of, the true changepoint and contribute to the TDR. Changepoint estimates more than 10 time points from the true changepoints or where another estimated changepoint is closer to the true changepoint are deemed false and contribute to the FDR. We seek a TDR as close to 1 as possible and an FDR as close to 0 as possible. As GeomCP estimates changepoints in both the distance and angle measures we apply the reconciling method from Section 3.2.4 with the threshold, $\xi = 10$. Then we apply the same TDR/FDR method to the reconciled changes.

3.3.1 Size of Changepoints

As we are interested in multivariate changepoints, we need to decide upon the size of a change in each series. If we fixed a specific change size in each series, then as p increases, the change becomes easier to identify due to multivariate power. If we fixed a total change size across all series, then as p increases, the change becomes considerably harder to detect. Hence, we set our simulated change sizes so that GeomCP has an approximately constant performance across p, in terms of TDR and FDR.

To achieve a constant performance in the change in mean scenario, we require the difference in the expected distance measure pre- and post-change to be constant across

p. If we assume unit variance and a set mean across all series before the change, $\tilde{\mu}_{\text{pre}}$ and after the change, $\tilde{\mu}_{\text{post}}$, using Theorem 3.2.1 the expected difference in the distance measure before and after a changepoint is,

$$\mathbb{E}(d_{\text{post}} - d_{\text{pre}}) = \sqrt{p}\left(\sqrt{\tilde{\mu}_{\text{post}}^2 + 1} - \sqrt{\tilde{\mu}_{\text{pre}}^2 + 1}\right) .$$

If we set the total mean change size in our simulated data as,

$$\sum_{j=1}^{p} \mu_{j,\text{post}} - \mu_{j,\text{pre}} = \sqrt{p}\Theta , \tag{3.3.1}$$

for some constant Θ and, again, assume the mean of each series is the same, we gain,

$$\begin{aligned}\Theta &= \sqrt{p}\,(\tilde{\mu}_{\text{post}} - \tilde{\mu}_{\text{pre}}) \\ &\approx \sqrt{p}\left(\sqrt{\tilde{\mu}_{\text{post}}^2 + 1} - \sqrt{\tilde{\mu}_{\text{pre}}^2 + 1}\right) \\ &= \mathbb{E}(d_{\text{post}} - d_{\text{pre}}) .\end{aligned}$$

Hence, for a constant Θ, using a total mean change size scaling as in (3.3.1) will result in the expected difference of the distance pre- and post-change, and therefore the performance of GeomCP, being approximately constant across p. As we re-scale our data before applying our two mappings, the pre- and post-change means will be large enough that this approximation is reasonable.

Similarly, to gain an approximately constant performance of GeomCP across p for a change in variance, we set the total variance change size in our simulated data as,

$$\prod_{j=1}^{p} \frac{\sigma_{j,\text{post}}}{\sigma_{j,\text{pre}}} = \Phi^{\sqrt{p}} , \tag{3.3.2}$$

for some constant Φ. When comparing methods, we shall use (3.3.1) and (3.3.2) to define the total change size for each scenario, with the change size being the same in all series that undergo a change.

3.3.2 GeomCP Investigation

First, we investigate how changes in mean, variance and a combination of the two, manifest themselves in the distance and angle measure within GeomCP. We set $n = 1000$ and $p = 200$ and simulate data with changepoints $\boldsymbol{\tau} = (250, 500, 750)$. At τ_1 we have a mean change of $+0.1$ in all series; at τ_2 we have a variance change of $\times 1.2$ in all series; and at τ_3 we have a mean change of -0.1 and a variance changes of $\times 1.2^{-1}$ in all series. Figure 3.3.1 shows 4 of the 200 series and shows the changepoints are undetectable by eye in the individual series. Applying the mappings within GeomCP results in the mapped series seen in Figure 3.3.1 where the changes are clearly identifiable in at least one of the distance or angle measure.

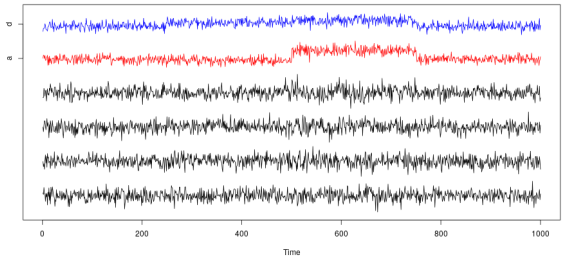
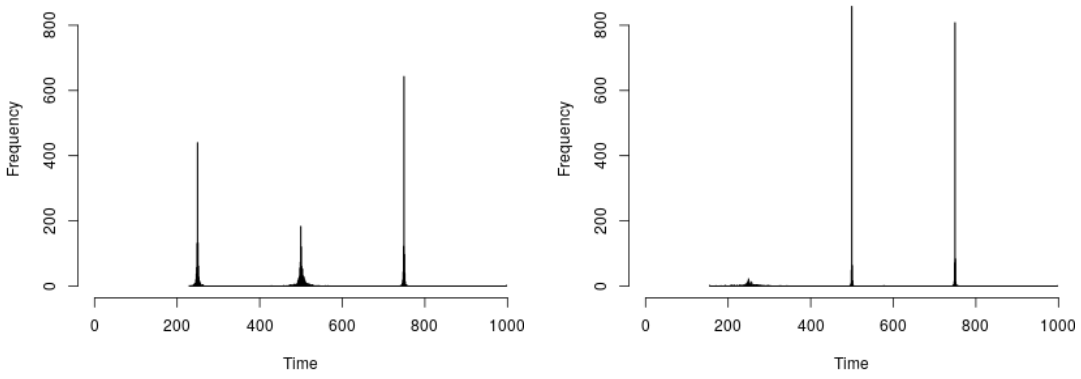


Figure 3.3.1: 4 series from the simulated data set with the distance (d) and angle (a) mappings showing 3 changepoints that are not obvious in the individual series

Figure 3.2(a) and 3.2(b) shows the position of identified changepoints in the distance and angle measure in 1000 replications of the current scenario using PELT. The

relatively small change in mean at time point 250 is only reliably picked up by the distance measure. The change in variance is picked up by the angle measure in almost all cases and is also seen in the distance measure, however, with less accuracy and less often. The change in mean and variance at time point 750 is reliably detected in both the distance and angle measures. These findings were similar for varying mean and variance changes. As such, this justifies setting the location of changepoints that occur in both series to be given by the angle changepoint location as stated in Section 3.2.4.



(a) Distance Changepoints (b) Angle Changepoints

Figure 3.3.2: Locations of detected changepoints in 1000 repetitions of simulated data set with changepoints at 250, 500 and 750, in mean, variance and, mean and variance, respectively

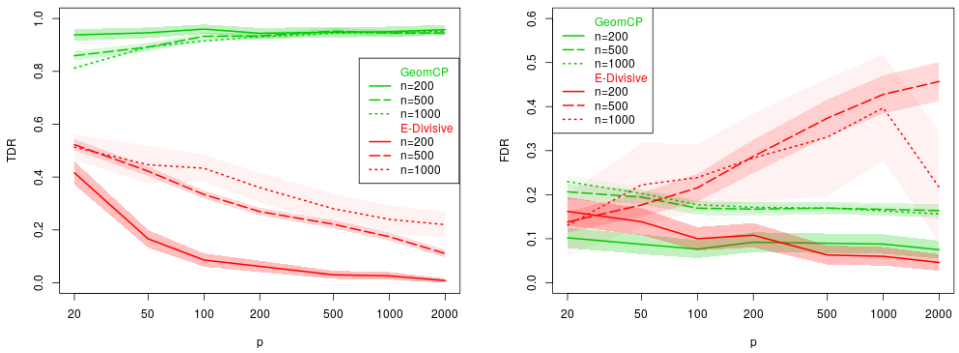
3.3.3 Dense Changepoints

Now we compare GeomCP's performance with E-Divisive and Inspect. We investigate dense variance changes here, with mean, and mean and variance changes given in the

Appendix A.3 and A.4.

We simulate data with variance changes that occur in all series for a wide range of n and p and show a subset of the results here. We keep the mean vector constant and we split the total change size defined in (3.3.2) evenly across all series. We display results with $\Phi = 3$ as this is shown to give a high TDR while maintaining a low FDR in Eckley et al. (2011) for $p = 1$. Similar findings occur with varying values of Φ; see Appendix A.8. We apply the GeomCP and E-Divisive methods to these simulated data sets and the TDR and FDR are shown in Figure 3.3.3.

Figure 3.3(a) shows the TDR across different numbers of dimensions and time points. It is clear that GeomCP outperforms E-Divisive in terms of TDR and the gap between the methods widens as the number of dimensions increases. Figure 3.3(b) shows that the improved TDR of GeomCP does not come at the expense of a higher FDR, which has similar rates across n and p.



(a) TDR (b) FDR

Figure 3.3.3: TDR and FDR for GeomCP and E-Divisive for simulated data sets containing variance changes that occur in all series for multiple n and p

In the mean, and mean and variance change scenarios, GeomCP similarly outperforms both E-Divisive and Inspect in terms of TDR whilst maintaining a low-level FDR across n and p. Results can be found in Appendix A.3 and A.4.

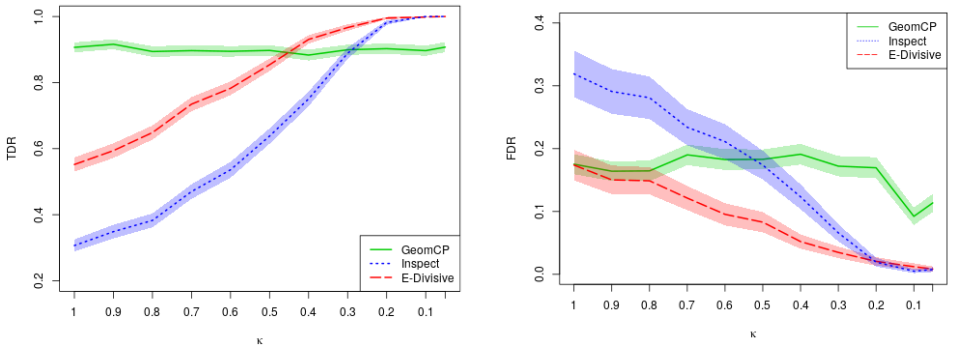
3.3.4 Sparsity Investigation

Thus far we assumed that all series undergo a change at each changepoint. We now explore the effect of the sparsity of the changepoint. We define $\kappa \in (0, 1]$ to be the probability that a series undergoes a change. We explore sparse mean changes here, with sparse variance changes included in Appendix A.5.

For the sparse changepoint scenarios, we set $n = 500$, $p = 200$ and vary κ; we note that there were similar findings for different n and p. We keep the variance vector constant and the change size in each series that undergoes a change, is the total change size defined in (3.3.1), split between the expected number of series to undergo a change. This means the expected total change size is the same as when all series undergo a change. We display results with $\Theta = 1.2$ and similar findings occur with varying values of Θ; see Appendix A.8. We apply the GeomCP, Inspect and E-Divisive methods to these scenarios and the TDR and FDR are shown in Figure 3.3.4.

Figure 3.4(a) shows that GeomCP maintains a constant TDR across κ as expected. This reflects the set up of the scenario where the expected total change size is constant across κ. For dense changepoints, GeomCP compares well as we might expect. Interestingly, E-Divisive also assumes dense changepoints but performs poorly in this scenario. Inspect is designed for sparse changes and as expected, for very sparse

changes the method performs the best. For sparse changepoints, the improved performance of Inspect and E-Divisive may be due to the size of change in each affected series increasing as κ decreases.



(a) TDR (b) FDR

Figure 3.3.4: TDR and FDR for GeomCP, Inspect and E-Divisive for simulated data sets with sparse mean changes for $n = 500$ and $p = 200$

3.3.5 Between-series Dependence

Now we will relax the assumption of a diagonal covariance matrix and investigate how this affects the performance of GeomCP. We will investigate how two different covariance matrix structures compare to the independent, diagonal covariance case. Here we will investigate variance changes in these covariance structures with mean changes explored in the supplementary material.

For these scenarios, we set $n = 200$, $p = 100$ and have one changepoint at $\tau = 100$. The pre-changepoint data will be distributed from a $N(\mathbf{0}, \Sigma)$ while the post-

changepoint data distributed from a $N(\mathbf{0}, \boldsymbol{\sigma}\Sigma)$. We will vary the change size, $\boldsymbol{\sigma}$, while each entry of $\boldsymbol{\sigma}$ will be identical for each change size. We will compare three structures for Σ:

1. Independent case: $\Sigma = I$.

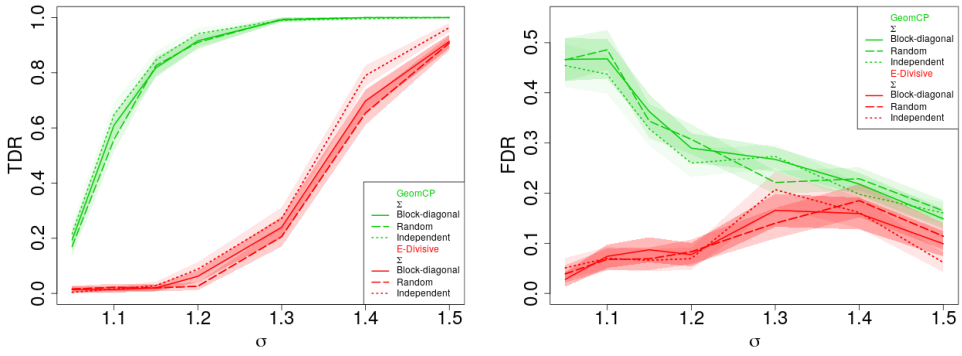
2. Block-diagonal case: Here Σ will be a block-diagonal matrix with block size of 2. The off-diagonal entries will be randomly sampled from a $U(-0.6, -0.3) \cup U(0.3, 0.6)$ distribution with the diagonal entries equal to 1.

3. Random case: Here we let $\Sigma = PDP'$ where P is an orthogonalized matrix of standard Normal random variables and D is a diagonal matrix with entries decreasing from 30 to 1.

As we no longer have independence between series we cannot assume Normality of the distance and angle measures within GeomCP. Hence, we use the empirical cost function (Haynes et al., 2017b) within PELT to detect changes in the distance and angle measures. We similarly use the empirical cost function in the independent case for comparability.

Figure 3.5(a) shows the TDR of GeomCP and E-Divisive for varying change sizes, σ, and the different covariance structures. GeomCP clearly has a greater TDR than E-Divisive for smaller change sizes. However, Figure 3.5(b) shows this comes at the expense of a higher FDR. This is to be expected when using the empirical cost function within PELT as this generally produces a higher FDR. By altering the penalty used within PELT this FDR could be reduced at the cost of some power in detecting

changes. Yet for $\sigma \geq 1.3$ GeomCP has a competitive FDR with E-Divisive while having a superior TDR. Interestingly, the covariance structure has very little impact on the performance of GeomCP, this follows our intuition from Section 3.2.5.



(a) TDR (b) FDR

Figure 3.3.5: TDR and FDR for GeomCP and E-Divisive for simulated data with a change in covariance for $n = 200$ and $p = 100$

3.3.6 Computational Speed

A major issue with high-dimensional changepoint detection is, as n and p grow large, many multivariate changepoint methods become computationally infeasible. Here, we compare the computational speeds of GeomCP, Inspect and E-Divisive for a range of n, p and m. We will compare the speeds in three scenarios:

1. n increasing while $p = 200$ and $m = \lceil \frac{n}{200} \rceil$.

2. n increasing while $p = 200$ and $m = 2$.

3. p increasing while $n = 500$ and $m = \lceil \frac{n}{200} \rceil = 3$.

The second scenario breaks PELT's assumption of a linearly increasing number of changepoints as the number of time points increases. This means the speed of detecting changepoints using GeomCP will no longer be linear in time. We performed simulations using the three scenarios defined above and only included mean changes so we can compare with Inspect. We set the mean change size to be $\theta_j = 0.8$ in all series so that the changes are obvious. For scenario 1 and 2, E-Divisive was computationally infeasible for $n \geq 1000$. For scenario 3, Inspect's speed is only shown for $p < 1000$ due to the excessive computational time of generating a data-driven threshold. Note that the data-driven thresholds needed for Inspect were calculated outside of the recorded times. In practice, if a threshold was required, then Inspect would take considerably longer to run especially as p increases. Within GeomCP we run the algorithm in serial, performing the mapping and changepoint identification for the distance and then for the angle. These could be processed in parallel, leading to a further reduction in computational time.

Figure 3.3.6 shows the computational speed of each method in the three scenarios. We can see from Figure 3.6(a) that, in scenario 1, GeomCP is the fastest of the three methods for all n. As n increases the difference between the speeds of GeomCP and Inspect increases (note the log scale on both axes). We can also see, E-Divisive is substantially slower than GeomCP and Inspect for all n and its run time increases rapidly as n gets large. Scenario 1 supports our claim that GeomCP has linear run time in n when the required assumptions of PELT are met.

In scenario 2, shown in Figure 3.6(b), we break the assumption within PELT that the

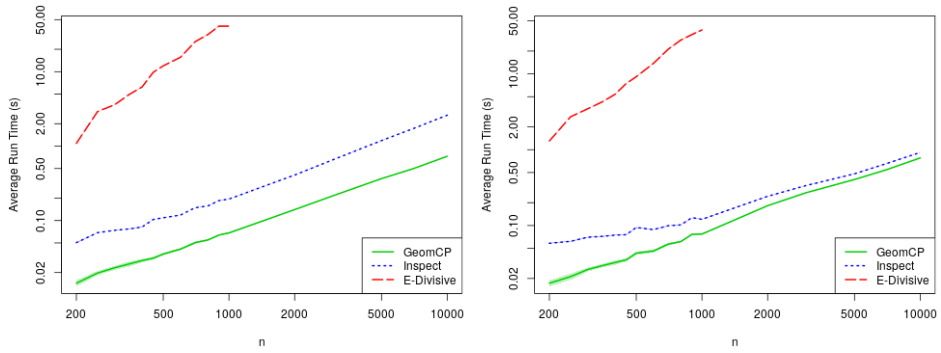
number of changepoints is increasing linearly in time. This results in a comparatively slower performance of GeomCP, although, it remains computationally faster than Inspect for all n shown. Similarly to scenario 1, E-Divisive has a much longer run time than both GeomCP and Inspect.

Finally, for scenario 3 Figure 3.6(c) shows that, for small p, Inspect is the fastest of the methods but as p increases above 50 GeomCP is computationally faster. Whilst Inspect is faster for $p < 50$, recall that this does not include the time for the calculation of the threshold. Interestingly, the run time of E-Divisive appears unaffected by p until $p \geq 1000$. This is likely due to it's computational cost being mainly affected by the number of changepoints and time points, which remain constant. Scenario 3 also supports our claim that GeomCP has linear run time as p increases, note the log scale that distorts the linearity of the plot.

3.4 Applications

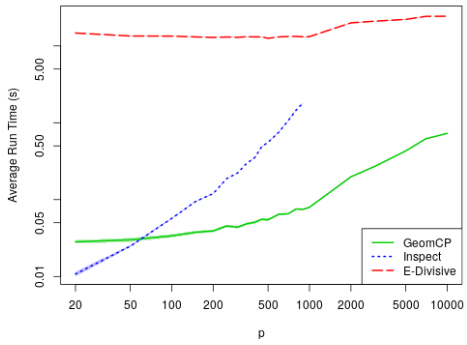
3.4.1 Comparative Genomic Hybridization

We study the comparative genomic hybridization microarray data set from Bleakley and Vert (2011). Comparative genomic hybridization allows the detection of copy number abnormalities in chromosomes by comparing the fluorescent intensity levels of DNA fragments between a test and reference sample. The data set contains log-intensity-ratio measurements from 43 individuals with bladder tumors with measurements taken at 2215 different positions on the genome. This data set is available



(a)

(b)



(c)

Figure 3.3.6: Average run time for each method when: (a) n is increasing, $p = 200$ and m is increasing; (b) n is increasing, $p = 200$ and $m = 1$; (c) $n = 500$, p is increasing and $m = 3$ by default

in the *ecp* R package (James and Matteson, 2014).

Copy number abnormalities come in regions on the genome and can either be specific to the individual or can be shared across several individuals. It is the latter that are of more interest as these are more likely to be disease-related. E-Divisive and Inspect have both been used to segment this data set with their results shown in Matteson and James (2014) and Wang and Samworth (2018) respectively. Under

the default settings, these two methods fitted a large number of changepoints, 93 and 254 respectively, which may not be representative of changes occurring across multiple individuals. Wang and Samworth (2018) suggest selecting the 30 most significant changepoints to counter this, however, the justification for choosing 30 is unknown.

To perform our analysis we first scale each series, similarly to Inspect, using the median absolute deviation to allow a better comparison. We then use the two mappings within GeomCP and apply the PELT algorithm, using the R package *changepoint.np* (Haynes and Killick, 2021), to the resulting mapped series. The mappings do not appear Normal for this application, hence, we use the empirical cost function and set the number of quantiles as $4\log(n)$, as suggested in Haynes et al. (2017b). We use the CROPS algorithm of Haynes et al. (2017a) to identify an appropriate penalty value with diagnostic plots shown in Appendix A.9. This lead to 37 changepoints being identified and these are shown in Figure 3.4.1 with the signal for 8 individuals from the study and the distance and angle mappings. Approximately 67.5% of the changepoints identified by GeomCP corresponded to those identified by E-Divisive (within 3 time points), with the majority of the rest corresponding to where E-Divisive fitted two changepoints. Also, the changepoints identified seem to be common across multiple individuals while changes specific to a series are not detected. It is promising that our proposed segmentation identifies similar changepoints as other methods, while only identifying those that seem common across multiple individuals. Other GeomCP segmentations, using different potential penalty values identified in CROPS, resulted

in more or less of the individual features from specific series being detected.

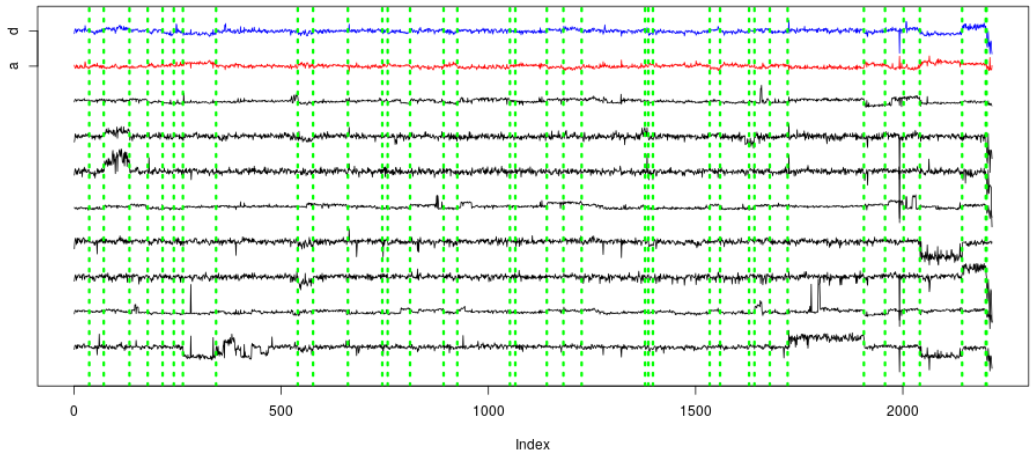


Figure 3.4.1: Log-intensity-ratio measurements of microarray data from 6 out of 43 individuals and distance (d) and angle (a) mappings with vertical lines showing the identified changepoints

3.4.2 SP500 Stock Prices

We now investigate the daily log-returns of the closing stock prices for 447 companies included in the S&P500 from January 2015 through to December 2016. This data set was created by Nugent (2018) and was loaded using the R package *SP500R* (Foret, 2019). The aim is to identify changes in log-returns that affect a large number of companies rather than changes that are specific to individual companies. First we scale each series using the median absolute deviation. Next we apply the mappings within GeomCP, before using the PELT algorithm from the *changepoint* package (Killick et al., 2016) to both mapped series using the Normal cost function. We used the CROPS algorithm of Haynes et al. (2017a) to identify an appropriate penalty

value for both series with diagnostic plots shown in Appendix A.9.

Using GeomCP, we identified 10 changepoints. These are shown in Figure 3.4.2 along with the log-returns of the first 10 companies from the S&P500 list and the mapped distance and angle measures. These changepoints correspond to key events that we would expect to impact the stocks of a large number of companies. For example, the changepoints in August 2015 correspond to large falls in the Chinese stock markets with the Dow Jones industrial average falling by 1300 points over 3 days. The changepoints in February and late June 2016 likely correspond to the announcement and subsequent result of the British referendum to leave the European Union. Applying the E-Divisive method, (with the minimum segment length set to 2 and the rest of the user defined parameters set as in Section 3.3) resulted in only 2 changepoints, both occurring in August 2015 similar to those detected in GeomCP.

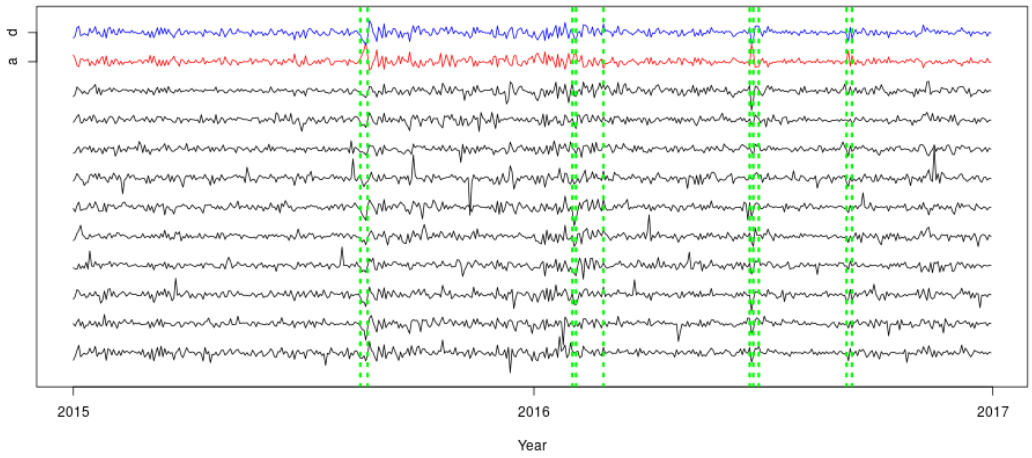


Figure 3.4.2: Log-returns of 10 out of 447 companies within the S&P500 and the distance (d) and angle (a) mappings with vertical lines showing the identified changepoints

3.5 Conclusion

We have presented a new high-dimensional changepoint detection method that can detect mean and variance changes in multivariate time series. This is achieved by implementing a univariate changepoint detection method on two related geometric mappings of the time series. We have shown that looking at the high-dimensional changepoint problem from a geometric viewpoint allows us to utilize relevant geometric structures to detect changepoints. We have displayed an improved performance in detecting and identifying the location of multiple changepoints over current state-of-the-art methods. Furthermore, we have shown an improved computational speed over competing methods when using the univariate changepoint method PELT. Finally, we have shown the effectiveness of GeomCP at detecting changepoints when applied to applications.

We have discussed how to extend this methodology to non-Gaussian data along with temporal and between-series dependence. However, a thorough investigation of how changes manifest in the distance and angle measure in the presence of these structures is left as future work.

The GeomCP method is implemented in the R package *changepoint.geo* available on CRAN `https://CRAN.R-project.org/package=changepoint.geo`.

Chapter 4

Subspace Changepoint Detection in Multivariate Time Series

4.1 Introduction

It is common for time series data to have abrupt structural changes that occur at certain time points, known as changepoints. Changepoint analysis is crucial to appropriately model, forecast or label time series data that contain structural changes. In the multivariate setting, the aim is to detect changes in multiple series simultaneously; this could be, but not limited to, changes in the mean vector or a change in the covariance structure. Multivariate changepoint analysis is a growing research area and has multiple applications including finance and economics (Maboudou-Tchao and Hawkins, 2013); genetics (Zhang et al., 2010); and internet security (Peng et al., 2004).

Until recently, the majority of changepoint research considered univariate time series with early work by Page (1954) and overviews in Eckley et al. (2011) and Brodsky and Darkhovsky (2013). Increasing attention is now being given to the multivariate extension, see Truong et al. (2020) for a recent review. In the multivariate setting, the vast majority of work focuses on the change in mean problem (Jirak, 2015; Wang and Samworth, 2018; Enikeeva and Harchaoui, 2019) while Grundy et al. (2020) allows for mean and variance changes simultaneously, assuming a fixed covariance structure between the series. There has been some work on identifying changes in covariance. Aue et al. (2009a) propose a version of the traditional CUSUM statistic to detect covariance changes, while Avanesov and Buzun (2018) and Wang et al. (2021) propose methods that allow for the high-dimensional setting, where the number of series can grow large.

Multivariate data often exhibits some low-dimensional structure, for example lying in a low-dimensional subspace. There is substantial literature aiming to exploit low-dimensional structures within multivariate data. Principal component analysis (PCA) (Pearson, 1901; Hotelling, 1933; Jolliffe, 2002) is a classic method for reducing the dimensionality of a data set to a low-dimensional subspace such that the variance within the subspace is maximized. Subspace clustering is another vast area of research where data points are assumed to lie in a union of low-dimensional subspaces and the aim is to cluster the points based on the subspace they lie in, see Parsons et al. (2004) for a review.

Despite the substantial literature in the clustering setting, there has been limited

work on exploiting low-dimensional subspace structures in the changepoint literature. Blythe et al. (2012) consider feature extraction to improve the detection of three current changepoint methods but do not consider subspace changepoints themselves, while Kawahara et al. (2007) present an algorithm based upon subspace identification using an observability matrix. The closest work to our setup is that of Jiao et al. (2018) who considered the subspace changepoint problem in an online setting where the aim is to identify a change in subspace as quickly as possible as additional data is collected. In this setting, historical data is used to estimate the current subspace and then a CUSUM style approach is used to identify changes to this underlying subspace. Xie et al. (2020) also present an online subspace changepoint method, however, their focus is on detecting the emergence of a subspace within the data. In contrast, our contribution is to the offline setting where all the data is known a priori and we aim to detect a change from one subspace to another prioritizing accuracy of the location of the change.

Time series data that lie in a subspace can also be viewed as exhibiting a spiked covariance structure, as introduced in Johnstone (2001). Thus, a subspace change is a special type of covariance change with additional assumptions on the eigenvalues of the covariance matrix. Detecting covariance changepoints is a challenging task, hence if the data is assumed to lie in a subspace we can exploit this additional information within the subspace changepoint framework; allowing for improved performance in changepoint detection. Applications for subspace changepoint detection are numerous including analysis of Motion Capture data (Jiao et al., 2018); seismic event detection

(Xie et al., 2020); and video segmentation (Vidal et al., 2005).

This paper proposes a novel method for detecting subspace changepoints in multivariate time series data. In Section 4.2, we formalize the subspace changepoint problem and propose a method for consistent subspace estimation which is used in our test statistic for detecting subspace changepoints. We also provide a method for calculating a data-driven threshold for determining the significance of potential changepoints. In Section 4.3, we perform an extensive simulation study and compare our method to existing state-of-the-art covariance changepoint methods. In Section 4.4 we suggest an extension of our method to allow for multiple changepoints and illustrate its effectiveness in an additional simulation experiment. Finally, Section 4.5 presents an application to Motion Capture data before Section 4.6 gives concluding remarks.

Note that during the peer-review process of this Chapter a link to factor models, in particular identifying changes in factor loadings was discovered. This is discussed in more detail in Section 6.2.

4.1.1 Notation

Before proceeding we first define some notation used throughout the paper. All matrices will be in capitalised bold font, $\boldsymbol{M} \in \mathbb{R}^{i \times j}$. Vectors will be in lower-case bold font, $\boldsymbol{v} \in \mathbb{R}^{j}$, while scalars will be in lower-case font, $s \in \mathbb{R}$. $\boldsymbol{I}$ will represent the identity matrix with dimension inferred from context. $\|\boldsymbol{M}\|_F$ represents the Frobenius norm of the matrix $\boldsymbol{M}$, while $\|\boldsymbol{v}\|_2$ represents the Euclidean norm of the vector $\boldsymbol{v}$.

4.2 Methodology

Let $\{\boldsymbol{y}_t\}_{t=1}^n$ be a multivariate time series with $\boldsymbol{y}_t \in \mathbb{R}^p$ for $t = 1, \ldots, n$. We assume these data lie in a latent low-dimensional linear subspace, $\mathcal{S}$, of known dimension $q < p$. Hence, our data at a single time point, t, takes the form,

$$\boldsymbol{y}_t = \boldsymbol{x}_t + \boldsymbol{\epsilon}_t \ ,$$

where $\boldsymbol{x}_t \in \mathcal{S}$ and $\boldsymbol{\epsilon}_t \in \mathbb{R}^p$ is independent additive noise. We can re-write $\boldsymbol{x}_t = \boldsymbol{Z}\boldsymbol{s}_t$ where $\boldsymbol{Z} \in \mathbb{R}^{p\times q}$ is an orthonormal basis of the subspace $\mathcal{S}$ and $\boldsymbol{s}_t \in \mathbb{R}^q$ are independent, mean-zero representations of the signal within the subspace. Furthermore, we assume the elements $s_{t,j}$ within $\boldsymbol{s}_t$ are independent with $\mathbb{E}[s_{t,j}] = 0$, $\mathbb{E}[s_{t,j}^2] = \sigma_{s,j}^2$ and $\mathbb{E}[s_{t,j}^4] < \infty$. We assume $\boldsymbol{\epsilon}_t$ is distributed similarly to $\boldsymbol{s}_t$ albeit with a common variance within $\boldsymbol{\epsilon}_t$, $\mathbb{E}[\epsilon_{t,j}] = \sigma_\epsilon^2$ for all $j = 1, \ldots, p$. Finally, we define $\rho = \min_{j\in[1,q]} \frac{\sigma_{s,j}^2}{\sigma_\epsilon^2}$ as the signal-to-noise ratio within the subspace and assume $\rho > 1$.

At an unknown time point, $\tau \in [1, n]$, we allow the latent subspace to potentially change (the changepoint), however we assume the subspace dimension, q, remains fixed. Denote the pre- and post-change subspace as $\mathcal{S}_1$ and $\mathcal{S}_2$ respectively with their bases represented by $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$. This leads to the changepoint formulation,

$$\boldsymbol{y}_t = \begin{cases} \boldsymbol{Z}_1\boldsymbol{s}_t + \boldsymbol{\epsilon}_t, & \text{if } t \leq \tau \\ \\ \boldsymbol{Z}_2\boldsymbol{s}_t + \boldsymbol{\epsilon}_t, & \text{if } t > \tau \ . \end{cases} \tag{4.2.1}$$

Our changepoint analysis aims to identify the most likely changepoint position τ and provide estimates of the two subspaces $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$. Note the case of $\tau = n$ represents the scenario where no change has occurred in the data and we would only obtain

an estimate of $\boldsymbol{Z}_1$. We focus on the at most one change (AMOC) setting described in (4.2.1) but provide an extension to detecting multiple changepoints in Section 4.4.

For a changepoint to be present in the data, there must be a difference between the pre- and post-change subspaces. Following Jiao et al. (2018), we define a metric space for the distance between two subspaces using the projection Frobenius norm distance (F-distance). The F-distance between two subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$ with orthonormal bases $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ is defined as

$$d(\mathcal{S}_1, \mathcal{S}_2) := \left\| \boldsymbol{Z}_1(\boldsymbol{Z}_1)^T - \boldsymbol{Z}_2(\boldsymbol{Z}_2)^T \right\|_F . \tag{4.2.2}$$

4.2.1 Subspace Estimation

The pre- and post-change subspace bases $\boldsymbol{Z}$ are unknown. Hence, we need a way to estimate $\boldsymbol{Z}$ based on some given data. Define the matrix $\boldsymbol{Y}_{b:e}$ to be a matrix with rows $\boldsymbol{y}_{b:e}$ corresponding to the time points we wish to estimate the subspace of. To gain an estimate of $\boldsymbol{Z}$ we use PCA. Specifically, we perform an eigendecomposition of the covariance matrix of $\boldsymbol{Y}$ yielding

$$\frac{1}{n}\boldsymbol{Y}^T\boldsymbol{Y} = [\hat{\boldsymbol{Z}}, \hat{\boldsymbol{Z}}_\perp]\boldsymbol{D}[\hat{\boldsymbol{Z}}, \hat{\boldsymbol{Z}}_\perp]^T . \tag{4.2.3}$$

Here $\hat{\boldsymbol{Z}} \in \mathbb{R}^{p\times q}$ is a matrix whose columns are the first q eigenvectors of the covariance matrix of $\boldsymbol{Y}$ and is an estimator for an orthonormal basis of the subspace. The estimator, $\hat{\boldsymbol{Z}}$, achieves the minimum squared distance of the data points, $\boldsymbol{Y}$, to a

q-dimensional subspace (Jolliffe, 2002);

$$\left\| \left(\boldsymbol{I} - \hat{\boldsymbol{Z}} \hat{\boldsymbol{Z}}^T \right) \boldsymbol{Y} \right\|_2^2 = \min_{\boldsymbol{U} \in \mathbb{R}^{p \times q},\, \boldsymbol{U}^T \boldsymbol{U} = \boldsymbol{I}} \left\| \left(\boldsymbol{I} - \boldsymbol{U} \boldsymbol{U}^T \right) \boldsymbol{Y} \right\|_2^2 ; . \tag{4.2.4}$$

Additionally, $\hat{\boldsymbol{Z}}_\perp \in \mathbb{R}^{p \times (p-q)}$ is a matrix whose columns are the remaining $p - q$ eigenvectors of the covariance matrix of $\boldsymbol{Y}$ and this is an orthonormal basis for the orthogonal compliment of the latent subspace. Finally, $\boldsymbol{D}$ is a diagonal matrix whose entries are the ordered eigenvalues, λ_j for $j = 1, \ldots, p$, of the covariance matrix of $\boldsymbol{Y}$.

Under the assumption that $\boldsymbol{s}_t$ and $\boldsymbol{\epsilon}_t$ are Normally distributed, Anderson (1963) showed that PCA results in a consistent estimator of $\boldsymbol{Z}$ (and $\boldsymbol{Z}_\perp$). If the error structure in the data is only present in the orthogonal directions to the subspace, then Shen et al. (2016) shows the consistency of the estimators $\hat{\boldsymbol{Z}}$ and $\hat{\boldsymbol{Z}}_\perp$.

Throughout, $\hat{\boldsymbol{Z}}^{(b:e)}$ and $\hat{\boldsymbol{Z}}_\perp^{(b:e)}$ will represent the estimate of the orthonormal basis of the subspace and orthogonal subspace, respectively based on the time points $\boldsymbol{y}_{b:e}$ with $1 \leq b < e \leq n$. Additionally, $\lambda_j^{(b:e)}$ will represent the j^{th}-largest eigenvalue of the covariance of the specified time points. For (4.2.4) to hold, the number of time points we are estimating the subspace of must be greater than the dimension of the data ($e - b > p$). Hence, we require a minimum segment length (msl) between changepoints and the boundaries of the data, of at least p. Setting a larger msl based upon prior knowledge of the problem at hand would be advisable for consistency of estimation.

4.2.2 Test Statistic

Section 4.2.1 demonstrated how we can estimate the subspace for a segment of data. We use this to define a cost function for a set of time points that describes the goodness-of-fit of the data to it's estimated subspace. A lower cost indicates a better fit and that the data truly lie in the estimated subspace. We use these costs to test for the presence of a changepoint such that at the true changepoint location,

$$\mathcal{C}(\boldsymbol{y}_{1:n}) \geq \mathcal{C}(\boldsymbol{y}_{1:\tau}) + \mathcal{C}(\boldsymbol{y}_{(\tau+1):n}) + \gamma \,, \tag{4.2.5}$$

where γ is a penalty to guard against over-fitting.

A natural cost function is the sum of the orthogonal distance of each time point to its estimated subspace. Specifically, we define the cost of time points $b:e$ as

$$\mathcal{C}(\boldsymbol{y}_{b:e}) = \sum_{i=s}^{e} \left\| \left(\boldsymbol{I} - \hat{\boldsymbol{Z}}^{(b:e)} \left(\hat{\boldsymbol{Z}}^{(b:e)} \right)^T \right) \boldsymbol{y}_t \right\|_2^2 = \sum_{i=s}^{e} \left\| \left(\hat{\boldsymbol{Z}}_{\perp}^{(b:e)} \right)^T \boldsymbol{y}_t \right\|_2^2 \,, \tag{4.2.6}$$

where $\hat{\boldsymbol{Z}}^{(b:e)}$ and $\hat{\boldsymbol{Z}}_{\perp}^{(b:e)}$ are calculated using (4.2.3). If the time points $i = b, \ldots, e$ all lie in the same subspace then $\hat{\boldsymbol{Z}}_{\perp}^{(b:e)}$ will be a good estimator and the projection of the points onto the subspace will be small. On the other hand, if the time points $i = b, \ldots, e$ belong to a union of subspaces (a changepoint is present) then the estimation of the subspace will be poor meaning the projections onto the subspace will be large.

To test for a change in subspace, we follow the traditional changepoint approach, comparing the cost of the whole data with the cost of adding a changepoint at any time point τ. This leads to the following test statistic,

$$\mathcal{T} = \mathcal{C}(\boldsymbol{y}_{1:n}) - \min_{\tau \in [\text{msl}, n-\text{msl}]} \left\{ \mathcal{C}(\boldsymbol{y}_{1:\tau}) + \mathcal{C}(\boldsymbol{y}_{(\tau+1):n}) \right\} \tag{4.2.7}$$

with the cost, $\mathcal{C}(\cdot)$ defined in (4.2.6). A large test statistic indicates the presence of a changepoint while a small test statistic indicates no changepoint is present. We deem the test statistic to be significant if it exceeds some threshold γ which we can see from (4.2.5) guards against over-fitting changepoints.

In the no-change scenario, the latent subspace will be the same for the whole data, hence, $\hat{\boldsymbol{Z}}_{\perp}^{(b:e)}$ should be similar for any $1 \leq b < e \leq n$. This means the test statistic should be close to zero and should not exceed the threshold γ. If a change is present, the estimator $\hat{\boldsymbol{Z}}_{\perp}^{(1:n)}$ will be poor as we are estimating a single subspace based on data that lie in a union of subspaces. This will make the cost of the whole data, $\mathcal{C}(y_{1:n})$ large. In comparison, by allowing a changepoint at some time point τ, the estimators $\hat{\boldsymbol{Z}}_{\perp}^{(1:\tau)}$ and $\hat{\boldsymbol{Z}}_{\perp}^{(\tau+1:n)}$ will be improved, meaning the cost of the two segments either side of the change will be small. This will result in a large test statistic which, if the change is large enough, should exceed the threshold γ. We determine that a changepoint has occurred if the test statistic exceeds the threshold, γ, and the estimated changepoint location is the value of τ which maximizes the test statistic, $\mathcal{T}$, in (4.2.7).

4.2.3 Threshold Choice

Once the test statistic in (4.2.7) has been calculated we need to compare it to some threshold to determine if the estimated changepoint is significant. Traditionally, we compare the test statistic to its true distribution under the scenario of no change. The type-1 error can be set to a pre-determined level, $\alpha \in [0, 1]$, and the test statistic is deemed significant if it exceeds the $(1 - \alpha)$-quantile of the true distribution. In

reality, the true distribution of the test statistic is unknown so there are two ways to proceed. Firstly, we could consider the asymptotic distribution of the test statistic with unknown subspaces. However, this is non-trivial and unlikely to maintain the type-1 error rate for small segments. The second approach, which we use in this paper, is to estimate the threshold using a data-driven approach. There are several advantages to using a data-driven threshold over a theoretical threshold. Most theoretical thresholds rely on asymptotics and can perform poorly in finite sample size examples. Moreover, theoretical thresholds often require strict distributional assumptions on the data generating process.

We propose using a permutation test to determine a suitable threshold, similar to that in Matteson and James (2014). Under the scenario of no change, the order of the time points is irrelevant as all time points lie in the same subspace and we assume the points are independent. We permute the time points, $\boldsymbol{y}_{1:n}$, creating a new time series, $\tilde{\boldsymbol{y}}_{1:n}$. We then calculate the test statistic in (4.2.7) using $\tilde{\boldsymbol{y}}_{1:n}$ and repeat this process for all possible permutations. We define the threshold, γ, based on a pre-specified α, by taking the $(1-\alpha)$-quantile of the permuted test statistics. If all possible permutations of the data are considered then the threshold would be exact in the sense that $\mathbb{P}(\mathcal{T} > \gamma) < \alpha$ if the data truly contained no changepoints. In reality, calculating all possible permutations is not computationally feasible and as such we gain an approximate threshold by performing P random permutations. The number of random permutations required to obtain an appropriate threshold that maintains the type-1 error rate, α, depends upon many factors including the length of

the data; the signal-to-noise ratio, ρ; and the subspace bases. The effectiveness of this permutation test, along with an exploration of the number of random permutations, P, required for an appropriate threshold is explored in Section 4.3.3.

4.2.4 Alternative Formulation of Cost Function

The calculation of the cost function defined in (4.2.6) requires the full eigendecomposition of the sample covariance matrix. This can be computationally intensive making the cost function infeasible to calculate for large data sets or when we come to consider multiple changepoints in Section 4.4. By considering the orthogonal distance from an alternative viewpoint, we find an algebraically equivalent cost function to (4.2.6) that is substantially faster to calculate at the expense of not recovering estimates of the subspace bases.

When calculating the orthogonal distance of each point to its estimated subspace we are effectively estimating the variance of the points in the directions of the orthogonal complement of the subspace. These variances are equivalent to the last $p - q$ eigenvalues of the sample covariance matrix. It can be shown that

$$\mathcal{C}(\boldsymbol{y}_{b:e}) = \sum_{i=b}^{e} \left\| \left(\hat{\boldsymbol{Z}}_{\perp}^{(b:e)} \right)^T \boldsymbol{y}_t \right\|_2^2 = (e-b) \sum_{j=q+1}^{p} \hat{\lambda}_j^{(b:e)} , \tag{4.2.8}$$

where both $\hat{\boldsymbol{Z}}_{\perp}$ and $\hat{\lambda}_j$ are calculated using the eigendecomposition in (4.2.3).

By using this equivalent definition of the cost function, we can calculate the test statistic in (4.2.7) substantially faster as we no longer need to estimate the eigenvectors of the covariance matrix, nor calculate the orthogonal distance for each point.

We do not get estimates of the bases of the underlying subspaces, however, these can be calculated post analysis, for a much smaller number of segments, if required. Simulations (not shown) verify the equivalence of these two formulations.

4.3 Simulation

In this section, we evaluate the performance of our test statistic on simulated datasets. First, we demonstrate the statistical power of our test statistic using ROC curves and show that changepoints estimated by our test statistic are tightly distributed around the true changepoint location. We then explore the performance of the permutation test proposed in Section 4.2.3 to provide an appropriate threshold for controlling the type-1 error while still maintaining strong statistical power. Next, we compare our test statistic to three state-of-the-art covariance changepoint methods to demonstrate that standard covariance methods are not always appropriate for subspace changepoints and a more problem-specific test statistic, such as ours, is required. Finally, we test the sensitivity of the test statistic to the assumed known subspace dimension q.

4.3.1 Data Generation

In order to generate data that contain subspace changepoints, we take the same approach as in Jiao et al. (2018). First, we generate the pre- and post-change subspaces the data are assumed to lie in. If $q \leq \frac{p}{2}$, we generate a random matrix, $\tilde{\boldsymbol{W}} \in \mathbb{R}^{p \times 2q}$ with i.i.d standard Gaussian entries. Upon orthonormalization, we gain a matrix

$\boldsymbol{W} = [\boldsymbol{W}_1, \boldsymbol{W}']$ where $\boldsymbol{W}_1$ and $\boldsymbol{W}'$ are the first and last q columns of $\boldsymbol{W}$ respectively. $\boldsymbol{W}_1$ is taken as the orthonormal basis of the pre-change subspace $\mathcal{S}_1$, while $\boldsymbol{W}'$ is an orthonormal basis of some subspace orthogonal to $\mathcal{S}_1$. In order to generate a basis for the post change subspace, $\mathcal{S}_2$, we take the scenario specific change size, $\Delta \in [0, \sqrt{q}]$, and generate an orthonormal basis, $\boldsymbol{W}_2$, by

$$\boldsymbol{W}_2 = \sqrt{1 - \frac{\Delta^2}{q}}\boldsymbol{W}_1 + \frac{\Delta}{\sqrt{q}}\boldsymbol{W}' \, . \tag{4.3.1}$$

It can be verified that the distance between subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$ is Δ, assuming the distance metric defined in (4.2.2). Intuitively, for $\Delta = 0$ then $\boldsymbol{W}_1 = \boldsymbol{W}_2$, hence $d(\mathcal{S}_1, \mathcal{S}_2) = 0$. Furthermore, for the maximum change size of $\Delta = \sqrt{q}$, then $\boldsymbol{W}_2 = \boldsymbol{W}'$ meaning $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are orthogonal implying a maximum distance between them. If $q > \frac{p}{2}$, we first generate the orthogonal complement spaces of $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$, as shown above, before transforming these to gain $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$.

The simulated dataset is then created by generating i.i.d $\boldsymbol{s}_t \sim N_q(\boldsymbol{0}, \sigma_s^2 \boldsymbol{I})$ and i.i.d $\epsilon_t \sim N_p(\boldsymbol{0}, \sigma_\epsilon^2 \boldsymbol{I})$. Then each time point i is defined as

$$\boldsymbol{y}_t = \begin{cases} \boldsymbol{W}_1 \boldsymbol{s}_t + \boldsymbol{\epsilon}_t, & \text{if } t \leq \tau \\ \\ \boldsymbol{W}_2 \boldsymbol{s}_t + \boldsymbol{\epsilon}_t, & \text{if } t > \tau \, . \end{cases} \tag{4.3.2}$$

Note, for simplicity we assume each entry of $\boldsymbol{s}_t$ is i.i.d with common variance σ_s^2.

Throughout all subspace bases, $\boldsymbol{W}$, will be generated randomly as described above. Moreover, all change sizes will be given as a function of the maximum change size in each scenario with $\Delta = \Theta\sqrt{q}$ where $\Theta \in [0, 1]$.

4.3.2 Power of Test Statistic

First, we use ROC curves to demonstrate the power of our proposed test statistic when detecting changes in subspace. We show one scenario here, with additional scenarios comparing different parameter setups given in Appendix B.1. ROC curves plot the true positive rate (TPR) for detecting a change against different false-positive rates (FPR). The aim is to achieve a curve as close to the top left corner as possible indicating a high TPR while maintaining a low FPR.

The ROC curves are generated by calculating our test statistic on 5000 repetitions of data with no changepoints ($n = 200$, $p = 20$, $q = 5$, $\sigma_s^2 = 1$, $\sigma_\epsilon^2 = 0.05$). These test statistics are used to obtain many different thresholds that correspond to different FPRs. Finally, we calculate our test statistic on 5000 equivalent data sets that contain a changepoint at $\tau = 100$ for different change sizes, Θ. For each threshold, the TPR is the number of repetitions where the test statistic exceeds the threshold. Note, as we are generating thresholds based on known null data we are assuming the full data generating process' for $\boldsymbol{s}_t$ and $\boldsymbol{\epsilon}_t$ are known.

Figure 4.3.1 shows the ROC curve for our specified scenario over different change sizes. For change sizes $\Theta > 0.1$ we have a detection rate above 75% while maintaining an FPR of only 0.05. This shows our test statistic has impressive power even for relatively small changes size. For comparison, Jiao et al. (2018) only consider change sizes $\Theta \geq 0.25$, although this is in the online changepoint setting.

Figure 4.3.1 shows our test statistic has substantial power to detect a change in

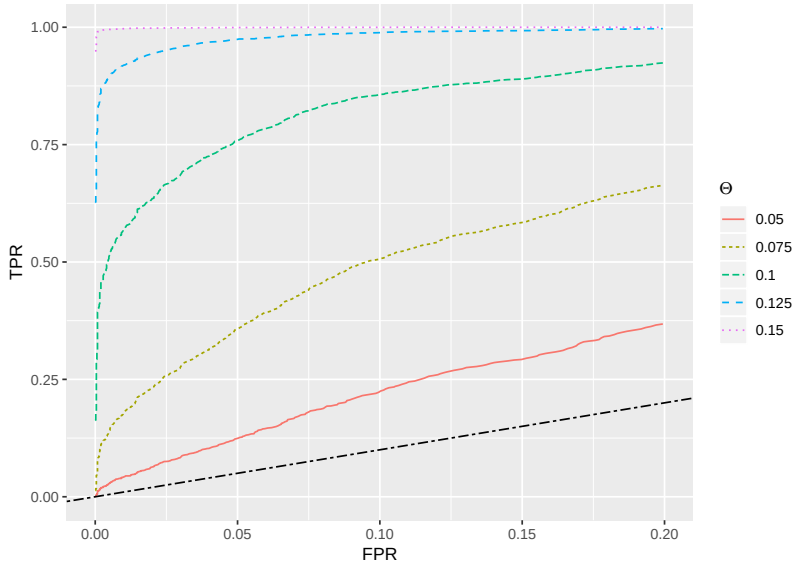


Figure 4.3.1: ROC curve showing the true positive rate for different false positive rates under varying change sizes

subspace, however, it only specifies whether a change is detected and not it's location. Figure 4.3.2 shows the locations of the significant changepoints in the above scenario assuming a FPR of 0.05. As expected, as the change size increases, the changepoint locations become increasingly concentrated around the true changepoint location, $\tau = 100$. For smaller change sizes the distribution of the changepoint locations is more widely spread and there are fewer changepoint estimates - this is to be expected as the TPR was lower in the ROC curves for the smaller change sizes.

The additional scenarios given in Appendix B.1 have some interesting conclusions. Firstly, if we keep the signal-to-noise ratio, $\rho = \frac{\sigma_s^2}{\sigma_\epsilon^2}$, constant we get similar ROC curves, indicating it is the ratio, ρ, and not the individual variances which affect the power of the test statistic. As expected, if we increase ρ we gain power and vice versa if we reduce ρ we lose power. Secondly, for fixed p increasing the subspace dimension

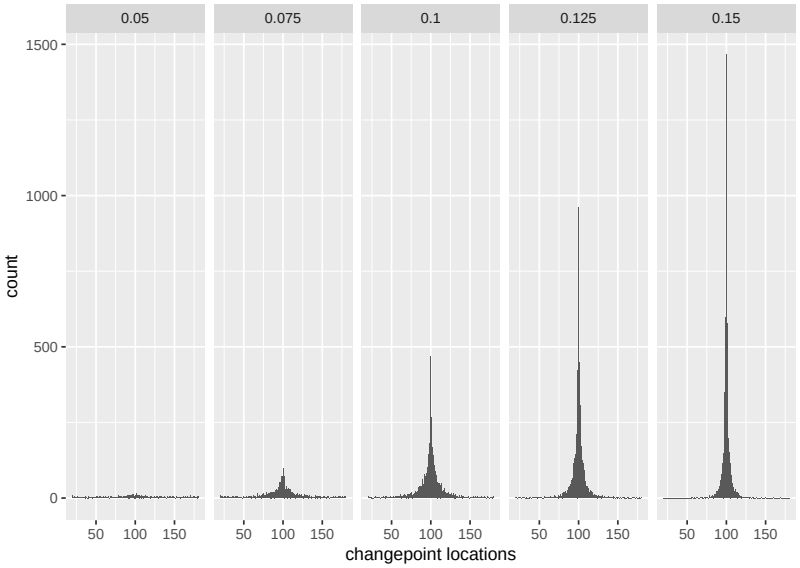


Figure 4.3.2: Histogram of changepoint positions for varying change sizes

q results in the test statistic having additional power. This is likely due to the change size Δ being a function of q and as such is larger as q increases. However, even with the difference or ratio between q and p being constant does not result in equal ROC curves. A thorough theoretical investigation into the power obtained for varying p and q would be interesting but is beyond the scope of this paper.

4.3.3 Permutation Test

In Section 4.3.2 we assumed the full data generating process' $\boldsymbol{s}_t$ and $\boldsymbol{\epsilon}_t$ are known in order to generate multiple thresholds. Here we use the data-driven permutation test described in Section 4.2.3 for determining a threshold. As this threshold is data-driven we don't require knowledge of the full data generating process' $\boldsymbol{s}_t$ and $\boldsymbol{\epsilon}_t$, however, this does mean we require larger change sizes Θ to take account of this extra uncertainty. First, we investigate the performance of the permutation test in controlling the FPR

under the scenario of no change, before demonstrating the method maintains a high TPR. Additionally, we explore the effects of the number of random permutations, P, on the performance of the test. Again, we show one parameter setup here ($p = 20$ and $q = 5$) with alternative values of p and q given in Appendix B.2.

First, we investigate the ability of the permutation test to control the FPR for different lengths of data sets, n, and a varying number of permutations, P. We generate 1000 repetitions of data with varying lengths n with no changepoints ($p = 20$, $q = 5$, $\sigma_s^2 = 1$ and $\sigma_\epsilon^2 = 0.05$). We calculate our test statistic and then generate a threshold for each repetition of data using the permutation test with $\alpha = 0.05$ and record if the test statistic exceeded the threshold, i.e. corresponds to a false positive. Figure 4.3.3 shows the FPR across all the repetitions for various n and P. This plot shows that even for a small number of permutations, P, the permutation test is controlling the FPR close to the desired level of 0.05.

As well as controlling the FPR, we also want our threshold to retain as much power as possible for correctly identifying changes. We simulated 200 repetitions of data with varying n and change sizes with a changepoint located at $\tau = \lceil 0.4n \rceil$ ($p = 20$, $q = 5$, $\sigma_s^2 = 1$, $\sigma_\epsilon^2 = 0.05$, $P = 200$, $\alpha = 0.05$). We deem a changepoint estimate to be correct if it is within a window of 20 on either side of the true changepoint. Figure 4.3.4 shows the TPR for the different scenarios.

Figure 4.3.4 shows that as n gets larger our test statistic gains power to detect changes, however even for smaller n if the change size is large enough we still obtain a good TPR. The change sizes Θ used here may seem much larger than those in Section 4.3.2,

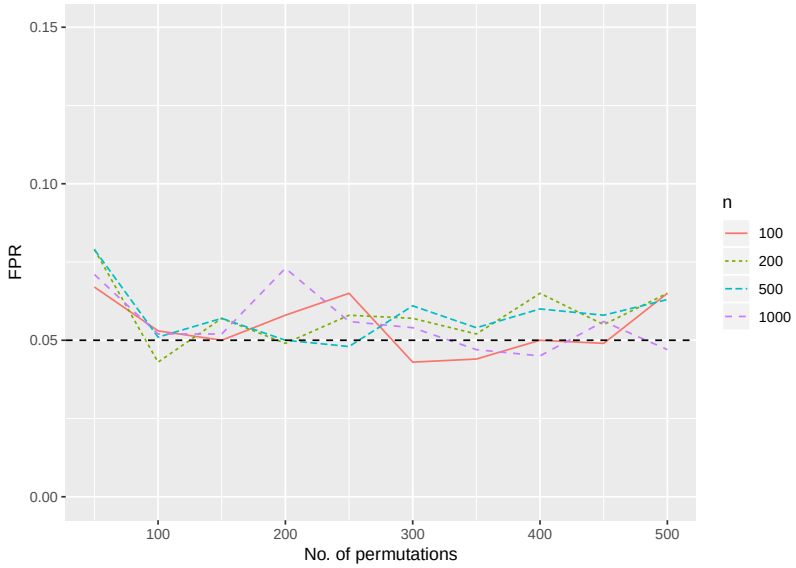


Figure 4.3.3: FPR for data containing no changepoints for varying n and number of permutations (P)

however, this is due to us now using the data-driven threshold rather than assuming σ_s^2 and σ_ϵ^2 are known which would be unlikely in practice. Note that for $n = 500$ we obtain a TPR of over 0.8 for $\Theta = 0.2$ and this change size is still smaller than any used in Jiao et al. (2018).

4.3.4 Method Comparison

We are not aware of any implementation of offline subspace changepoint detection which we can compare our method to. Hence, we compare our method to state-of-the-art covariance changepoint methods noting that subspace changepoints can be viewed as covariance changepoints where we impose conditions on the eigenvalues of the covariance matrix. As our method assumes the subspace dimension is known we would expect to outperform the covariance methods. Therefore, we will present

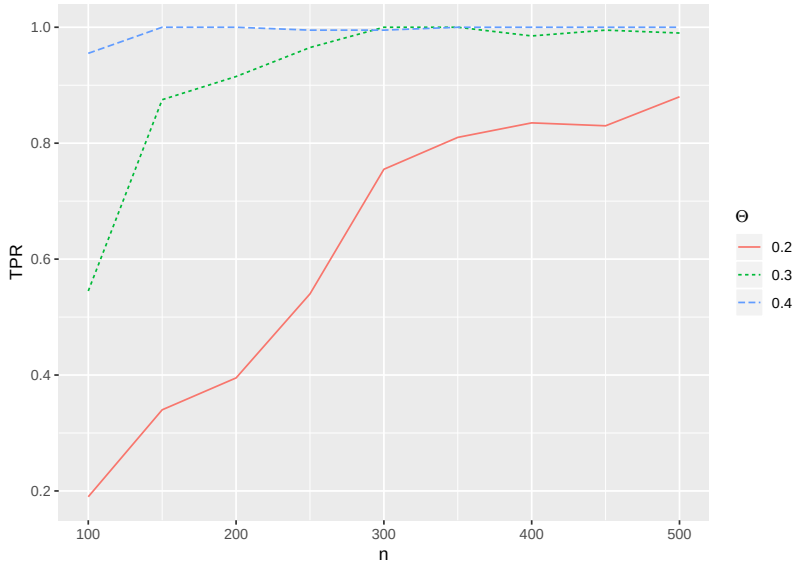


Figure 4.3.4: TPR for varying n and different Θ.

3 scenarios: one generic example and then two additional examples which are more suited to the covariance changepoint methods and show our method outperforms or is at least competitive.

We will compare our subspace changepoint method to that of Aue et al. (2009a), Avanesov and Buzun (2018) and Wang et al. (2021) and these methods will be referred to as *Aue*, *Av-Buz* and *Wang* throughout, while our method will be referred to as *Subspace*. The simulation studies performed in Aue et al. (2009a), Avanesov and Buzun (2018) and Wang et al. (2021) are limited meaning appropriate thresholds to determine significant changepoints are not obvious for each method. Furthermore, the choice of threshold for each method can drastically affect its perceived performance, therefore, to perform a fair comparison we will only consider the estimated changepoint locations assuming the data contains one changepoint. Hence, we don't need to choose appropriate thresholds for each method, we simply select the most likely

changepoint location for all regardless of significance.

We set the msl allowed for each method to $2p\log(n)$ as this is the msl allowed in *Wang* and is the largest among the methods. For the *Aue* method an estimate of the long-run covariance matrix is needed and we use the Bartlett estimator (Andrews, 1991) as suggested in Aue et al. (2009a). For the *Av-Buz* method a window size is needed and we use the minimum segment length $2p\log(n)$ as this is the minimum segment length in our simulation but is larger than the required minimum segment length for the method - we note similar results with varying window lengths. Additionally, *Av-Buz* requires a hyperparameter for the graphical lasso used in the method. We follow the advice given in Avanesov and Buzun (2018) and set this as $\sqrt{\frac{\log(p)}{n}}$.

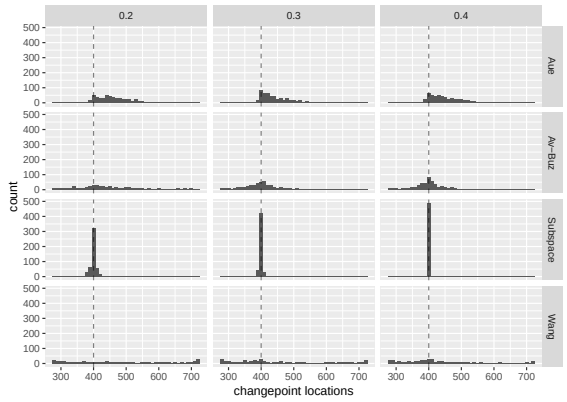
Scenario 1: For the first scenario, we simulate 1000 data sets for each of $\Theta \in \{0.2, 0.3, 0.4\}$ with a single changepoint at $\tau = 0.4n$ ($n = 1000$, $p = 20$, $q = 5$, $\sigma_s^2 = 1$, $\sigma_\epsilon^2 = 0.05$). Figure 4.5(a) shows the estimated changepoint locations in each of the methods over varying change sizes. Clearly, the *Subspace* method outperforms the other methods in terms of estimated changepoint location. The covariance changepoint methods only start to show slight peaks around the true changepoint for the largest change size.

Scenario 2: In the next scenario, we set the subspace dimension as $q = 1$ and keep the rest of the parameters the same as in Scenario 1, apart from the changepoint location which is now $\tau = 0.6n$. This setup should suit the *Wang* method as their test statistic uses the operator norm and therefore focuses on the first eigenvalue. Figure 4.5(b) shows the *Wang* method does perform much better in this scenario, however it

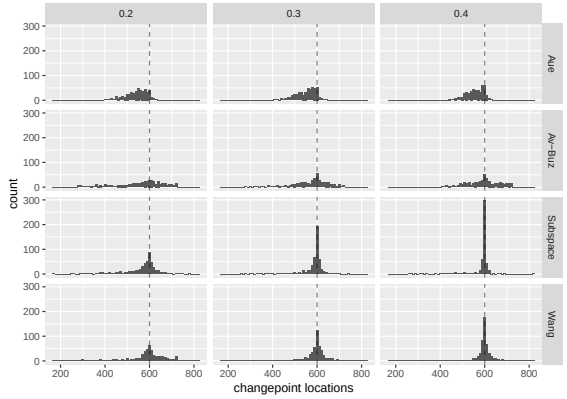
is still not as accurate as the *Subspace* method in detecting the changepoint location. Interestingly, the *Aue* method, which was right-skewed in Figure 4.5(a), is now left-skewed showing the method appears to be biased towards equal segment lengths if the changepoint location is not in the centre. Again the *Av-Buz* method only starts getting a slight peak around the true changepoint location for $\Theta = 0.4$.

Scenario 3: In the final scenario, we reduce the dimension of the data so that $p = 10$, $q = 5$ and $\tau = 0.6n$ - the rest of the parameters are the same as in Scenario 1. The *Aue* method is not designed to perform well in high-dimensional settings unlike the *Av-Buz* and *Wang* methods and as such we would expect it to perform better in this lower-dimensional setting. Figure 4.5(c) shows that the *Aue* method performs extremely well in this scenario and is comparable to the *Subspace* method. However, the *Aue* method cannot be used in higher-dimensional settings, which are becoming increasingly common in practice, as will be seen in Section 4.5.

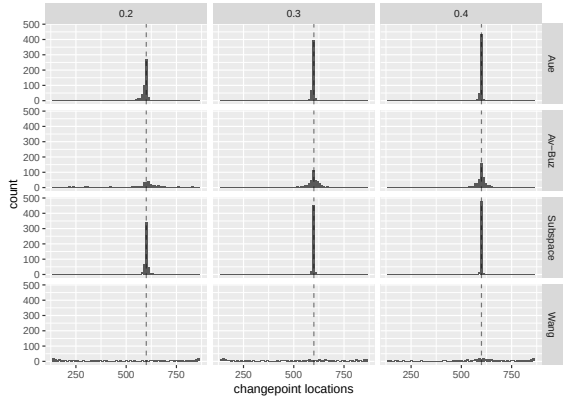
It is no surprise that Figure 4.3.5 shows our method outperforms the covariance changepoint methods as we assume we know the subspace dimension size q and our test statistic is specifically designed for this covariance structure. Yet, this does show a need for a subspace changepoint method to detect these types of changes as generic change in covariance methods are unable to locate the changes accurately across scenarios.



(a) Scenario 1



(b) Scenario 2



(c) Scenario 3

Figure 4.3.5: Estimated changepoint locations of the four methods for 3 different change sizes $(0.2, 0.3, 0.4)$ in the three scenarios described.

4.3.5 Sensitivity to Subspace Dimension

Throughout, we have assumed that the dimension of the subspace, q, is known. In certain applications we may not know the true subspace dimension, hence we explore the sensitivity of our test statistic to the assumed subspace dimension. We will generate data with a true subspace size and create ROC curves, similar to Section 4.3.2, assuming different subspace sizes. Let q be the true subspace dimension and $\hat{q}$ be the assumed subspace dimension. Figure 4.3.6 shows the ROC curves for varying $\hat{q}$ ($n = 200$, $p = 20$, $q = 5$, $\tau = 100$, $\sigma_s^2 = 1$ and $\sigma_\epsilon^2 = 0.05$).

Figure 4.3.6 clearly shows if we overestimate the subspace dimension this has a far less drastic effect on the power than if we underestimate the subspace dimension. This can be explained from our test statistic using the cost defined in (4.2.8). In the cost of a segment, we sum the last $p - q$ eigenvalues which correspond to the variance in the orthogonal complement of the subspace. Now if we underestimate q, there will be eigenvalues contributing to the cost that are large as they correspond to the variance in one or more directions within the subspace. Hence, the cost of the segment will be large. Now, even if we add a changepoint to get a better estimation of the true underlying subspaces the inflated costs, due to the large eigenvalues, will mean the test statistic still has minimal power. On the other hand, if we overestimate the true subspace then we have less of the eigenvalues that correspond to the variance in the orthogonal complement to the subspace, meaning when we add a change then the cost of the correct segmentations will still be low but the estimation of the union of the subspaces in the cost of the whole data will be improved meaning we have slightly

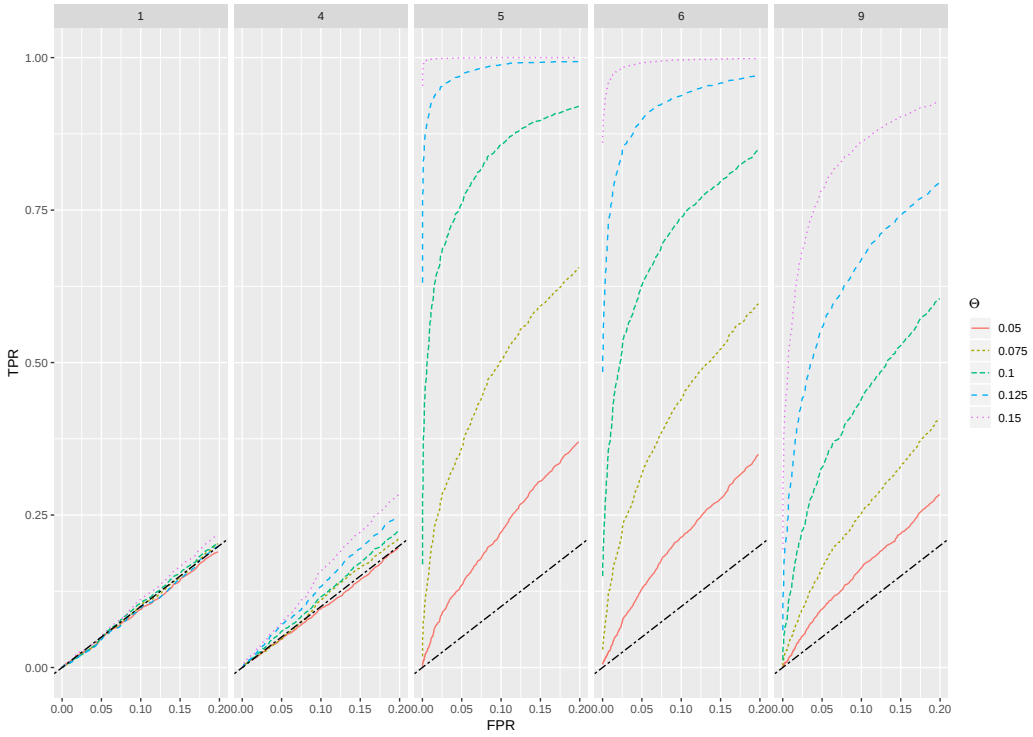


Figure 4.3.6: ROC curves for data with different assumed subspace dimensions. The true subspace dimension for all plots is 5 and each facet shows the ROC for different assumed subspace dimensions.

less power.

In applications, the separation of the eigenvalues corresponding to the subspace and it's orthogonal complement will likely be less explicit than in simulated examples. Therefore, the drop in power by underestimating the subspace dimension would be less dramatic. However, if the true underlying subspace is unknown then we would advise being cautious and overestimate the subspace dimension.

4.4 Extension: Multiple Changepoints

So far we have considered the AMOC scenario where we assume a maximum of one changepoint in a data set. In many situations, it is possible for data sets to contain multiple changepoints and hence we need a way to handle this. In the changepoint literature, there are two main approaches for handling multiple changepoints. One solution is to take a dynamic programming style approach (Auger and Lawrence, 1989; Jackson et al., 2005; Killick et al., 2012) where a cost function is defined for a segment and these algorithms find an exact minimization of the total cost of the data, given some penalty for adding a change. Choosing an appropriate penalty can be challenging and usually requires some strict assumptions placed on the cost function and data generating process. The alternative approach is to use a Binary Segmentation procedure (Scott and Knott, 1974; Vostrikova, 1981), which is a heuristic approach. Here a single changepoint is proposed in the data set, if this changepoint is significant we split the data and look for a changepoint on either side of the detected change. This approach makes it easy to extend our AMOC scenario to allow for multiple changepoints.

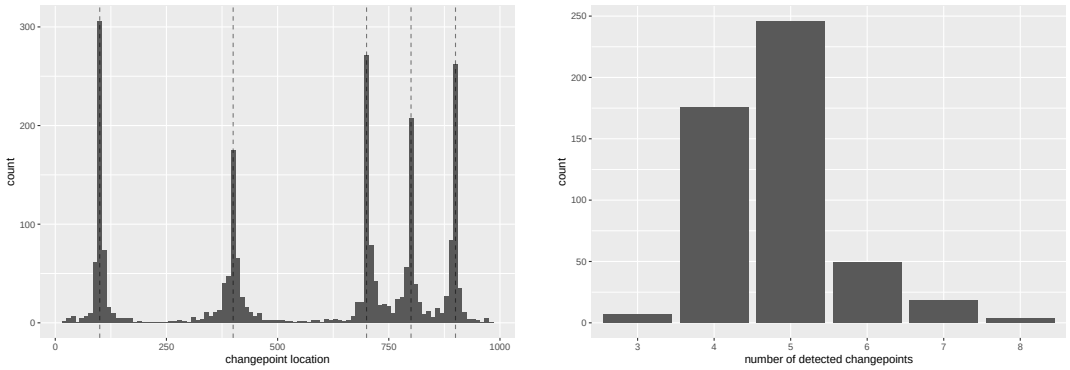
If the true number of changepoints, m, is known we can run the Binary Segmentation algorithm with our test statistic to gain the first m most significant changepoints. If m is unknown, we can perform the permutation test described in Section 4.2.3 to provide a stopping criterion. First, we find the most likely changepoint position in the data and perform the permutation test on the data to determine if the test statistic for this changepoint is significant. If significant, we then search for the

most likely changepoint position in the two segments on either side of the detected changepoint. We then perform a permutation test on the segment containing the most likely changepoint position. This process of splitting the data is repeated until the permutation test deems a changepoint insignificant in which case we stop and return all detected changepoints. Note that recent variations to Binary Segmentation (Fryzlewicz, 2014; Kovács et al., 2020) could equally be used but produce a larger FPR especially when the Binary Segmentation assumptions are satisfied (Shi et al., 2021).

4.4.1 Simulation

Here we explore how our method performs when detecting multiple changepoints using the Binary Segmentation approach. In order to generate multiple subspaces we generate $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$, as described in Section 4.3.1. Next, we find a basis that is orthogonal to $\boldsymbol{W}_2$ and use this in (4.3.1) to generate the next subspace. This process is iterated until the required number of subspaces is reached. The change sizes between each subspace are set by $\Delta = \Theta\sqrt{q}$. The data is then generated similarly to (4.3.2) for all segments.

We simulate a data set with $n = 1000$, $p = 20$, $q = 5$, $\sigma_s^2 = 1$, $\sigma_\epsilon^2 = 0.05$ and $\tau = (100, 400, 700, 800, 900)$ with corresponding change sizes $\Theta = (0.25, 0.15, 0.25, 0.25, 0.25)$. The number of changepoints was not known to our algorithm, hence, the permutation test stopping criteria detailed above was used to determine the number of changepoints. We used $P = 200$ random permutations and set $\alpha = 0.05$. We per-



(a) (b)

Figure 4.4.1: (a) Histogram of estimated changepoint locations with dashed lines showing the true changepoint locations. (b) Number of estimated changepoints in each repetition, the true number is 5.

formed 500 repetitions of this scenario and Figure 4.4.1 shows a histogram of the estimated changepoint locations and the number of estimated changepoints in each repetition.

Figure 4.1(a) shows that the detected changepoints are distributed around the true changepoint locations. The changepoint at $\tau = 400$ is detected with less accuracy due to the change being smaller despite the larger gap between the changepoints. Even the three changepoints that are close together still have three distinct peaks around the true changepoint locations which implies the Binary Segmentation implementation is working well. Additionally, Figure 4.1(b) shows the majority of repetitions returned either the correct number of changepoints or one fewer. This shows using the permutation test as a stopping criterion for the method performs well.

4.5 Application: Motion Capture Data

In this section, we aim to detect changepoints in the Carnegie Mellon Motion Capture dataset available at http://mocap.cs.cmu.edu. The data set was created by tracking multiple sensors placed on the human body while subjects performed different activities. The aim is to detect time points at which subjects change from one activity to another. We examine 56 of these sensors and focus on trial 1 for subject 86 which consists of activities including walking, jumping, punching and kicking. It is assumed that during the same activity the 56 series will lie in a subspace and when this activity changes we observe a change in subspace.

Before running our algorithm we standardised each series by subtracting the median and dividing by the median absolute deviation. We assume a subspace dimension of $q = 9$ and stored the 30 most significant changepoints. The raw permutation test is not appropriate to derive a threshold as the data contains temporal dependence. Hence, we examined the total cost of the data, $\sum_{k=1}^{m+1} \mathcal{C}(y_{\tau_{k-1}+1:\tau k})$ including each changepoint sequentially in terms of their significance. To find an appropriate number of changepoints we find the total cost whereby adding another changepoint results in only a minimal decrease in total cost - this implies that adding more changes does not significantly improve the model fit. Using this method we found $m = 6$ to be the most appropriate number of changepoints. Justification for this, and choosing the subspace dimension $q = 9$, is given in Appendix B.3.

Figure 4.5.1 shows the location of the 6 identified changepoints on a range of se-

Figure 4.5.1: Selection of series relating to different sensors with identified changepoints shown by red solid vertical lines and transition windows shown by dashed vertical lines.

ries relating to sensors from different parts of the subject. Additionally, one-second windows are shown which correspond to the subject transitioning between activities, these were independently identified by eye before analysis. As can be seen, the 6 identified changepoints fall within these one-second windows showing the method has successfully segmented the data into the different activities.

4.6 Conclusion

A novel method for detecting changepoints in potentially high-dimensional time series that are assumed to lie in low-dimensional subspaces has been presented. We have provided a consistent way to estimate the latent subspaces, determine if a subspace

changepoint has occurred in the data and find its most likely position. A data-driven threshold for determining the significance of a potential changepoint is provided and we discussed an extension of the methodology to the multiple changepoint scenario. Additionally, an extensive simulation study in the AMOC and multiple changepoint scenarios is presented that displays the effectiveness of our method in both settings. Moreover, we have shown the need for a subspace changepoint method through comparison to current state-of-the-art covariance changepoint methods. Finally, we demonstrated our approach can detect activity changes in Motion Capture data.

The current methodology has some limitations which could provide future research opportunities. Firstly, a thorough theoretical exploration may discover a theoretical threshold that could be used as an alternative to the data-driven threshold. The major assumption in the paper is the knowledge of the underlying subspace dimension; finding a way to automate the identification of the subspace dimension would be of interest in addition to allowing the subspace dimension to vary between changepoint segments.

The methodology presented in this Chapter is implemented in the R package *changepoint.cov* available on GitHub `https://github.com/grundy95/changepoint.cov`.

Chapter 5

Identifying Sequential Changes in Mean and Variance Within More Complex Model Structures

5.1 Introduction

Accurate forecasting is crucial for planning and marketing decisions for many companies. For example, within inventory management, the optimization of stock levels crucially depends upon accurate supply and demand forecasts (see, for example reviews by Fildes, 1985; Fildes et al., 2008; Syntetos et al., 2016). In hospitals, accurate forecasts are crucial for bed management (Jones et al., 2002) and ensuring adequate numbers of clinicians, nurses and equipment in A&E departments (Ordu et al., 2020). In general, with accurate forecasts, informed decisions can be made but what happens

if over time these forecasts become poor? If forecasts (or prediction intervals based upon them) start performing poorly this can have severe consequences, starting from product shortages in the retail sector, to a shortage of medical supplies in hospitals. Hence, it is crucial to quickly identify when forecasting models start deteriorating in performance, in order to react and make necessary adjustments.

A common cause of forecasts becoming inaccurate is due to a changepoint (also known as a structural break) occurring in the underlying process. Throughout we call the data generated from the underlying process the raw data. After a changepoint in the raw data, the current forecasting model may become inappropriate and the model needs rebuilding. Without the knowledge of changepoints, it is disputed how long forecasting models can be used before updating (Krasheninnikov et al., 2018). If a model is created and left to run for long periods of time without regular checks, practitioners could be making key decisions based upon inaccurate information. Alternatively, if parameters are frequently updated then changepoints that affect the choice of forecasting model would be missed. Furthermore, the updated parameters may be based on data containing a changepoint and therefore the estimation will be contaminated by the pre-change data thus damaging the resulting forecasts (Chapman and Killick, 2020).

A common approach to identify changes is to add a changepoint component to the modelling paradigm; this is easier in some paradigms than others. Even when it is clear how to add changes to our modelling paradigm, if our model is complex (for example containing seasonality, temporal dependence and/or trends) then our ability

to identify a change quickly deteriorates with increasing complexity. In this paper, we propose a framework for sequential changepoint detection within forecast errors to detect if a model becomes inaccurate. As mentioned, the most likely cause for forecasts becoming inaccurate is a change in the raw data being forecast. A change in the mean of the raw data could result in bias in the forecast and this would show as a mean change in the forecast errors. A change in the variance of the raw data, may not cause bias in the forecasts but could lead to miss-calibrated prediction intervals. This change would be reflected by a variance change in the forecast errors. Hence, by detecting changes in the forecast errors, we can automatically identify when the forecasting model needs to be reconstructed and/or re-estimated. This eliminates the need for manual checks of the forecasting models and allows practitioners to focus on other tasks.

Sequential changepoint analysis aims to detect changes in a data generating process in an online manner. As more data becomes available, they are sequentially checked to determine if the distribution of the data has changed. Sequential changepoint analysis dates back to Page (1954) and traditionally aims to identify changes in the mean of the data generating process under the assumption of independent observations and constant variance, see Basseville and Nikiforov (1993), Csörgö and Horváth (1997) and Tartakovsky et al. (2014) for overviews. Our approach to sequential changepoint detection stems from the work of Chu et al. (1996), where sequential changepoint procedures are considered in two phases. In phase 1, we assume a fixed amount of data is readily available and is generated from the same distribution; we call this

the training data. The training data is used to estimate the data generating process. Sequential monitoring begins in phase 2; i.e, when the forecasting model goes live. As more data becomes available we make a decision, based upon some rule, checking whether the new data is still being generated from the same distribution as Phase 1. If the new data significantly differs from the data in Phase 1, then we deem a changepoint to have occurred and stop monitoring. Hence, here our model is trained once using the fixed amount of training data and is not re-trained; either a changepoint is detected and the monitoring stops or the monitoring continues indefinitely. Note that the proposed approach can be complimentary to the existing forecasting techniques in practice: while the main model used for decision making can be re-estimated as soon as the new data arrives, the secondary model used for changepoint detection would need to be estimated only once (initially), and its output can be tracked independently of the first model.

The work of Chu et al. (1996) has been extended in various ways including sequential monitoring of linear models (Horváth et al., 2004); autoregressive processes (Gombay and Serban, 2009); and financial time series (Aue et al., 2009b). Here we utilize sequential changepoint methods to detect changes in forecast errors to quickly identify when forecasts become inaccurate. In particular, we examine one-step-ahead forecast errors, which, within models with an additive error structure, are the model residuals. There has been previous work utilizing the residuals of a model to detect changepoints. In particular, Horváth et al. (2004) use the residuals from estimated linear models to detect changes in regression co-efficients. Furthermore, Aue et al. (2015) use the

residuals from estimated ARMA processes to detect changes in mean, variance or regressive parameters. Moreover, within the streaming classification literature there are numerous methods that use classification errors to detect when changes have occurred (Gama et al., 2004; Ross et al., 2012). In the offline changepoint setting where all data is available a priori, Robbins et al. (2011) proposed using residuals within the CUSUM procedure and showed that this gave an improved performance over using the raw data when temporal dependence was present in the data. All the methods mentioned above, assume a specified, known forecasting model. Our framework is far more general as any forecasting approach could be used, including model-based, machine learning and expert judgement forecasts - we only require the forecast errors.

The novelty in our work is two-fold. From a forecasting viewpoint, by combining forecasts and sequential changepoint detection we create a novel framework for automatically identifying when a forecasting model needs re-evaluating. Moreover, we show theoretically how some common changes in raw data (which are a likely cause of inaccurate forecasts) manifest in the forecast errors and therefore can be detected within our framework. From a sequential changepoint viewpoint, this methodology can provide an improved performance over model-based sequential changepoint techniques. If the data generating process is complex, then large amounts of data are needed to get an adequate model. Hence, after a change has occurred, we would expect a longer detection delay before signalling a change as more data is required to model the post change distribution. In contrast, performing sequential changepoint

detection on the forecast errors is a much simpler task as the complex modelling issues are eradicated by the forecasting model.

The paper proceeds as follows. In Section 5.2, we introduce the sequential changepoint detection framework and significant results from the literature that are used in this paper. Building on this, in Section 5.3 we introduce our framework for monitoring forecast errors and show theoretically how different changes in potentially complex data models manifest in these forecast errors. Section 5.4, shows how two common forecasting models fit into our framework before Section 5.5, demonstrates the improved performance of monitoring the forecast errors over monitoring the raw data in a number of simulated examples. Finally, Section 5.6 shows the effectiveness of our method in two applications, one from NHS A&E admissions data and the other from Royal Mail delivery volumes, before Section 5.7 gives concluding remarks and suggests some future research directions.

5.2 Sequential Changepoint Detection

Before proceeding with our framework, we first introduce the sequential changepoint detection paradigm used in this paper for identifying a mean change in a generic sequence of data $\{Y_t : t = 1, 2, \ldots\}$. We present the general setup introduced in Chu et al. (1996), along with the detector used for detecting changes; Page's CUSUM detector (Page, 1954). Moreover, we highlight the key large-sample results from the literature which we utilize within our framework.

Let $\{Y_t : t = 1, 2, \ldots\}$ be a univariate time series that follows a change in mean model,

$$Y_t = \begin{cases} \lambda + \epsilon_t, & t = 1, \ldots, m + k^*, \\ \lambda + \epsilon_t + \Delta_m, & t = m + k^* + 1, m + k^* + 2, \ldots \ . \end{cases} \tag{5.2.1}$$

Here $\lambda \in \mathbb{R}$ is the pre-change mean level; $\Delta_m \in \mathbb{R}$ is the size of the mean change; $1 \leq k^* \leq \infty$ is the time of change after the training period, and $\{\epsilon_t : t = 1, 2, \ldots\}$ are zero-mean random variables. Here $m \in \mathbb{N}$ is the length of the training period, which is used to estimate model parameters and we assume the data within the training period have no changepoints and exhibit a stationary structure. All asymptotics considered are with respect to $m \to \infty$ throughout. Note the model is trained once using the m data points and is not re-trained once the monitoring period has begun.

To be as general as possible, we allow for the random variables $\{\epsilon_t : t = 1, 2, \ldots\}$ to exhibit different dependence structures, hence we assume they satisfy some weak invariance principles.

Assumption 5.2.1. $\left| \sum_{t=1}^{m} \epsilon_t \right| = \mathcal{O}_P\left(\sqrt{m}\right)$.

Assumption 5.2.2. *There exists a sequence of Wiener processes* $\{W_m(i) : i \geq 0, m \geq 1\}$ *and a constant* $\sigma \in \mathbb{R}^+$ *such that*

$$\sup_{1/m \leq i < \infty} \frac{1}{(mi)^{1/\nu}} \left| \sum_{t=m+1}^{m+mi} \epsilon_t - \sigma W_m(mi) \right| = \mathcal{O}_P(1) \quad \textit{for some } \nu > 2 \ .$$

Aue and Horváth (2004) give examples of sequences that satisfy Assumptions 5.2.1 and 5.2.2. Importantly for our framework, we show that the random variables will be i.i.d and therefore will satisfy Assumptions 5.2.1 and 5.2.2.

To detect a change in mean, we test the hypothesis,

$$H_0 : \Delta_m = 0 \,, \tag{5.2.2}$$

$$H_A : \Delta_m \neq 0 \,, \tag{5.2.3}$$

and define a stopping rule τ_m such that

$$\lim_{m\to\infty} \mathbb{P}\left(\tau_m < \infty\right) = \alpha \quad \text{under } H_0 \,,$$

$$\lim_{m\to\infty} \mathbb{P}\left(\tau_m < \infty\right) = 1 \quad \text{under } H_1 \,,$$

where α is the user-specified false alarm rate. Note here that by assuming that we have a large number of training samples, we control the false alarm rate indefinitely. This differs from the traditional, control chart based, sequential changepoint detection where we generally set a false alarm rate based upon the average run length of the process (Tartakovsky et al., 2014).

To define Page's CUSUM detector, $D(m,k)$, we first define the CUSUM detector, $Q(m,k)$. Let $k \in \mathbb{N}$, be the current time point, then

$$Q(m,k) = \sum_{t=m+1}^{m+k} Y_t - \frac{k}{m}\sum_{t=1}^{m} Y_t \,.$$

Using this we gain

$$D(m,k) = \max_{0\leq i\leq k} \left|Q(m,k) - Q(m,i)\right| , \tag{5.2.4}$$

and define the correspond stopping time as

$$\tau_m = \min\left\{k \geq 1 : D(m,k) \geq \hat{\sigma}_m c_\alpha g(m,k,\gamma)\right\} .$$

Here, $\hat{\sigma}_m$ is a weakly consistent estimator for σ in Assumption 5.2.2 and is estimated using the m training samples. The critical value, c_α, is a constant derived from the

large sample distribution of the stopping time under H_0 and $g(m, k, \gamma)$ is a weight function defined as

$$g(m, k, \gamma) = \sqrt{m}\left(1 + \frac{k}{m}\right)\left(\frac{k}{m+k}\right)^{\gamma} \quad \text{for } \gamma \in [0, 1/2)\ . \tag{5.2.5}$$

The hyper-parameter γ allows a practitioner to tune the stopping time dependent upon when the change is likely to take place. If the change is likely to occur early in the process then increasing γ will lead to a faster detection time, however, if the change occurs later in the process then having a larger γ will decrease the detection time. Throughout this paper, we assume $\gamma = 0$ for simplicity but for a more detailed discussion, see Horváth et al. (2004).

The critical constant, c_α, can be derived from the limiting distribution of the stopping time under H_0 shown in Fremdt (2015).

Theorem 5.2.3. *(Fremdt, 2015)*

Let $\{Y_t : t \geq 1\}$ *follow* (5.2.1) *and Assumptions 5.2.1 and 5.2.2 hold. Then under* H_0*, for* $c \in \mathbb{R}$*,*

$$\lim_{m\to\infty} \mathbb{P}\left(\frac{1}{\hat{\sigma}_m} \sup_{1\leq k<\infty} \frac{D(m,k)}{g(m,k,\gamma)} \leq c\right) = \mathbb{P}\left(\sup_{0<i<1} \sup_{0\leq j\leq i} \frac{1}{i^\gamma} \left|W(i) - \frac{1-i}{1-j} W(j)\right| \leq c\right)\ ,$$

where $\{W(i) : i \in [0, 1]\}$ *denotes a standard Brownian motion.*

Under H_A, Fremdt (2014) derived the limiting distribution of the stopping time, τ_m, under certain additional assumptions.

Assumption 5.2.4. *There exists a* $\theta > 0$ *such that the changepoint* $k^* = \lfloor \theta m^\beta \rfloor$ *with* $0 \leq \beta < 1$.

Assumption 5.2.5. $\sqrt{m}\,|\Delta_m| \to \infty$ *as* $m \to \infty$.

Assumption 5.2.6. $\Delta_m = \mathcal{O}(1)$.

These assumptions ensure the change size is bounded and is large enough relative to the size of the training data and the time of the changepoint. For a detailed discussion of these assumptions see Fremdt (2014).

The form of the large sample distribution under H_A depends on the asymptotic behavior of the sequence $|\Delta_m| m^{\gamma-1/2} k^{*1-\gamma}$, hence, due to Assumption 5.2.4, depends on the asymptotic behavior of the term

$$\tilde{\Delta}_m = |\Delta_m| \, m^{\beta(1-\gamma)-1/2+\gamma} \, .$$

This leads to three scenarios as $m \to \infty$:

(I) $\tilde{\Delta}_m \to 0$

(II) $\tilde{\Delta}_m \to \tilde{C} \in (0, \infty)$

(III) $\tilde{\Delta}_m \to \infty$

For scenario (II), from Assumption 5.2.4, we have

$$|\Delta_m| m^{\gamma-1/2} k^{*1-\gamma} \to \theta^{1-\gamma} \tilde{C} = C \in (0, \infty) \, .$$

Here, for any real c, we define $d = d(c)$ as the unique solution of

$$d = 1 - \frac{c}{C} d^{1-\gamma} \, .$$

The large sample distribution of the stopping time depends upon the three scenarios (I)-(III). To define the limit distribution, the function $\tilde{\Psi}(x)$ is introduced for all real

x,

$$\tilde{\Psi}(x) = \begin{cases} \Phi(x) & \text{under scenario (I)} \\ \mathbb{P}\left(\sup\limits_{d<i<1} W(i) \leq x\right) & \text{under scenario (II)} \\ \mathbb{P}\left(\sup\limits_{0<i<1} W(i) \leq x\right) = \begin{cases} 0 & x < 0 \\ 2\Phi(x) - 1 & x \geq 0 \end{cases} & \text{under scenario (III)} \end{cases}$$

where $\Phi(x)$ denotes the standard Normal distribution function.

Theorem 5.2.7. *(Fremdt, 2014)*

Let $\{Y_t : t = 1, 2, \ldots\}$ *be a sequence of random variables according to* (5.2.1) *that satisfies Assumptions 5.2.1-5.2.6. Then, for all real* x *under* H_A *and with* $\Psi(x) = 1 - \bar{\Psi}(-x)$,

$$\lim_{m\to\infty} \mathbb{P}\left(\frac{\tau_m - a_m(c_\alpha)}{b_m(c_\alpha)} \leq x\right) = \Psi(x) \ ,$$

where $a_m(c)$ *is the unique solution of*

$$a_m(c) = \left(\frac{\sigma c m^{1/2-\gamma}}{|\Delta_m|} + \frac{k^*}{(a_m(c))^\gamma}\right)^{1/(1-\gamma)} \ ,$$

and

$$b_m(c) = \frac{\sigma\sqrt{a_m(c)}}{|\Delta_m|}\left(1 - \gamma\left(1 - \frac{k^*}{a_m(c)}\right)\right)^{-1} \ .$$

This allows us to quantify the size of changepoints that can be detected within a certain time frame and shows that given a certain change size, asymptotically in m, we will always detect certain changes eventually.

5.3 Monitoring Forecast Errors

The main aim of this paper is to introduce a framework that can quickly identify if a forecasting model becomes inaccurate. A common cause of inaccurate forecasts is a change in the potentially complex data structure. For simple data structures, we can detect mean changes by performing sequential changepoint detection on the data, namely Page's CUSUM detector, as described in Section 5.2. We can also adapt this detector to detect variance changes by monitoring the (centered) squared data as described in Inclan and Tiao (1994). Yet, it is increasingly common for data to exhibit more complex patterns than this setting allows. Some common examples of such structure include seasonality and trend with the data. Moreover, if the data exhibits temporal dependence then convergence to the limiting distributions in Section 5.2 are slow and the performance of the detector deteriorates. There has been some work to identify changepoints in some specific complex models such as data exhibiting trend (Horváth et al., 2004) and ARMA models (Aue et al., 2015), however, our work is more general and encompasses these models and more by allowing for a generic forecasting model.

Alternatively, model-based sequential changepoint techniques could be used for more complex models. These fit a model to the data before and after a potential changepoint and use a likelihood ratio style approach to determine if a changepoint has occurred. Even if the model is known this approach has a number of drawbacks. Firstly, we need to have at least as many observations as parameters in the model. This means if we have seasonality in the model, we require at least one full season worth of data just

to fit the post-change model; this can greatly increase the detection delay before a change is identified. Additionally, as we get more data points, there are more potential changepoint locations to check for a change, thus increasing the computational time of the method. Finally, the choice of threshold is non-trivial, so identifying significant changepoints while controlling the false alarm rate is challenging.

Our framework bypasses the above issues by monitoring forecast errors, instead of the raw data, for changepoints. We show theoretically that mean and variance changes in potentially complex data structures will manifest in the forecast errors (under certain assumptions on the forecasting model) and adapt Page's CUSUM detector (Section 5.2) to detect the manifested changes in the forecast errors. The detection of a mean change in the forecast errors indicates that the forecasting model needs re-evaluating as it has become biased, while a variance change indicates that prediction intervals from the forecasting model will no longer be precise and would need to be re-constructed.

Let $\{Y_t : t = 1, 2, \ldots\}$ be a univariate time series following some unknown distribution, F, with an associated forecasting model, $\mathcal{F}$. Assuming a known forecasting model, $\mathcal{F}$, we define the point forecast for time point t made h time steps previously as,

$$\hat{y}_t(h) = \mathbb{E}_{\mathcal{F}}\left[Y_t | y_{t-h}, \ldots, y_1\right]. \tag{5.3.1}$$

Without loss of generality, we focus on one-step-ahead forecasts, hence $\hat{y}_t(1)$ is the forecast for time point t made at time point $t-1$. Using (5.3.1), we define the forecast errors, e_t, as

$$e_t = Y_t - \hat{y}_t(1).$$

First, we consider theoretically how mean changes in the potentially complex data structures manifest in the forecast errors. We highlight the desirable properties of the forecast errors and describe how we can incorporate them into Page's CUSUM detector. Secondly, we consider the more general setting where mean and/or variance changes in the raw data can occur. Again, we show theoretically how these changes manifest in the forecast errors and describe an alternative adaption of Page's CUSUM detector to detect these changes.

5.3.1 Mean Changes

First, we consider how mean changes in potentially complex data structures manifest in the forecast errors. Assume there is a changepoint at some unknown location k^*,

$$\begin{aligned} Y_t|y_{t-1},\ldots,y_1 &\sim F\ , \quad \text{for } t = 1,\ldots,m+k^*\ , \\ Y_t|y_{t-1},\ldots,y_1 &\sim G\ , \quad \text{for } t = m+k^*+1, m+k^*+2,\ldots\ , \end{aligned} \tag{5.3.2}$$

where F and G differ in expectation,

$$\mathbb{E}_G\left[Y_t|t > m+k^*, y_{t-1},\ldots,y_1\right] - \mathbb{E}_F\left[Y_t|t \leq m+k^* y_{t-1},\ldots,y_1\right] = \delta_{\mu,m} \neq 0. \tag{5.3.3}$$

To ease notation, we write $\tilde{y}_t$ to represent the whole past upon which Y_t is dependent. Moreover, the relevant pre- and post-changepoint times are inferred from the distribution we are taking expectations with respect to. Note, (5.3.2) encompasses the mean change model in (5.2.1) from Section 5.2 and additionally allows the raw

data to have many different properties including, but not limited to, being temporally dependent, exhibiting trends or containing seasonality.

For the mean change in (5.3.3) to manifest in the forecast errors, we must place the following assumptions on the forecasting model being used.

Assumption 5.3.1. *We have a forecasting model, $\mathcal{F}$, such that*

$$\mathbb{E}_F\left[Y_t|\tilde{y}_t\right] - \mathbb{E}_{\mathcal{F}}\left[Y_t|\tilde{y}_t\right] = b_\mu \quad \text{for } t = 1, \ldots, m + k^* \,,$$

where b_μ is some unknown (potentially zero) bias in the forecasts.

Assumption 5.3.2. *For a forecasting model, $\mathcal{F}$, and data that follows* (5.3.2) *then,*

$$\mathbb{E}_G\left[Y_t|\tilde{y}_t\right] - \mathbb{E}_{\mathcal{F}}\left[Y_t|\tilde{y}_t\right] = b_\mu + f(\delta_{\mu,m}) = b_\mu + \Delta_{\mu,m} \quad \text{for } t = m + k^* + 1, m + k^* + 2, \ldots, \,,$$

where $\Delta_{\mu,m} \neq 0$ and is some function of the change size in the raw data.

For generality, these assumptions state that, in the pre-change regime, the forecasting model has a constant bias, b_μ, which in practise would ideally be equal to zero. When a change in the raw data occurs, many forecasting models will adjust in some way to try and account for this change. For example, in an AR(1) model with autoregressive parameter, ϕ, a change size of $\delta_{\mu,m}$ in the raw data would result in a change size of $\Delta_{\mu,m} = \delta_{\mu,m}(1 - \phi)$ in the forecast errors after one time point; assuming a perfectly specified forecasting model. Assumption 5.3.2 states that, like the AR(1) example above, the forecasting model cannot fully adjust to the change in the raw data and some constant change size still exists. This seems intuitive because if the forecasting model fully adjusts to the changepoint then we would not want to signal a changepoint as the forecasting model has adapted to the change and is still performing well.

Under Assumptions 5.3.1 and 5.3.2, the change in mean in (5.3.2) will result in a change in expectation in the forecast errors that is similar to (5.2.1),

$$Y_t - \hat{y}_t(1) = e_t = \begin{cases} b_\mu + \epsilon_{\mu,t}, & t = 1, \ldots, m + k^*, \\ b_\mu + \epsilon_{\mu,t} + \Delta_{\mu,m}, & t = m + k^* + 1, m + k^* + 2, \ldots \ . \end{cases}$$

with relevant null and alternative hypothesis defined in (5.2.2) and (5.2.3) with $\Delta_m = \Delta_{\mu,m}$.

As the forecasting model accounts for the dependence in the raw data, the zero-mean random variables $\{\epsilon_{\mu,t} : t = 1, 2, \ldots\}$ will be i.i.d. Note, even if the forecasting model is misspecified, the general assumptions on $\{\epsilon_t : t = 1, 2, \ldots\}$ given in Assumptions 5.2.1 and 5.2.2 are likely to hold.

To detect the manifested mean change in the forecast errors, we use Page's CUSUM detector, defined in (5.2.4), on the forecast errors,

$$Q_\mu(m, k) = \sum_{t=m+1}^{m+k} e_t - \frac{k}{m} \sum_{t=1}^{m} e_t \ ,$$

$$D_\mu(m, k) = \max_{0 \le i \le k} |Q_\mu(m, k) - Q_\mu(m, i)| \ ,$$

with corresponding stopping time,

$$\tau_{\mu,m} = \min \{k \ge 1 : D(m, k) \ge \hat{\sigma}_{\mu,m} c_{\mu,\alpha} g(m, k, \gamma)\} \ .$$

As the random variables, $\{\epsilon_t : t = 1, 2, \ldots\}$, are i.i.d, $\hat{\sigma}_{\mu,m}$, can be estimated using the traditional estimate of the standard deviation of the forecast errors within the training period. The weight function, $g(m, k, \gamma)$ is defined as in (5.2.5) and the critical constant, $c_{\mu,\alpha}$ can be derived from the limiting distribution of the stopping time which is known due to the following corollary.

Corollary 5.3.3. *From Theorem 5.2.3*

Let $\{Y_t : t \geq 1\}$ *follow* (5.3.2) *and* $\mathcal{F}$ *be a forecasting model that satisfies Assumption 5.3.1 and 5.3.2. Under* H_0*, the limit distribution in Theorem 5.2.3 holds for* $D(m,k) = D_\mu(m,k)$ *and* $\hat{\sigma}_m = \hat{\sigma}_{\mu,m}$.

Furthermore, the limiting distribution of $\tau_{\mu,m}$ under H_A is also known from the following corollary.

Corollary 5.3.4. *From Theorem 5.2.7*

Let $\{Y_t : t \geq 1\}$ *follow* (5.3.2) *and* $\mathcal{F}$ *be a forecasting model that satisfies Assumption 5.3.1 and 5.3.2. Furthermore, assume* $\Delta_{\mu,m}$*, from Assumption 5.3.2, satisfies Assumptions 5.2.4-5.2.6, then under* H_A*, the limit distribution in Theorem 5.2.7 holds with* $\tau_m = \tau_{\mu,m}$ *and* $c_\alpha = c_{\mu,\alpha}$.

These corollaries state that as long as we have a forecasting model that satisfies the given assumptions, a mean change in potentially complex data structures will manifest as mean changes in the forecast errors and will be detectable under our framework.

5.3.2 Mean and Variance Changes

So far we only have considered detecting mean changes using Page's CUSUM detector. Inclan and Tiao (1994) extend this to detect variance changes by considering the CUSUM detector of the (centered) squared data. This can be viewed as searching for a mean change in the variance estimates of the data. Here we employ a similar idea

and show that mean and/or variance changes in the raw data will manifest as mean changes in the (centered) squared forecast errors.

We extend the mean change model in (5.3.2) to allow for mean and variance changes. Again assume there is a changepoint at some unknown location, k^*,

$$\begin{aligned} Y_t|y_{t-1},\ldots,y_1 &\sim F\ , \quad \text{for } t=1,\ldots,m+k^*\ , \\ Y_t|y_{t-1},\ldots,y_1 &\sim G'\ , \quad \text{for } t=m+k^*1, m+k^*+2,\ldots\ , \end{aligned} \tag{5.3.4}$$

where F and G' differ in expectation and variance,

$$\mathbb{E}_{G'}\left[Y_t|t>m+k^*, y_{t-1},\ldots,y_1\right] - \mathbb{E}_F\left[Y_t|t\leq m+k^* y_{t-1},\ldots,y_1\right] = \delta_{\mu,m}\ ,$$

$$\mathrm{Var}_{G'}\left[Y_t|t>m+k^*, y_{t-1},\ldots,y_1\right] - \mathrm{Var}_F\left[Y_t|t\leq m+k^* y_{t-1},\ldots,y_1\right] = \delta_{\xi,m}.$$

Under Assumption 5.3.1, the change in mean and/or variance will result in a mean change in the squared forecast errors,

$$(Y_t-\hat{y}_t(1)-b_\mu)^2 = (e_t-b_\mu)^2 = \begin{cases} b_\xi+\epsilon_{\xi,t}, & t=1,\ldots,m+k^*, \\ b_\xi+\epsilon_{\xi,t}+\Delta_{\xi,m}, & t=m+k^*+1, m+k^*+2,\ldots\ . \end{cases}$$

where $b_\xi=\mathbb{E}[\epsilon_{\mu,t}^2]$ and $\{\epsilon_{\xi,t}: t=1,2,\ldots\}$ are zero-mean random variables. Again, due the forecasting model accounting for the potentially complex dependence structure in the raw data, $\{\epsilon_{\xi,t}: t=1,2,\ldots\}$, will be i.i.d.

The change size in the (centered) squared forecast errors, $\Delta_{\xi,m}$, can be shown to be a combination of the mean and/or variance change in the raw data.

Proposition 5.3.5. *Let* $\{Y_t: t=1,2,\ldots\}$ *follow* (5.3.4) *with* $\delta_{\mu,m}=0$ *and* $\mathcal{F}$ *be a forecasting model that satisfies Assumption 5.3.1. Then,*

$$\Delta_{\xi,m} = \mathrm{Var}_{G'}\left[Y_t|\tilde{y}_t\right] - \mathrm{Var}_F\left[Y_t|\tilde{y}_t\right]\ . \tag{5.3.5}$$

Moreover, if $\delta_{\mu,m} \neq 0$ *and Assumption 5.3.2 holds, then*

$$\Delta_{\xi,m} = \operatorname{Var}_{G'}\left[Y_t|\tilde{y}_t\right] - \operatorname{Var}_{F}\left[Y_t|\tilde{y}_t\right] + \Delta^2_{\mu,m} \; . \tag{5.3.6}$$

Proof. We start by noting that the change size, $\Delta_{\xi,m}$, is the difference between the expectation of the squared forecast errors assuming the post change distribution G' and the pre-change distribution F. From this, we can show the equivalences given in Proposition 5.3.5 by substituting relevant definitions of the forecast errors and using Assumption 5.3.1. □

To detect the manifested changes in the centered squared forecast errors, we can again adapt Page's CUSUM detector,

$$Q_\xi(m,k) = \sum_{t=m+1}^{m+k} \left(e_t - \hat{b}_{\mu,m}\right)^2 - \frac{k}{m}\sum_{t=1}^{m}\left(e_t - \hat{b}_{\mu,m}\right) \; ,$$

$$D_\xi(m,k) = \max_{0\leq i\leq k} \left|Q_\xi(m,k) - Q_\xi(m,i)\right| \; ,$$

with corresponding stopping time,

$$\tau_{\xi,m} = \min\left\{k \geq 1 : D_\xi(m,k) \geq \hat{\sigma}_{\xi,m} c_{\xi,\alpha} g(m,k,\gamma)\right\} \; .$$

Here $\hat{b}_{\mu,m}$ is the mean estimate of the forecast errors in the m training samples and $\hat{\sigma}_{\xi,m}$ is the traditional estimate of the standard deviation of the centered squared forecast errors within the training period. Again, $g(m,k,\gamma)$ is defined as in (5.2.5) and the critical constant $c_{\xi,\alpha}$ can be derived from the limiting distribution of the stopping time under H_0, which is known due to the following corollary.

Corollary 5.3.6. *From Theorem 5.2.3*

Let $\{Y_t : t = 1, 2, \ldots\}$ *follow* (5.3.4) *and* $\mathcal{F}$ *be a forecasting model that satisfies Assumption 5.3.1. Under* H_0*, the limit distribution in Theorem 5.2.3 holds for* $D(m, k) = D_\xi(m, k)$ *and* $\hat{\sigma}_m = \hat{\sigma}_{\xi,m}$.

Using Proposition 5.3.5, we gain the limiting distribution of $\tau_{\xi,m}$ under H_A from the following proposition.

Proposition 5.3.7. *Let* $\{Y_t : t = 1, 2, \ldots\}$ *follow* (5.3.4) *with* $\delta_{\mu,m} = 0$ *and* $\mathcal{F}$ *be a forecasting model that satisfies Assumption 5.3.1. Furthermore, assume* $\Delta_{\xi,m}$ *in* (5.3.5) *from Proposition 5.3.5 satisfies Assumptions 5.2.4-5.2.6. Then under* H_A*, the limit distribution in Theorem 5.2.7 holds with* $\tau_m = \tau_{\xi,m}$ *and* $c_\alpha = c_{\xi,\alpha}$*. Moreover, the limit distribution holds if* $\delta_{\mu,m} \neq 0$ *and* $\mathcal{F}$ *also satisfies Assumption 5.3.2 with* $\Delta_{\xi,m}$ *in* (5.3.6) *from Proposition 5.3.5 satisfying Assumptions 5.2.4-5.2.6.*

Proof. From Proposition 5.3.5, given a certain change size in the raw data, we know the change size in the forecast errors. This, in combination with Theorem 5.2.7, yields the above result. □

Propositions 5.3.5 and 5.3.7 and Corollary 5.3.6 show that provided the mean and variance changes in the potentially complex data structures are large enough, they will result in a mean change in the squared forecast errors that is detectable using our framework.

5.4 Common Forecasting Models

Here we investigate some common forecasting models and show how these can be used within our framework to detect different types of changes. We consider ARMA and ETS forecasting models, showing when they satisfy our forecasting model assumptions and give two small simulation examples illustrating the effectiveness of their forecast errors for detecting changepoints. Note a more detailed simulation study is performed in Section 5.5.

5.4.1 ARMA Models

We denote $\{Y_t : t = 1, 2, \ldots\}$ by the ARMA(p, q) process,

$$\phi(B)(Y_t - \lambda_t) = \theta(B)\epsilon_t, \qquad t = 1, 2, \ldots \,, \tag{5.4.1}$$

where λ_t are the mean parameters, $\phi(B) = 1 - \phi_1 B - \ldots - \phi_p B^p$ and $\theta(B) = 1 + \theta_1 B + \ldots + \theta_q B^q$ are the autoregressive and moving average polynomials respectively and B is the back-shift operator. The error terms, ϵ_t, are assumed to be independent zero-mean random variables with variance, σ_t^2. We assume the ARMA process is causal and invertible meaning,

$$\phi(x) \neq 0 \quad \text{and} \quad \theta(x) \neq 0 \qquad \text{for all } |x| \leq 1 \,.$$

Interest lies in testing whether the time-dependent parameters (specifically λ_t and σ_t) remain constant or change at some time point. A change in λ_t corresponds to a mean change in the data while a change in σ_t corresponds to a variance change. Here

we investigate how an ARMA forecasting model would react to changes in λ_t and σ_t.

First, we present a small simulated example to display the effectiveness of using forecast errors to detect a mean change in an ARMA(1,1) model. For details on how the data was generated and the creation of the forecasting model see Section 5.5. Figure 5.4.1 shows some raw data generated from an ARMA(1,1) model with $\lambda = 0$, $\phi = -0.8$, $\theta = 0.2$ and a change in mean at time point 400. Using an ARMA forecasting model and sequentially producing one-step-ahead forecasts we obtain the forecast errors shown. These forecast errors appear i.i.d and clearly display the changepoint in the raw data. Moreover, the detector (here $D_\mu(m, k)$ with $m = 300$) rises sharply after the changepoint and hits the required threshold (the dashed line) to detect the change just 19 time points after it has occurred. Figure 5.4.1 demonstrates that mean changes in ARMA models will manifest in the forecast errors. Additionally, as the forecast errors are i.i.d this makes detecting the changepoint easier using $D_\mu(m, k)$ than using $D(m, k)$ on the raw data. This is demonstrated further in Section 5.5.

Detecting changes in ARMA model parameters was also considered in Aue et al. (2015). They proposed monitoring ϵ_t and ϵ_t^2 as changes in the ARMA model parameters would manifest in ϵ_t and ϵ_t^2. This approach is similar to ours, as under a perfectly estimated ARMA model the residuals and forecast errors would be identical. Note, the method in Aue et al. (2015) is exclusively for ARMA models where as our framework is more general and can be used with many different forecasting models.

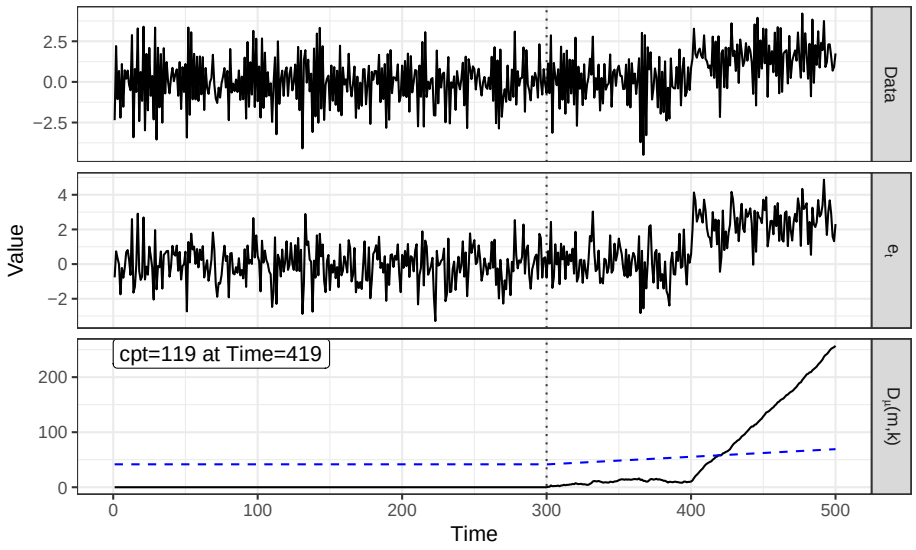


Figure 5.4.1: Data generated from an ARMA(1,1) model with a change in mean at time point 400; the resulting forecasting errors when using an ARMA forecasting model; and the detector, $D_\mu(m, k)$, where the dashed line shows the associated threshold for detecting a change. The vertical dotted line represents the start of the monitoring period.

In the mean change scenario, Aue et al. (2015) showed that a change size of $\delta_{\mu,m}$ in the mean parameter λ would manifest in ϵ_t (and therefore our forecast errors) as

$$\Delta_{\mu,m} = \delta_{\mu,m} \left(1 - \sum_{j=1}^{p} \phi_j\right) \sum_{l=0}^{\infty} \psi_l(\theta) \ ,$$

where

$$\frac{1}{\theta(z)} = \sum_{l=0}^{\infty} \psi_l(\theta) \ .$$

Therefore, Assumptions 5.3.1 and 5.3.2 are satisfied and assuming $\Delta_{\mu,m}$ satisfies Assumptions 5.2.4-5.2.6 then Corollaries 5.3.3 and 5.3.4 hold, giving us the limit distribution of the stopping times under H_0 and H_A.

Similarly, for a change in variance in the data of size $\delta_{\xi,m}$, Aue et al. (2015) showed this would produce a change size in ϵ_t^2 (and therefore our squared forecasting errors) as $\Delta_{\xi,m} = \delta_{\xi,m}$. Note, as expected this matches with Proposition 5.3.5.

Again, if $\Delta_{\xi,m}$ satisfies the Assumptions 5.2.4-5.2.6 and $\Delta_{\mu,m} = 0$ then Corollary 5.3.6 and Proposition 5.3.7 hold.

5.4.2 ETS Models

ETS models cover a wide range of model structures. Here we show that generally ETS forecasting models do not satisfy Assumption 5.3.2 when detecting mean changes, however our methodology will still allow for the detection of variance changes. For demonstration purposes, for the mean change setting, we consider the local level model, ETS(A,N,N), with an additive error structure, no trend and no seasonality. We can produce one-step-ahead forecasts using the ETS(A,N,N) model by the recursive

relation,

$$\hat{y}_t(1) = \hat{y}_{t-1}(1) + \alpha e_{t-1} \,, \tag{5.4.2}$$

where e_{t-1} is the forecast error at time point $t-1$ and $0 < \alpha < 1$ is the smoothing parameter. A larger value of α will cause the model to adapt to the new data more quickly, while a smaller value will place more weight on historical values making the model less sensitive to recent values.

ETS models adapt well to mean changes in raw data and in the majority of scenarios return to being unbiased after a few time steps as shown in the following proposition.

Proposition 5.4.1. *Let* $\{Y_t : t = 1, 2, \ldots\}$ *follow the change in mean model in* (5.3.2) *and* $\mathcal{F}$ *be a forecasting model that follows* (5.4.2). *Assuming* $\mathbb{E}_F[Y_t|\tilde{y}_t] = \lambda$, *the expectation of the forecast errors, after the changepoint, is*

$$\mathbb{E}[e_{k^*+s}] = \begin{cases} \delta_{\mu,m} & \text{for } s = 1 \\ (\lambda + \delta_{\mu,m}) \left(1 - \sum_{i=0}^{s-2} \alpha(1-\alpha)^i\right) - \lambda\left(\alpha(1-\alpha)^{s-1} - (1-\alpha)^s\right) & \text{for } s = 2, 3, \ldots. \end{cases}$$

Proof. This follows from the repeated substitution of the previous forecast errors into (5.4.2) until the pre-change regime is reached. □

Proposition 5.4.1 shows that as s increases the forecast errors will return to zero assuming $\alpha > 0$. Hence, Assumption 5.3.2 is not satisfied unless $\alpha = 0$. Moreover, we can see for larger α this convergence to zero will occur faster.

Thus mean changes in the raw data will only manifest in the forecast errors if α is small. However, our framework can still be used to detect variance changes in the

raw data. These variance changes will affect the prediction interval resulting from the ETS model. Here we show a small example illustrating how a variance change in raw data can make prediction interval poor and how this can be detected using the squared forecast errors from an ETS model.

Figure 5.4.2 shows raw data that exhibits a trend and a variance change at time point 400. We use an ETS(A,A,N) model to forecast this data and the prediction interval (based on ± 2 standard errors) are shown by the dashed lines. Clearly, after the changepoint there are many data points lying outside the prediction interval and hence we need to detect the change in variance so the prediction interval can be re-calibrated. The squared forecast errors show a clear mean change after the variance change in the data has occurred and the detector (here $D_{\xi}(m, k)$ with $m = 300$) reaches the threshold, shown by the dashed line, just 7 time points after the change has occurred.

This mean change in the squared forecast errors is to be expected due to Proposition 5.3.5. Hence, assuming the variance change size in the raw data, $\delta_{\xi,m}$, satisfies Assumptions 5.2.4-5.2.6 then the limit distribution of the stopping time under H_0 and H_A is given in Corollary 5.3.6 and Proposition 5.3.7.

5.5 Simulation

We now examine the performance of our framework in a selection of simulation examples. First, we examine the use of forecast errors when a mean change occurs in

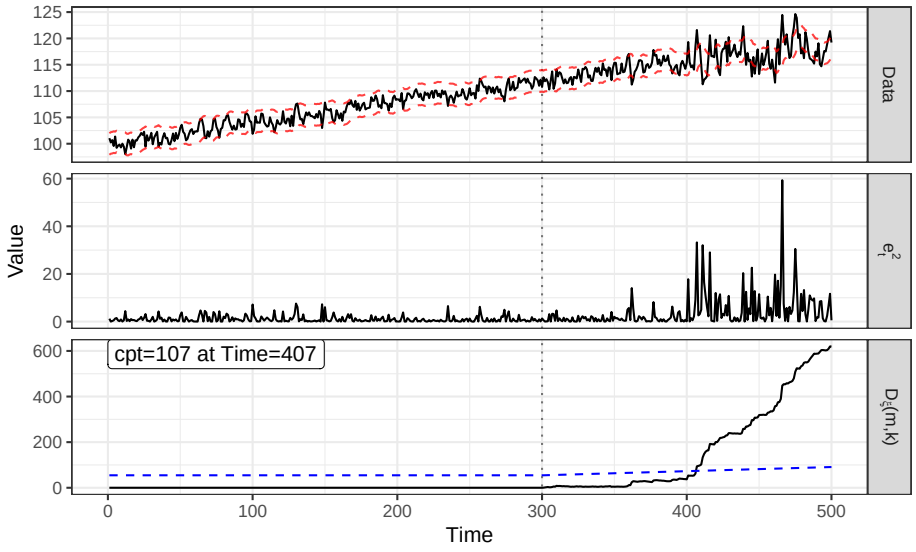


Figure 5.4.2: Data with a variance change at time point 400, with dashed lines showing prediction intervals from an ETS forecasting model. The squared forecast errors are shown along with the Detector with the dashed line being the threshold. The vertical dotted line shows the start of the monitoring period.

the underlying data generating process before exploring the effectiveness of using the squared forecast errors for detecting variance changes in the data generating process. We then examine multiple scenarios where we generate more complex data with seasonality and aim to detect changes in: 1) mean; 2) when a trend develops in the data generating process; and 3) the temporal dependence structure of the data generating process. For the later two scenarios, Assumptions 5.3.1 and 5.3.2 required for the theoretical validity of the stopping rule are not necessarily satisfied, however, we still show satisfactory performance.

For all changes in first-order structure (mean changes and trend change) we compare the detector $D_{\mu}(m, k)$ with forecast errors generated from an appropriate 1) ARMA model against 2) using the detector $D(m, k)$ with the raw data. For all other changes, we compare the detector $D_{\xi}(m, k)$ with the squared forecast errors from an appropriate 1) ARMA and 2) ETS model against 3) the detector $D(m, k)$ with the (centered) squared raw data as in Inclan and Tiao (1994). Note we do not include the ETS model in the first-order changes as we showed in Proposition 5.4.1 that Assumption 5.3.2 is not satisfied meaning we would not expect to see a changepoint in the forecast errors. Throughout we will refer to $D(m, k)$ based upon the raw data as *Raw CUSUM*; $D_{\cdot}(m, k)$ based upon the ARMA forecast errors as *ARMA Forecast Errors*; and $D_{\cdot}(m, k)$ based upon the ETS model forecast errors as *ETS Forecast Errors* with the choice of $D_{\mu}(m, k)$ or $D_{\xi}(m, k)$ dependent on the type of change.

The forecasting models are trained on an initial proportion of the data of length n_{train}. Then the forecasting models sequentially produce one-step-ahead forecasts yielding

the required forecast errors. To generate the ARMA forecasting models we use the *Arima* function within the *forecast* R package (Hyndman et al., 2021) and for the ETS forecasting model we use the *es* function within the *smooth* R package (Svetunkov, 2021). To determine when a change has occurred we use the theoretical thresholds from Theorem 5.2.3 and Corollaries 5.3.3 and 5.3.6 with $\gamma = 0$ and false alarm rate set at 0.05. For the *Raw CUSUM*, to estimate $\hat{\sigma}_m$ within in the threshold we use the Bartlett long run variance estimator (Andrews, 1991) from the *sandwich* R package (Zeileis et al., 2020). For the *ARMA Forecast Errors* and *ETS Forecast Errors*, we assume the forecast errors are i.i.d so the traditional standard deviation estimator is used. Throughout, we repeat each scenario 1000 times and report the average detection delay (ADD) with error bars showing 2 empirical standard errors either side of the mean; detection probability (DP) and false detection probability (FDP). The ADD is the mean of the detection delays given a change was signalled within the length of the data. The DP is the proportion of simulations where a change was detected within the length of the monitoring data. Finally, the FDP is the proportion of simulations where a change was detected before the true change had occurred. Hence, we seek a ADD as close to 0 as possible; a DP close to 1 and a FDP close to 0. Unless otherwise specified, we assume the random errors (innovations) are Normally distributed but note our framework does not require this assumption.

Throughout, we generate data with an ARMA error structure introduced in (5.4.1). In practice, it is rare to know the exact data generating process and this raises questions regarding the misspecification of our forecasting model. Clearly, given the data

generating process is ARMA, the most appropriate forecasting model should be an ARMA model with the same order. In the first two scenarios, we assume the order is known and as such we expect the *ARMA Forecast Errors* to perform best. For the final more complex scenarios, we use the *auto.arima* function from the *forecast* package to pick the order of the ARMA model as well as estimating the model parameters. This means the forecasting model used to generate the *ARMA Forecast Errors* may also be misspecified. We show that despite these misspecifications, a good forecasting model generally still produces better detection results than using the *Raw CUSUM*. Lastly, we note the data generating process used varies across the scenarios to demonstrate our method works well on a wide variety of processes.

5.5.1 Mean Change

First, we explore mean changes in the underlying data generating process. We generate two scenarios where the data is generated from:

1. An AR(2) process with $\phi = (0.5, -0.3)$, $\lambda = 0$ and $\sigma^2 = 1$.

2. An ARMA(1,1) process with $\phi = 0.6$, $\theta = 0.3$, $\lambda = 0$ and $\sigma^2 = 1$.

For both scenarios we generate data of length $n = 1000$, with $n_{\text{train}} = 200$. We set $m = 200$ as the training period for the detectors and once monitoring begins, we introduce a changepoint of size δ_μ to the process at $k^* = 100$. This leaves 500 time points for the methods to identify if a change has occurred. Note as *Raw CUSUM* does not require any training of the model the first $n_{\text{train}} = 200$ time points are not used

by this method. For clarity, to generate the data we used the *arima.sim* function from the *stats* package with an extra 50 time points for burn-in that is removed before analysis begins. For *ARMA Forecast Errors*, we assume the order of the data generating process is known and the parameters are estimated within the training period.

Figure 5.5.1 shows the DP, FDP and ADD of the 2 methods. For both models, we can clearly see the ADD when using the *ARMA Forecast Errors* is much lower than using the *Raw CUSUM* across all change sizes. Additionally, the DP for the *ARMA Forecast Errors* is at least as big as for the *Raw CUSUM* while both have comparable FDP. These metrics show that for the mean changes presented here using *ARMA Forecast Errors* has a distinct advantage over the *Raw CUSUM*.

5.5.2 Variance Change

Now we examine variance changes in the underlying data generating process. Again we generate two scenarios, where the data is generated from

1. An AR(1) process with $\phi = 0.9$, $\lambda = 0$ and $\sigma^2 = 1$.

2. An ARMA(2,1) process with $\phi = (0.5, -0.3)$, $\theta = 0.6$, $\lambda = 0$ and $\sigma^2 = 1$.

Here n, n_{train}, m and k^* are the same as in Section 5.5.1 ($n = 1000$, $n_{\text{train}} = 200$, $m = 200$, $k^* = 100$). Again, for the *ARMA Forecast Errors*, we assume the order of the ARMA process is known but the parameters need to be estimated. For the *ETS Forecast Errors*, we assume an ETS(A,N,N) model and the smoothing parameter is

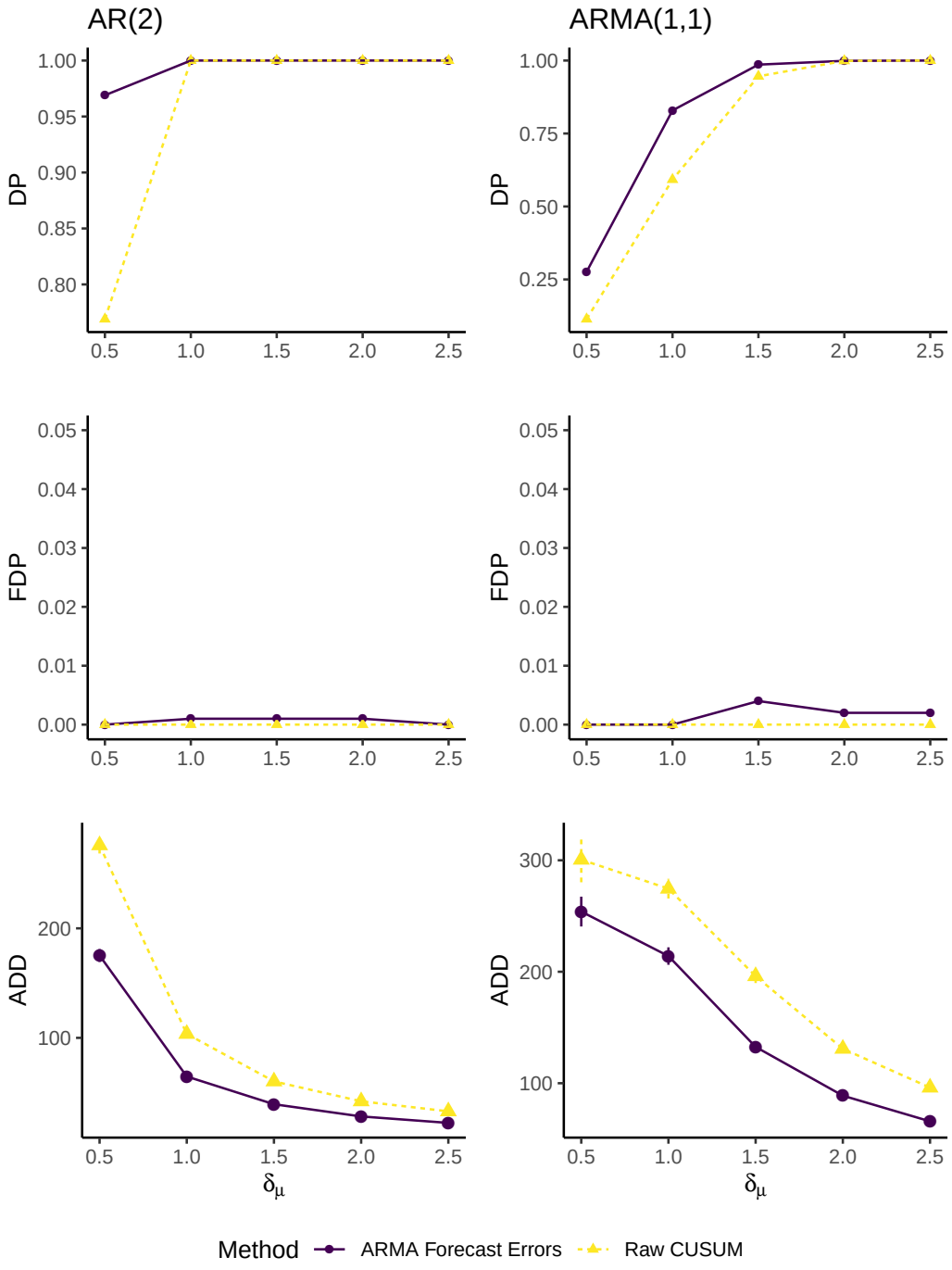


Figure 5.5.1: Detection Probability (DP), False Detection Probability (FDP) and Average Detection Delay (ADD), with error bars showing 2 standard errors either side of the mean, for two scenarios with mean changes of varying sizes.

to be estimated. The change size, δ_ξ, represents the new variance in the underlying data process after the change has occurred.

Figure 5.5.2 shows that in both scenarios the ADD is nearly always smaller when using *ARMA Forecast Errors* or *ETS Forecast Errors* than the *Raw CUSUM*. The reason the ADD for the *Raw CUSUM* is lower in the AR(1) scenario with change size $\delta = 1.5$ is probably due to the small DP. This is also evidenced by the standard error bars overlapping with those from *ARMA Forecast Errors*. The *Raw CUSUM* also has a much higher FDP in the AR(1) scenario while in the ARMA(2,1) scenario *ETS Forecast Errors* has a higher FDP; both are still conservative (< 0.05).

Again these scenarios show the benefit of using forecast errors rather than the raw data to detect changepoints in the underlying data.

5.5.3 Mean and Trend Changes in Seasonal Data

Next, we examine two different scenarios, one where a mean shift occurs and one where a trend emerges in the underlying data that includes a seasonal component, to mimic applications. For the mean change scenario, we generate errors that follow an ARMA(2,1) model with $\phi = (-0.6, 0.3)$, $\theta = -0.3$, $\lambda = 0$ and $\sigma^2 = 1$. We add a changepoint of size δ_μ at time point $k^* = 100$. We add a seasonal component to the data with frequency 12, mimicking monthly data. Within each season the data follows the curve $y = 10\sin(x)$ with $x \in [0, \pi]$ so the mean of each seasonal component is taken from 12 equally spaced points along this curve.

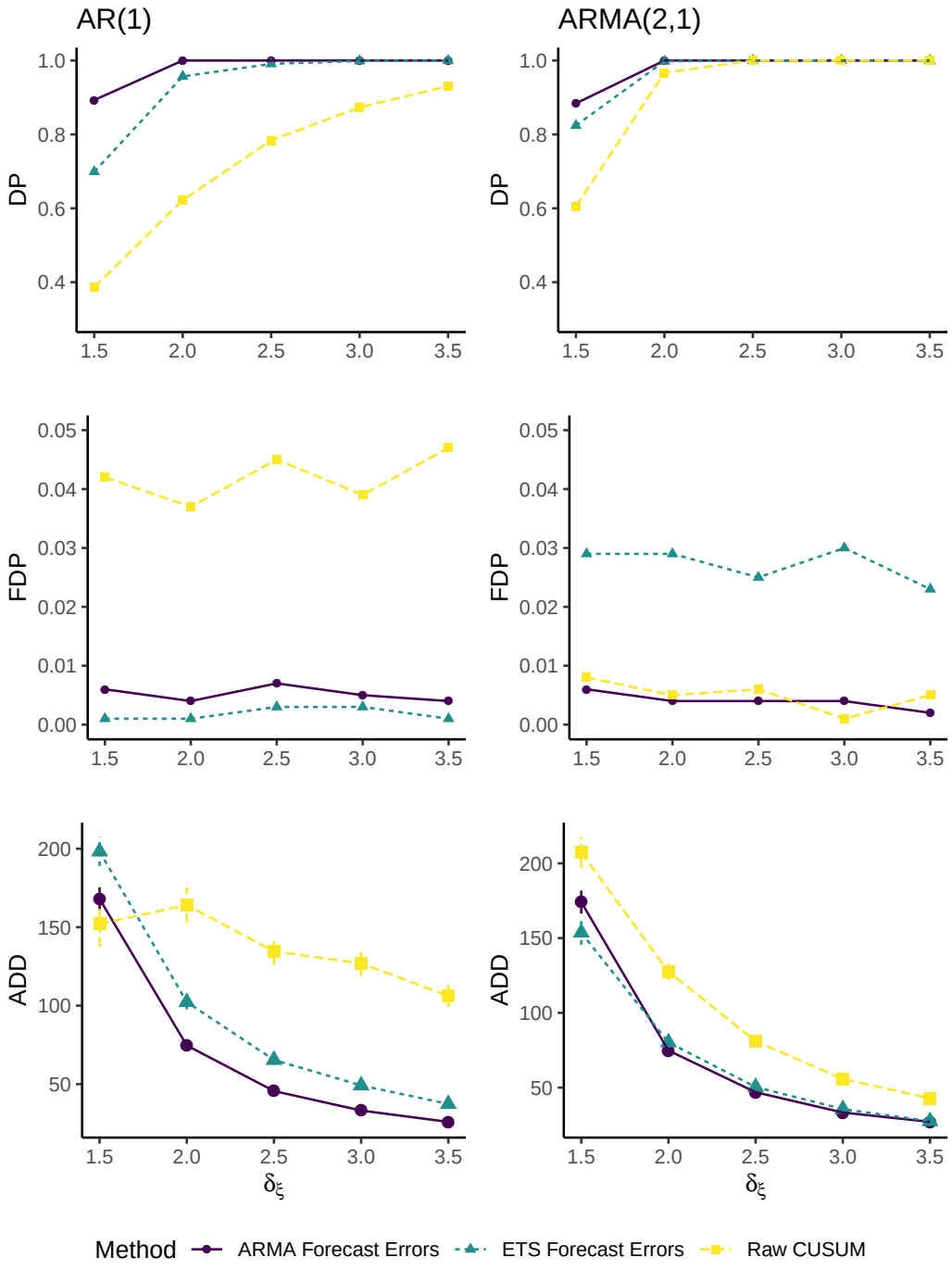


Figure 5.5.2: Detection Probability (DP), False Detection Probability (FDP) and Average Detection Delay (ADD), with error bars showing 2 standard errors either side of the mean, for two scenarios with variance changes of varying sizes.

For the trend change scenario, we simulate errors, ϵ_t, following an ARMA(1,1) model with $\phi = 0.2$, $\theta = 0.2$, $\lambda = 0$ and $\sigma^2 = 1$. We simulate the data with the form $y_t = \mathbb{1}[t > m + k^*]\beta t + \epsilon_t$ for varying gradients β. Hence, the pre-change data has no trend and the post-change data has a trend with specified gradient, β. Moreover, we add a seasonal component to the data with frequency 7, mimicking daily data with day of the week effects. Each 7 time point cycle follows the curve $y = 10\cos(x)$ with $x \in [0, 2\pi]$, so the mean of each seasonal component is taken from 7 equally spaced point along this curve.

Due to the added seasonal component, we can no longer specify the exact models for the *ARMA Forecast Errors*, without using dummy regressors. Hence, we use the *auto.arima* function from the *forecast* package for the *ARMA Forecast Errors* to estimate the seasonal ARMA forecasting model and specify the model should have no differencing. This function estimates the most appropriate seasonal ARMA model using AIC.

Figure 5.5.3 shows some interesting results. For the mean change scenario, *ARMA Forecast Errors* has a high DP and extremely low ADD across all scenarios and outperforms the *Raw CUSUM*. In the change in trend scenario, both *ARMA Forecast Errors* and *Raw CUSUM* have extremely high DP, however, the *ARMA Forecast Errors* has a lower ADD across all change sizes. The FDP is low across all scenarios.

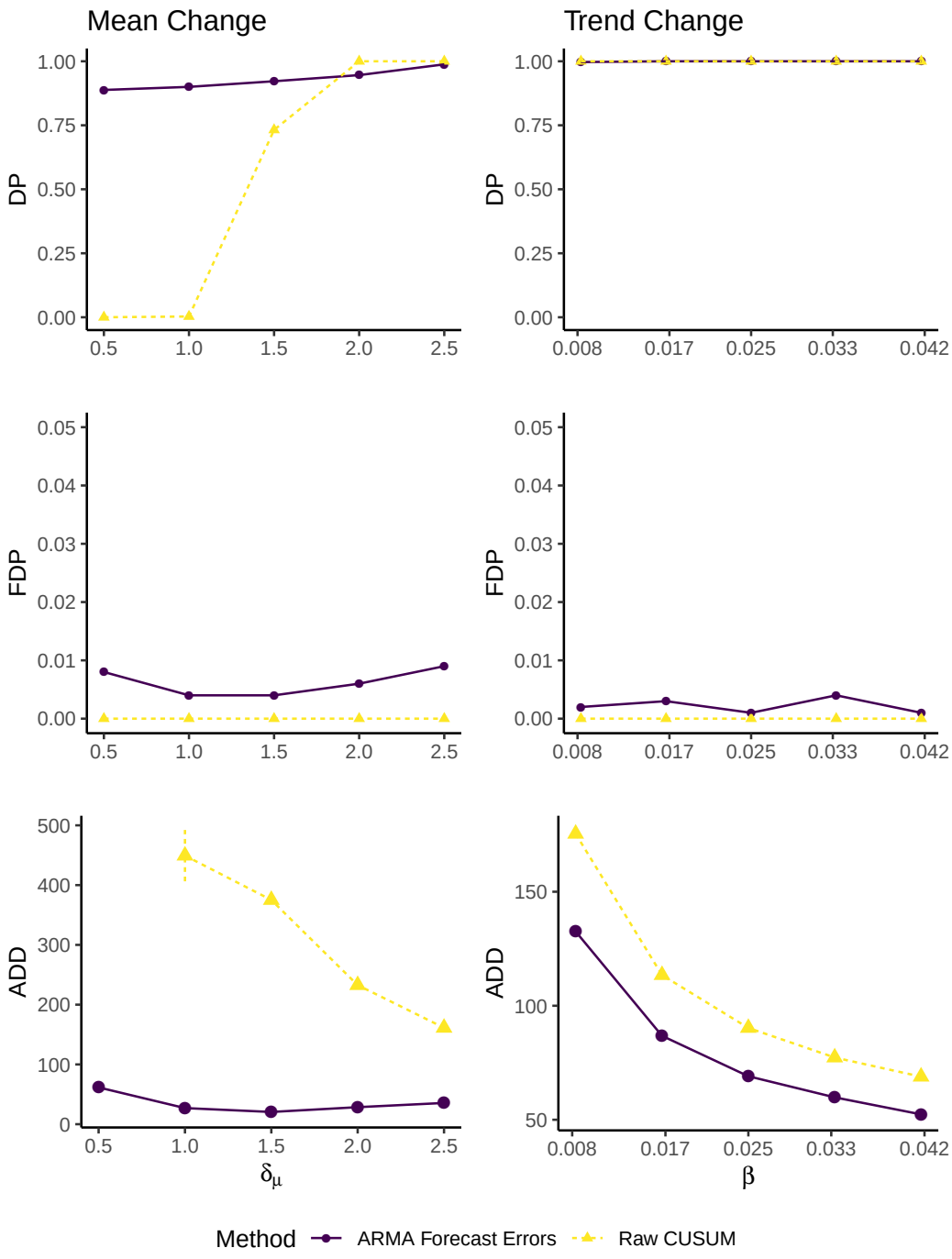


Figure 5.5.3: Detection Probability (DP), False Detection Probability (FDP) and Average Detection Delay (ADD), with error bars showing 2 standard errors either side of the mean, for two scenarios; a change in mean and a change in trend within seasonal data.

5.5.4 Change in Autoregressive Parameter in Seasonal Data

We now examine scenarios where we have data with AR(1) generated errors which undergoes a change in its autoregressive parameter at the specified changepoint. For this section, we include seasonality with a frequency of 4 mimicking quarterly data. The length of the data, training sizes etc remain the same as in the previous scenarios and we test two scenarios:

1. The pre-change data has AR(1) errors with $\phi_{\text{pre}} = 0.2$ and $\lambda = 0$. The post-change errors remain AR(1) but with varying parameters, ϕ_{post}.

2. The pre-change data has AR(1) errors with $\phi_{\text{pre}} = 0.8$ and $\lambda = 0$. The post-change errors remain AR(1) but with varying parameters, ϕ_{post}.

The seasonality is created by adding dummy variables making the mean of each quarter $(-2, 5, 7, -10)$ respectively. We use the *auto.arima* function from the *forecast* package to fit a seasonal ARMA model and we specify the model should have no differencing. For the *ETS Forecast Errors* we use an ETS(A,N,A) model with smoothing parameters to be estimated.

Figure 5.5.4 shows that the DP and ADD is better for all methods when the dependence increases after the change. In all scenarios, using forecast errors gives better results than the *Raw CUSUM* - this is to be expected as the *Raw CUSUM* does not take the seasonality into account and therefore overestimates the variance in the training period making changes much harder to detect. The FDP is conservative for all methods (< 0.075).

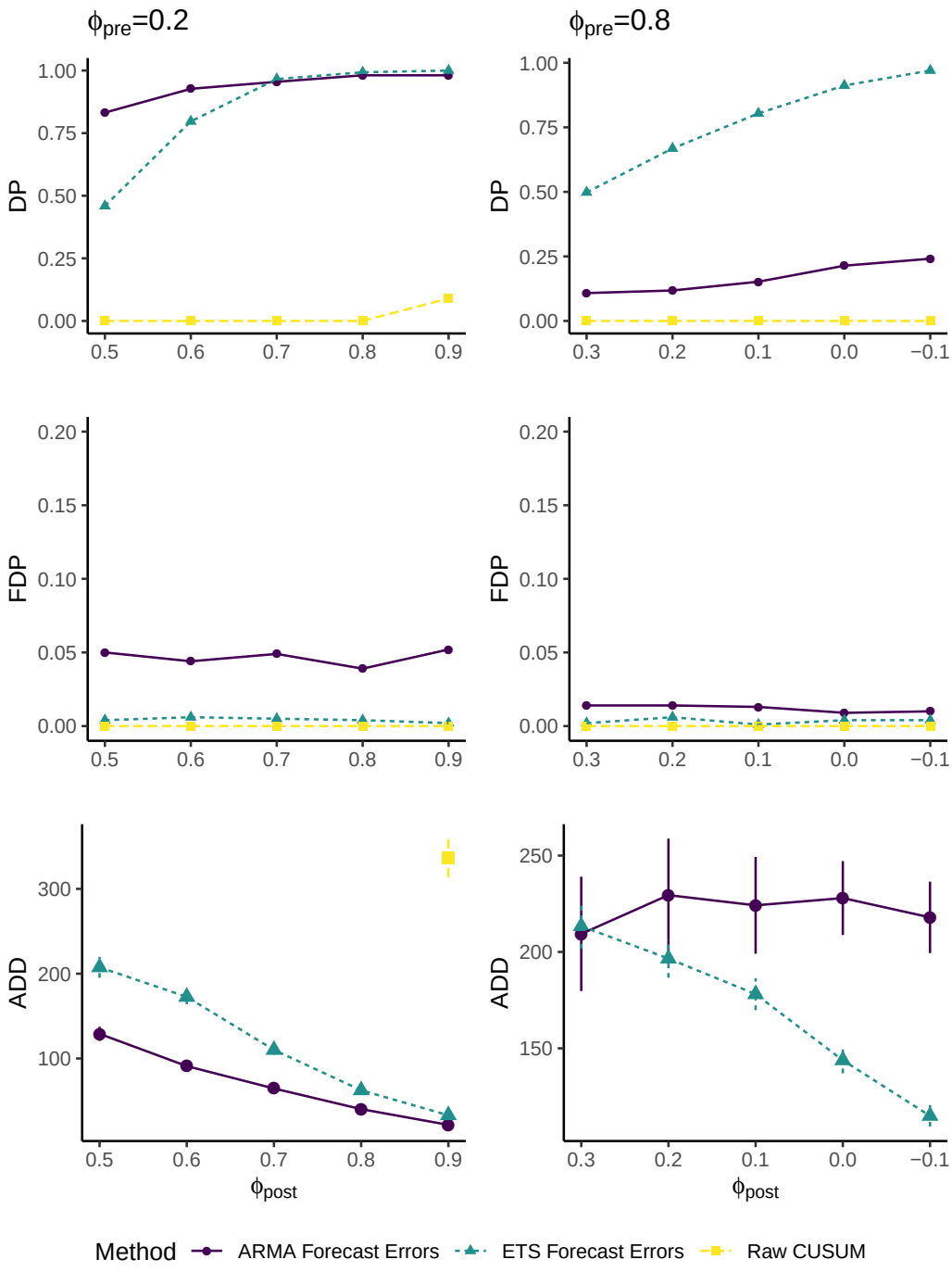


Figure 5.5.4: Detection Probability (DP), False Detection Probability (FDP) and Average Detection Delay (ADD), with error bars showing 2 standard errors either side of the mean, for two scenarios with a change in autoregressive parameter of different sizes

5.6 Application

We now apply this approach to two different applications. Firstly, we seek to determine a change in mail volumes indicating the start of the Christmas peak season within a Royal Mail delivery office by examining the forecast errors of the number of parcels being delivered each day. This is an example where only the forecast errors are known and the underlying data and forecasts are unknown to us. Secondly, we monitor NHS A&E admissions for Gallstone related pathology (GRP) and seek to identify any changepoints by fitting an appropriate ARMA forecasting models.

5.6.1 Parcel Delivery Volumes

Each day Royal Mail forecast the number of parcels that need to be delivered from each delivery office across the UK. With accurate forecasts, the Delivery Office managers can make informed decisions for each day such as ensuring there are enough staff on a shift to meet demand. Each year the number of parcels being delivered increases around the Christmas period, however the exact date can differ from year to year and between different delivery offices. Hence, to maintain accurate staffing levels Royal Mail needs to quickly identify when the Christmas peak season begins.

Here we aim to identify the start of the Christmas peak season in one specific delivery office. We take the forecast errors received from Royal Mail, which date from the 1st July 2020 through to 31st December 2020. We perform our analysis assuming we receive the data in a sequential fashion and aim to identify the changepoint as

soon as possible. The Christmas peak season can start anytime from the start of November, hence we will use the data from July to October as our training sample and begin monitoring the forecast errors from 1st November. We will use both detectors $D_\mu(m,k)$ and $D_\xi(m,k)$ with the theoretical thresholds from Corollaries 5.3.3 and 5.3.6 to determine when a change has occurred.

Figure 5.6.1 shows the forecast errors along with the detectors $D_\mu(m,k)$ and $D_\xi(m,k)$. Both the detectors identify a changepoint on the 2nd of December and looking at the forecast errors this corresponds to a rise in the mean and variance, indicating the start of the Christmas peak season. Upon examination of the raw volumes (not shown due to confidentiality) this identified changepoint corresponds to a jump in the mean and variance of the number of parcels being delivered from this specific delivery office.

Using this information Royal Mail would be able to adjust their forecasting model to account for the increase in parcels being delivered around the Christmas period in the knowledge that the Christmas peak season has begun.

5.6.2 GRP Admissions

We examine the proportion of A&E admissions related to GRP from a number of participating hospitals across England. This dataset is based on NHS Hospital Episode statistics and was initially analyzed in Taib et al. (2021). The data consists of the monthly proportion of GRP A&E admissions and is shown in the top panel of Figure 5.6.2. There is a clear changepoint in the trend that we wish to identify that is marked

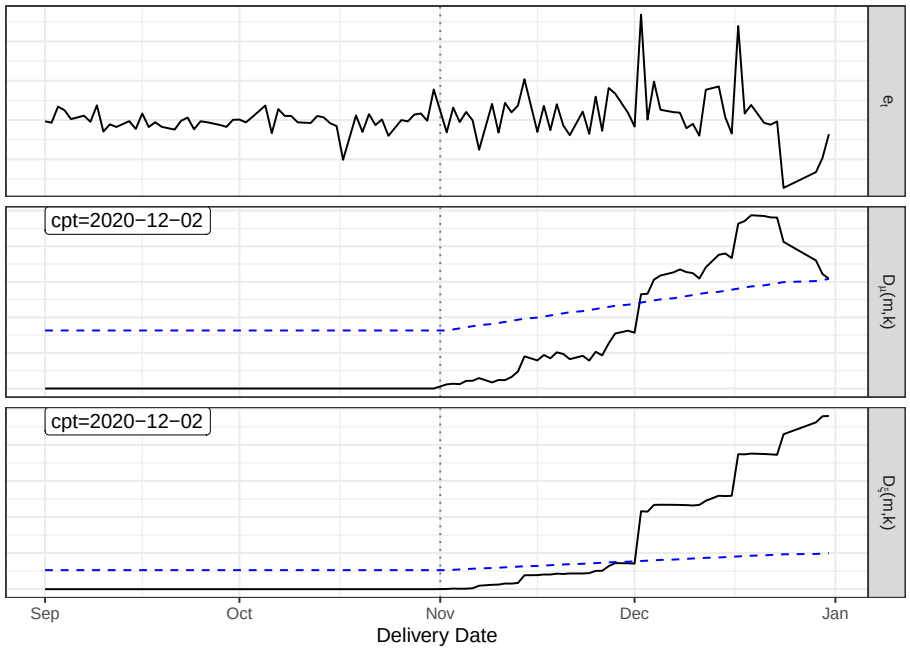


Figure 5.6.1: Forecast errors for the number of parcels to be delivered from a specific Royal Mail delivery office in 2020; the detctors $D_\mu(m,k)$ and $D_\xi(m,k)$ with associated thresholds shown by the dashed lines.

by the vertical dashed line. The analysis of this data set in Taib et al. (2021) was performed retrospectively after all the data from 2010-2020 was available, however to allow for accurate forecasting, using a sequential changepoint method could detect this changepoint in a much shorter time frame.

Taib et al. (2021) found that an SARMA$(1,0,0)(1,0,0)_{12}$ model was appropriate for the data with a time regressor to account for the trend. To identify a change in a sequential manner, using a model-based approach, would require at least two years worth of data to account for the yearly seasonality - this could be far too long and result in poor forecasts of GRP admissions.

We demonstrate the effectiveness of our method by fitting a SARMA$(1,0,0)(1,0,0)_{12}$ on the data up to 2013 and use this as our forecasting model. We used this forecasting model to generate one-step-ahead forecasts for the remaining data up to 2020. We used the data from 2013-2016 as our training data of size $m = 36$ and performed our analysis using both detectors $D_\mu(m,k)$ and $D_\xi(m,k)$ with monitoring beginning in 2016.

Figure 5.6.2 shows the forecast errors and the detectors along with the associated theoretical thresholds from Corollaries 5.3.3 and 5.3.6. We can see that the changepoint is identified in $D_\mu(m,k)$ just four data points after a changepoint identified in Taib et al. (2021) (marked by the dashed line). Additionally, $D_\xi(m,k)$ identified the changepoint just 2 time points after the change occurred. Both of these detectors have a much shorter detection delay than using a model-based sequential changepoint algorithm which has a minimum possible detection delay of 24 months due to the

yearly seasonality in the data.

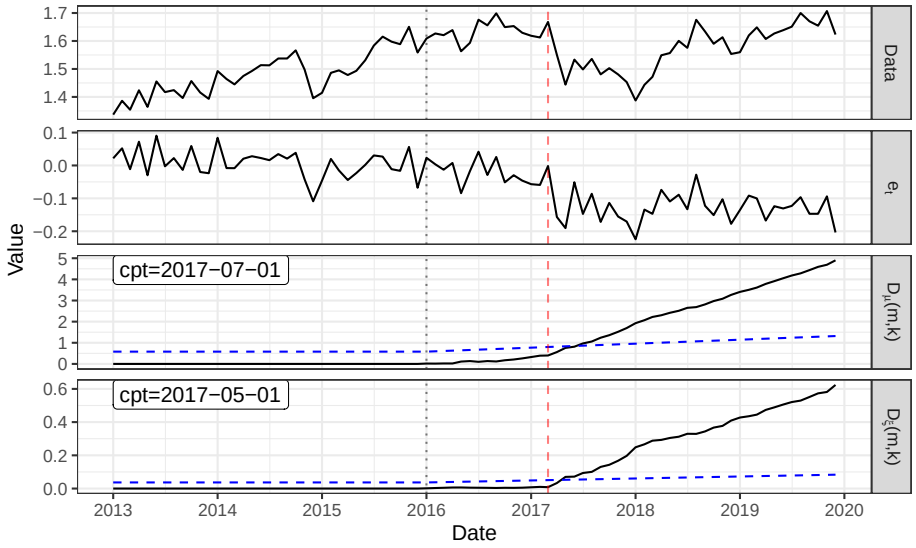


Figure 5.6.2: Data showing the proportion of GRP A&E admissions; the forecast errors from a SARIMA$(1,0,0)(1,0,0)_{12}$ model; and the detectors $D_\mu(m,k)$ and $D_\xi(m,k)$ with associated thresholds shown by the horizontal dashed lines. The dotted line represents the start of the monitoring period and the vertical dashed line indicates a changepoint identified in the retrospective analysis.

5.7 Conclusion

This paper presents a new framework for monitoring forecasting models through the use of sequential changepoint detection. We have shown that common changepoints in underlying, potentially complex, data generating processes manifest in the one-step-ahead forecast errors under certain assumptions and can therefore be detected using sequential changepoint analysis. Moreover, we have shown by monitoring the forecast errors rather than the raw data for changepoints, we can greatly reduce the

detection delay of changepoints while maintaining a low false positive rate. When using ARMA forecasting models, we have shown that changes in mean and variance in the underlying data can be detected. ETS models often adapt to mean changes, but can still be used to detect variance changes, which can cause prediction intervals to become miscalibrated. Applying our framework to delivery volume forecast errors from a Royal Mail delivery office allowed for the rapid detection of the Christmas peak season thus allowing future forecasts and staffing to be appropriately adjusted. Furthermore, we showed using a forecasting model on NHS A&E admission data allowed us to detect a changepoint in the proportion of GRP admissions related in a shorter time frame than would be possible using retrospective analysis or a model-based sequential changepoint approach.

There are a number of extensions that could be made to this framework. Firstly, it is becoming increasingly common for multivariate forecasts to be made where many variables/products are forecasted at once. Incorporating a multivariate sequential changepoint algorithm on the resulting vector of forecast errors would be an interesting extension but raises additional questions such as the sparsity of the change and is therefore left as future work. Additionally, here we focused on one-step-ahead forecast errors but the framework is general, for h-steps ahead, and combining information from multiple forecast horizons would be an interesting extension.

Chapter 6

Conclusion

In this thesis, we have presented novel methods and frameworks for identifying changepoints in time series. Here we summarize the key findings within this thesis, as well as highlighting some of the limitations, open problems and future work associated with the proposed methodologies.

6.1 Key Findings

In Chapter 3, we presented a novel, computationally efficient, method for detecting mean and variance changes in high-dimensional time series, GeomCP. There are a limited number of methods for detecting different types of changes simultaneously in the fully multivariate changepoint setting. We have shown empirically that GeomCP provides a viable solution in this setting and furthermore, we have shown the consistency of GeomCP for detecting changes in mean using our distance projection. In

the fully multivariate changepoint setting our methodology outperforms the current state-of-the-art methods and remains competitive in the sparse changepoint setting. In two industry examples, we showed that GeomCP detected multiple changepoints from S&P500 stock data that can be incorporated in future modelling. Secondly, GeomCP identified numerous changepoints in genomic data that highlights potential mutations along the genome that could be related to bladder tumours.

Chapter 4 proposed a new cost function and a changepoint framework for detecting changes in subspace. We demonstrated how high-dimensional data can often lie in a low-dimensional linear subspace and by exploiting this structure we can provide more accurate changepoint estimates. We highlighted that subspace changepoints can be seen as covariance changes with conditions on the eigenvalues of the covariance matrix. By comparing our methodology to generic covariance methods we showed, when the condition on the eigenvalues is satisfied, using our method gives an improved performance at detecting changes. One industry example where data lie in a low-dimensional subspace is based on Motion Capture data. We showed how our methodology can be used to identifying changes in human activity based on the movement of different parts of their body. This analysis can be used to automate the process of someone manually watching multiple videos and identifying the changepoints.

Finally, Chapter 5 introduced a novel framework for monitoring forecast models. We showed, when using an appropriate forecast model, changes in potentially complex data structures manifest in the forecast errors and can be detected using sequential changepoint techniques. This framework provides a way to automatically monitor the

accuracy of a forecasting model and identify if it starts performing poorly. Moreover, from a sequential changepoint detection viewpoint, if an appropriate forecasting model can be created for the data, this can be used to detect changes in potentially complex data structures in a more timely manner than current methods.

6.2 Open Problems and Future Work

Despite the effectiveness of the GeomCP method presented in Chapter 3 for detecting fully multivariate changepoints, there are a few open questions that remain. Firstly, the theoretical guarantees for the distance projection were given in the Gaussian setting while assuming temporal independence and a diagonal covariance matrix. This is a limited setting and these assumptions are often violated in industrial applications. Therefore, extending the theoretical results to other distributions and incorporating temporal dependence would be desirable. Moreover, a thorough theoretical understanding of the angle projection and how it preserves changes in covariance would give practitioners added confidence for using GeomCP in a wider range of data scenarios. Aside from theoretical extensions, exploring the geometry of other types of distributional changepoints could lead to other geometrically inspired mappings. We have shown a first step in utilizing geometry in a changepoint setting and believe this could be extended further into a more general framework.

During the peer-review process of the change in subspace methodology presented in Chapter 4, the connection of our work to factor models was highlighted. Factor mod-

els arose from the economics literature and recently interest has grown in detecting changes in the factor loadings which correspond to our subspace changepoints. The cost function we proposed was independently developed Chen (2015) for the localization of a single change in factor loadings. Recently, there have been other papers examining changes in factor loadings (Barigozzi et al., 2018; Duan et al., 2021; Jiao et al., 2021), however, there are some key differences between this and our work. Firstly, many of these methods assume a single changepoint is present in the data and aim to identify its location. This is subtly different from our framework where we first identify if a change is present, and if so, then we locate the change. Moreover, we provide an extension to multiple changepoints and provide an alternative formulation of the cost function, which in practice leads to significant computational savings. Theoretical guarantees are missing from our proposed method, however, Bai et al. (2020) details key theoretical results regarding the cost function we use in the factor model setting. Extending these results to the multiple changepoint setting could provide theoretical results for our method and back up the strong empirical results shown in Chapter 4.

Finally, our framework for monitoring forecasting models only considered one-step-ahead forecasts, however, the framework is general for h-steps-ahead. Despite this, combining information from multiple forecast horizons could lead to an extension of our work which may produce faster detection rates for identifying changepoints. Moreover, the framework is limited to univariate forecasts but the extension to multivariate forecasts would be feasible with additional considerations. Using a multivariate se-

quential changepoint detection technique on the multivariate forecasts would be a sensible starting point. However, this raises many of the issues regarding multivariate changepoints discussed in Chapter 2.2.

Multivariate changepoint analysis and sequential changepoint analysis are ever-growing research areas with a vast amount of open problems. The methodologies and frameworks presented here aim to provide some initial solutions to a few of these problems, yet further development is needed to truly solve the problems at hand. The open problems and suggest extensions to our methods will help further in our understanding of multivariate and sequential changepoint problems along with further investigation of open problems not tackled in this thesis.

Appendix A

High-Dimensional Changepoint Detection via a Geometrically Inspired Mapping

A.1 Preliminary Lemmas

Lemma A.1.1. *Suppose we have independent random variables,* $Y_i \sim \Gamma(\alpha, \beta_i)$*, with a common shape parameter,* $\alpha > 0 \ \forall i$*, and varying scale parameters,* $\beta_i > 0 \ \forall i$*. Let* $\mathbb{E}(Y_i) = \mu_i$*,* $Var(Y_i) = \sigma_i^2$ *and* $s_n^2 = \sum_{i=1}^{n} \sigma_i^2$*. Then,* $\exists \delta > 0$ *such that,*

$$\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{n} \mathbb{E}\left[|Y_i - \mu_i|^{2+\delta}\right] = 0 \, .$$

Proof. Consider the moment generating function of $Y_i - \mu_i$,

$$\begin{aligned} M_{Y_i-\mu_i}(t) &= \mathbb{E}\left[e^{t(Y_i-\mu_i)}\right] \\ &= e^{-t\mu_i}\mathbb{E}\left[e^{tY_i}\right] \\ &= \frac{e^{-t\mu_i}}{(1-t\beta_i)^{\alpha}} \, . \end{aligned}$$

By considering the 4th derivative of $M_{Y_i-\mu_i}(t)$ we gain,

$$\begin{aligned} \sum_{i=1}^{n}\mathbb{E}\left[|Y_i-\mu_i|^4\right] =& \sum_{i=1}^{n} M^{(\mathrm{IV})}_{Y_i-\mu_i}(0) \\ =& \sum_{i=1}^{n}(\alpha+3)(\alpha+2)(\alpha+1)\alpha\beta_i^4 - 4(\alpha+2)(\alpha+1)\alpha\beta_i^3\mu_i \\ &+ 6(\alpha+1)\alpha\beta_i^2\mu_i^2 - 4\alpha\beta_i\mu_i^3 + \mu_i^4 \\ =& c_1\sum_{i=1}^{n}(\beta_i^4) - c_2\sum_{i=1}^{n}(\beta_i^3\mu_i) + c_3\sum_{i=1}^{n}(\beta_i^2\mu_i^2) - c_4\sum_{i=1}^{n}(\mu_i^3\beta_i) + \sum_{i=1}^{n}\mu_i^4 \\ \leq& c_1 n\beta_{\max}^4 - c_2 n\beta_{\min}^3\mu_{\min} + c_3 n\beta_{\max}^2\mu_{\max}^2 - c_4 n\beta_{\min}\mu_{\min}^3 + n\mu_{\max}^4 \\ =& nc_5 \, , \end{aligned}$$

where $c_1, c_2, c_3, c_4 \in \mathbb{R}^+$ and $c_5 \in \mathbb{R}$.

Now,

$$\begin{aligned} s_n^4 &= \left(\sum_{i=1}^{n}\sigma_i^2\right)^2 \\ &\geq n^2\sigma_{\min}^4 \, . \end{aligned}$$

Hence, for $\delta = 2$,

$$0 \leq \lim_{n\to\infty}\frac{1}{s_n^{2+\delta}}\sum_{i=1}^{n}\mathbb{E}\left[|Y_i-\mu_i|^{2+\delta}\right] \leq \lim_{n\to\infty}\frac{nc_5}{n^2\alpha^2\beta_{\min}^4} = 0 \, ,$$

as required. □

Lemma A.1.2. *Suppose we have independent random variables,* $Y_i \sim N(0, \sigma_i^2)$*, and* $X = \sum_{i=1}^{n} Y_i^2$*. Then, as* $n \to \infty$*,*

$$\frac{X - \sum_{i=1}^{n} \sigma_i^2}{\sqrt{2 \sum_{i=1}^{n} \sigma_i^4}} \xrightarrow{\mathcal{D}} N(0,1) \; .$$

Proof. Let $Y_i = \sigma_i Z_i$, where Z_i are independent standard normal random variables. Then,

$$\left(\frac{Y_i}{\sigma_i}\right)^2 = Z_i^2 \; ,$$

and $Z_i^2 \sim \chi^2(1) \sim \Gamma(\frac{1}{2}, 2)$. Hence,

$$Y_i^2 \sim \Gamma(\frac{1}{2}, 2\sigma_i^2) \; .$$

Using Lemma A.1.1 with $\alpha = \frac{1}{2}$, $\beta_i = 2\sigma_i^2$ and $s_n^2 = 2 \sum_{i=1}^{n} \sigma_i^4$, we have,

$$\lim_{n \to \infty} \frac{1}{s_n^4} \sum_{i=1}^{n} \mathbb{E}\left[|Y_i^2 - \sigma_i^2|^4\right] = 0 \; ,$$

satisfying Lyapunov's condition for the Lyapunov Central Limit Theorem (Lyapunov, 1901) for $\delta = 2$. Applying Lyapunov's Central Limit Theorem gains the required result. □

A.2 Proof of Theorem 1

Proof. We can re-write X as,

$$X = \sqrt{\sum_{i=1}^{p} \mu_i^2 + 2 \sum_{i=1}^{p} \mu_i \sigma_i Z_i + \sum_{i=1}^{p} \sigma_i^2 Z_i^2} \; ,$$

where Z_i are independent standard normal random variables. Let Z^* and Z^{**} be dependent standard normal random variables with unknown correlation ρ. Using Lemma A.1.2 and the binomial expansion, as $p \to \infty$,

$$
\begin{aligned}
X \xrightarrow{\mathcal{D}} & \left(\sum_{i=1}^{p} \left(\mu_i^2 + \sigma_i^2 \right) + 2 \sqrt{\sum_{i=1}^{p} (\mu_i \sigma_i)^2} Z^* + \sqrt{2 \sum_{i=1}^{p} \sigma_i^4} Z^{**} \right)^{\frac{1}{2}} \\
&= \left(\sum_{i=1}^{p} \left(\mu_i^2 + \sigma_i^2 \right) \right)^{\frac{1}{2}} \left(1 + \frac{2 \sqrt{\sum_{i=1}^{p} (\mu_i \sigma_i)^2} Z^* + \sqrt{2 \sum_{i=1}^{p} \sigma_i^4} Z^{**}}{\sum_{i=1}^{p} \left(\mu_i^2 + \sigma_i^2 \right)} \right)^{\frac{1}{2}} \\
&= \left(\sum_{i=1}^{p} \left(\mu_i^2 + \sigma_i^2 \right) \right)^{\frac{1}{2}} \left(1 + \frac{2 \sqrt{\sum_{i=1}^{p} (\mu_i \sigma_i)^2} Z^* + \sqrt{2 \sum_{i=1}^{p} \sigma_i^4} Z^{**}}{2 \sum_{i=1}^{p} \left(\mu_i^2 + \sigma_i^2 \right)} + \mathcal{O}\left(\frac{1}{p} \right) \right) \\
&= \left(\sum_{i=1}^{p} \left(\mu_i^2 + \sigma_i^2 \right) \right)^{\frac{1}{2}} \\
&\quad + \sqrt{\frac{2 \sum_{i=1}^{p} (\mu_i \sigma_i)^2 + \sum_{i=1}^{p} \sigma_i^4 + 2\rho \left(2 \sum_{i=1}^{p} \sum_{j=1}^{p} \mu_i^2 \sigma_i^2 \sigma_j^4 \right)^{\frac{1}{2}}}{2 \sum_{i=1}^{p} \left(\mu_i^2 + \sigma_i^2 \right)}} Z^{***} + \mathcal{O}\left(\frac{1}{\sqrt{p}} \right) ,
\end{aligned}
$$

where Z^{***} is a standard normal random variable, thus giving the required result. □

A.3 Dense Mean Changepoints

We simulate data with changes in mean that occur in all series for a wide range of n and p and show a subset of the results. We keep the variance vector constant and the total mean change size is,

$$
\sum_{j=1}^{p} \mu_{j,\text{post}} - \mu_{j,\text{pre}} = \sqrt{p} \Theta \, . \tag{A.3.1}
$$

We split the total change size evenly across all series and display results with $\Theta = 1.2$. Similar findings occur with varying values of Θ. As we assume changepoints are dense, the change size in each series violates the minimum change size assumption in Inspect; we include the results for interest. We apply the GeomCP, Inspect and E-Divisive methods to these scenarios and the true detection rate (TDR) and false detection rate (FDR) are shown in Table A.3.1. For $p \geq 1000$, we are unable to calculate an appropriate stopping threshold for Inspect and, therefore, exclude it from such scenarios.

Table A.3.1 clearly shows GeomCP has a greater TDR than both Inspect and E-Divisive in all scenarios. As p increases, the performance of Inspect and E-Divisive drastically decreases, whereas GeomCP maintains a similar TDR. The FDRs shown in Table A.3.1 indicate GeomCP is either better or competitive with Inspect and E-Divisive. Taking both TDR and FDR into account it is clear that GeomCP outperforms both Inspect and E-Divisive in detecting mean changes that occur in all series.

A.4 Dense Mean and Variance Changepoints

We simulate data with changes in mean and variance that occur in all series for multiple n and p and show a subset of the results. We set the total mean change size as in (A.3.1) and the total variance change size to be,

$$\prod_{j=1}^{p} \frac{\sigma_{j,\text{post}}}{\sigma_{j,\text{pre}}} = \Phi^{\sqrt{p}}\,, \tag{A.4.1}$$

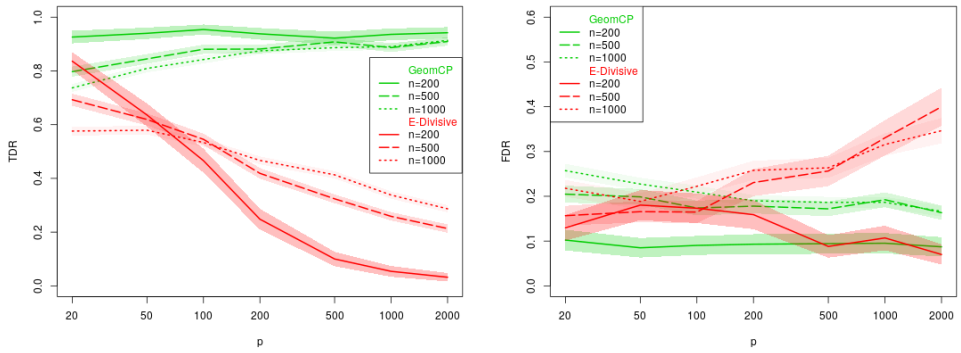
Table A.3.1: TDR and FDR for GeomCP, Inspect and E-Divisive for simulated data sets containing mean changes that occur in all series

		GeomCP		Inspect		E-Divisive	
n	p	TDR	FDR	TDR	FDR	TDR	FDR
200	50	**0.956**	**0.086**	0.484	0.163	0.812	0.095
	100	**0.928**	**0.101**	0.304	0.150	0.614	0.159
	500	**0.952**	0.080	0.018	**0.026**	0.160	0.118
	1000	**0.922**	**0.088**	-	-	0.076	0.094
	2000	**0.952**	0.086	-	-	0.046	**0.065**
500	50	**0.877**	0.189	0.613	0.191	0.790	**0.119**
	100	**0.913**	0.168	0.470	0.247	0.689	**0.138**
	500	**0.891**	**0.189**	0.203	0.462	0.389	0.211
	1000	**0.891**	**0.172**	-	-	0.304	0.285
	2000	**0.896**	**0.172**	-	-	0.230	0.380
1000	50	**0.904**	0.188	0.637	0.203	0.807	**0.125**
	100	**0.900**	0.184	0.419	0.280	0.713	**0.154**
	500	**0.900**	**0.188**	0.318	0.446	0.507	0.200
	1000	**0.906**	**0.184**	-	-	0.367	0.311
	2000	**0.908**	**0.175**	-	-	0.333	0.292
2000	50	**0.880**	0.200	0.634	0.213	0.747	**0.131**
	100	**0.894**	0.187	0.561	0.262	0.730	**0.127**
	500	**0.892**	**0.190**	0.343	0.392	0.463	0.284
	1000	**0.887**	**0.200**	-	-	0.447	0.253
	2000	**0.889**	**0.197**	-	-	0.360	0.281

and split the total change sizes evenly across the series. As we have a change in both mean and variance, we would expect a smaller change in both to still be detectable. Here we set $\Theta = 1$ and $\Phi = 2$ and note similar finding occurred for different Θ and Φ. We apply the GeomCP and E-Divisive to these data sets and the TDR and FDR are shown in Figure A.4.1.

Figure A.1(a) shows that for all scenarios GeomCP has a better TDR than E-Divisive and the gap between the methods widens as p increases. Additionally, the FDR, shown in Figure A.1(b), shows that GeomCP has an improved or competitive FDR with E-Divisive across all scenarios. Considering this, it is clear that for detecting

changes that occur in mean and variance simultaneously, GeomCP is to be preferred over E-Divisive.



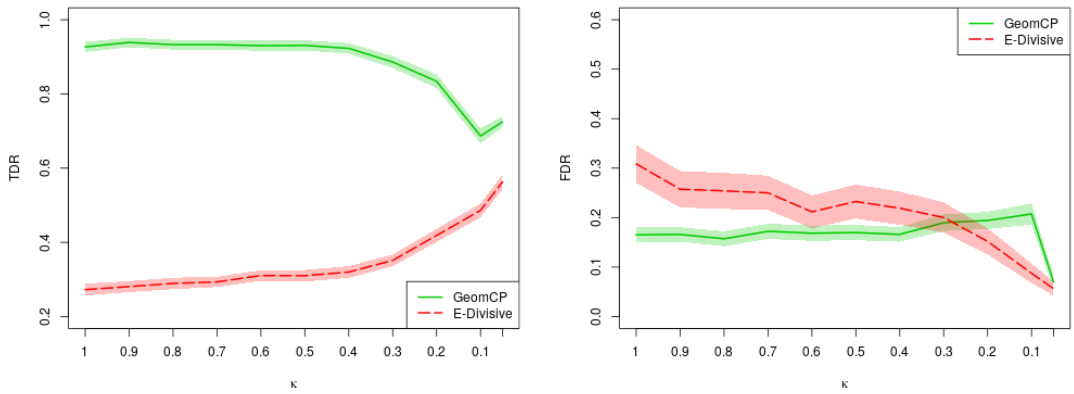
(a) TDR (b) FDR

Figure A.4.1: (a) TDR and (b) FDR for GeomCP and E-Divisive for simulated data sets containing mean and variance changes that occur in all series for multiple n and p

A.5 Sparse Variance Changepoints

To investigate sparse variance changes we set $n = 500$, $p = 200$ and varied κ. We keep the mean vector constant and the variance change in each series that undergoes a change, is the total variance change size defined in (A.4.1), split between the expected number number of series to undergo a change. We display results with $\Phi = 3$ and note similar findings occur with varying values of Φ. We apply the GeomCP and E-Divisive methods to these scenarios and the TDR and FDR are shown in Figure A.5.1.

Figure A.1(a) shows that for all levels of sparsity, GeomCP has a far greater TDR than E-Divisive. Figure A.1(b) shows that GeomCP has a competitive, if not lower, FDR than E-Divisive across all sparsity levels. This shows that for sparse variance changes GeomCP has an improved performance over E-Divisive.



(a) TDR (b) FDR

Figure A.5.1: (a) TDR and (b) FDR for GeomCP and E-Divisive for simulated data sets with sparse variance changes with $n = 500$ and $p = 200$

A.6 Between-series Dependence: Mean Change

We now investigate the performance of GeomCP in scenarios where there is underlying covariance structure and a mean change occurs. For these scenarios, we set $n = 200$, $p = 100$ and have one changepoint at $\tau = 100$. The pre-changepoint data will be distributed from a $N(\mathbf{0}, \Sigma)$ while the post-changepoint data distributed from a $N(\boldsymbol{\mu}, \Sigma)$. We will vary the change size, $\boldsymbol{\mu}$, while the entries of $\boldsymbol{\mu}$ will be identical for each change size. We will compare three structures for Σ:

1. Independent case: $\Sigma = I$.

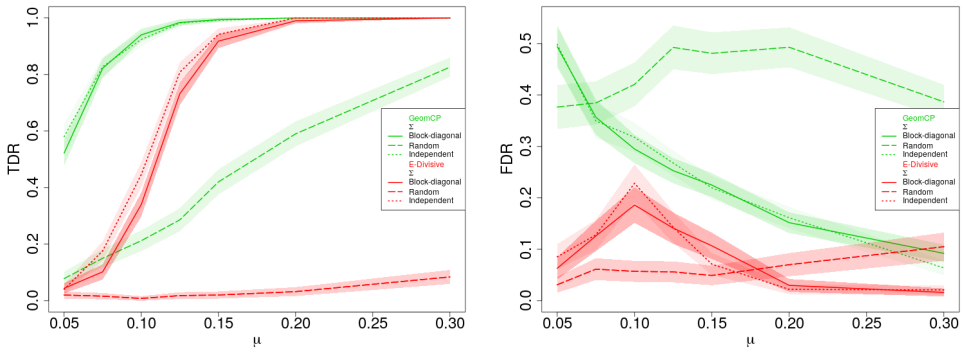
2. Block-diagonal case: Here Σ will be a block-diagonal matrix with block size of 2. The off-diagonal entries will be randomly sampled from a $U(-0.6, -0.3) \cup U(0.3, 0.6)$ distribution with the diagonal entries equal to 1.

3. Random case: Here we let $\Sigma = PDP'$ where P is an orthogonalized matrix of standard Normal random variables and D is a diagonal matrix with entries decreasing from 30 to 1.

As we no longer have independence between series we cannot assume Normality of the distance and angle measures within GeomCP. Hence, we use the empirical cost function (Haynes et al., 2017b) within PELT to detect changes in the distance and angles measures. We similarly use the empirical cost function in the independent case for comparability.

Figure A.1(a) shows GeomCP has a superior TDR over E-divisive for smaller change sizes μ. Interestingly, the TDR for the random covariance structure is poor for both methods. Similarly, to the case of a variance change, Figure A.1(b) shows by using the empirical cost function within PELT we get a worse FDR for smaller change sizes. However, this trade-off between TDR and FDR could be improved by tuning the penalty used within PELT.

A.7 Performance under the Null



(a) TDR (b) FDR

Figure A.6.1: TDR and FDR for GeomCP (using the empirical cost function) and E-Divisive for simulated data with an underlying covariance structure and a change in mean for $n = 200$ and $p = 100$

Here we investigate the performance of GeomCP on null data. The aim is to keep the number of false positive as low as possible. We simulated Normal data with no changepoints for varying n and p and ran the GeomCP method using the Normal cost function within PELT. We calculated the false positive rate (FPR) by taking the total number of detected changepoints across all replications and dividing by the total number of replications.

Figure A.7.1 shows that for the majority of scenarios the FPR stays below 0.05 indicating a conservative performance in terms of the type 1 error.

A.8 Dense Change Size Investigation

Now we investigate how the performance of GeomCP, and the competing methods, vary as we alter Θ and Φ for the dense change in mean and change in variance

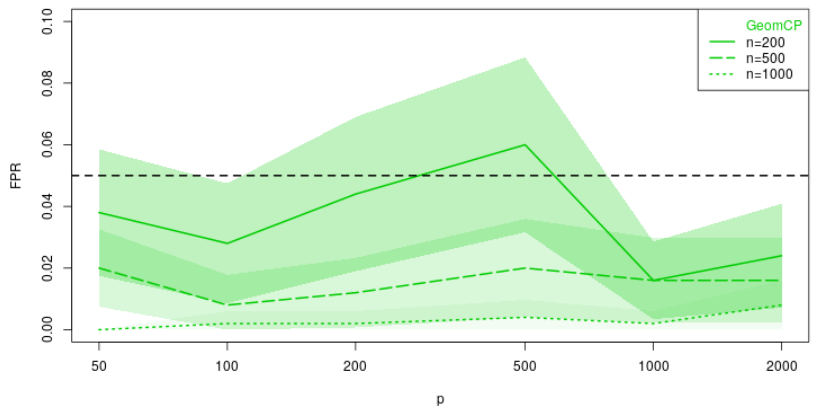


Figure A.7.1: FPR for GeomCP (using the Normal cost function) for simulated data sets with no changepoints for varying n and p

scenarios. Table A.8.1 shows the TDR and FDR of GeomCP, Inspect and E-Divisive for multiple n and p for different change size values. GeomCP's performance remains an improvement upon the competing methods for all change sizes. Table A.8.2 shows similar finding for the change in variance scenarios.

A.9 CROPS diagnostics plots

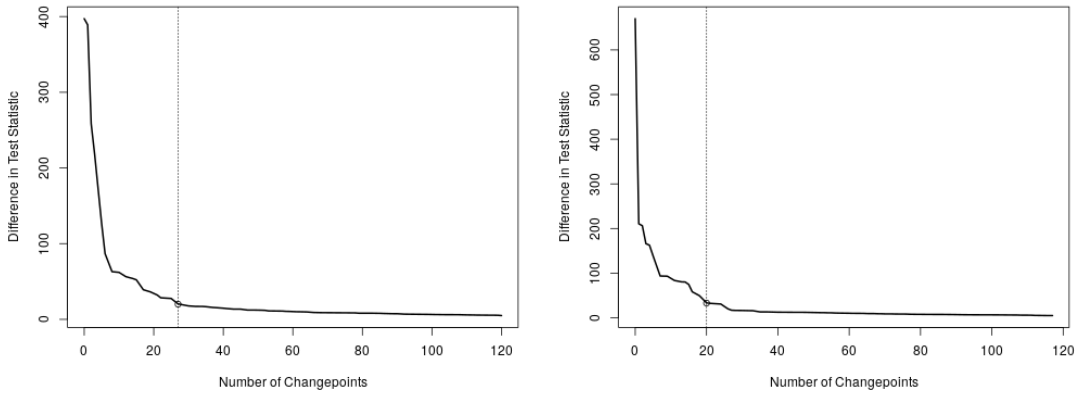
Here we show the CROPS diagnostic plots used to find the optimal number of changepoints in both our applications, the CGH data set and the S&P500 data set. The circled points indicates the elbow of the plot and the number of changepoints we used.

Table A.8.1: TDR and FDR for GeomCP, Inspect and E-Divisive for simulated data sets containing mean changes that occur in all series. Θ relates to the size of the change.

			GeomCP		Inspect		E-Divisive	
n	p	Θ	TDR	FDR	TDR	FDR	TDR	FDR
200	100	1	**0.794**	**0.092**	0.130	**0.092**	0.384	0.181
		1.2	**0.928**	**0.101**	0.304	0.150	0.614	0.159
		1.5	**0.988**	**0.073**	0.610	0.180	0.908	0.076
	500	1	**0.758**	0.144	0.010	**0.014**	0.078	0.111
		1.2	**0.952**	0.080	0.018	**0.026**	0.160	0.118
		1.5	**0.986**	0.072	0.088	**0.049**	0.372	0.153
1000	100	1	**0.764**	0.252	0.329	0.362	0.567	**0.184**
		1.2	**0.900**	0.184	0.419	0.280	0.713	**0.154**
		1.5	**0.983**	0.146	0.581	0.167	0.820	**0.092**
	500	1	**0.751**	**0.263**	0.228	0.538	0.327	0.369
		1.2	**0.900**	**0.188**	0.318	0.446	0.507	0.200
		1.5	**0.976**	0.144	0.393	0.412	0.667	**0.136**

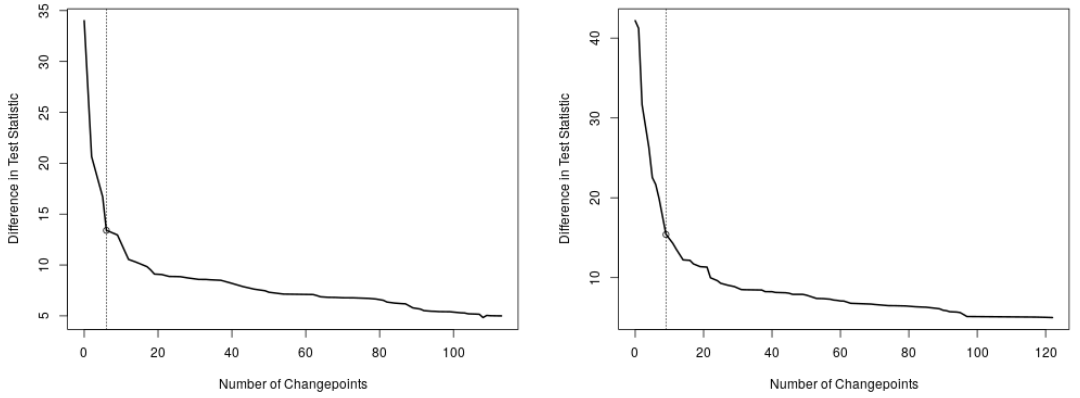
Table A.8.2: TDR and FDR for GeomCP and E-Divisive for simulated data sets containing variance changes that occur in all series. Φ relates to the size of the change.

			GeomCP		E-Divisive	
n	p	Φ	TDR	FDR	TDR	FDR
200	100	2.5	**0.844**	0.113	0.054	**0.064**
		3	**0.960**	**0.076**	0.086	0.100
		3.5	**0.966**	**0.072**	0.188	0.148
	500	2.5	**0.882**	0.100	0.016	**0.070**
		3	**0.948**	0.090	0.030	**0.063**
		3.5	**0.976**	**0.061**	0.044	0.079
1000	100	2.5	**0.804**	0.225	0.373	**0.247**
		3	**0.915**	**0.177**	0.433	0.239
		3.5	**0.962**	0.166	0.500	**0.156**
	500	2.5	**0.844**	**0.204**	0.173	0.506
		3	**0.944**	**0.169**	0.280	0.331
		3.5	**0.982**	**0.157**	0.367	0.267



(a) Distance (b) Angle

Figure A.9.1: CROPS diagnostics plots for distance and angle measure of comparative genomic hybridization data where the vertical line and circled point indicates the elbow of the plot we use for the number of changepoints



(a) Distance (b) Angle

Figure A.9.2: CROPS diagnostics plots for distance and angle measure of S&P500 log-returns data where the vertical line and circled point indicates the elbow of the plot we use for the number of changepoints

Appendix B

Subspace Changepoint Detection in Multivariate Time Series

B.1 Additional Simulation Results: ROC Curves

In this appendix we show additional ROC curves, accompanying those in Section 4.3.2.

First, we show the ROC curves for varying values of σ_s^2 and σ_ϵ^2. The variation of σ_s^2 and σ_ϵ^2 makes the estimation of the subspace easier or harder. If we increase the signal-to-noise ratio, $\rho = \frac{\sigma_s^2}{\sigma_\epsilon^2}$ the estimation of the subspace will be easier as the signal-to-noise ratio within the subspace is greater. We kept the remaining parameters the same as in Section 4.3.2 ($n = 200$, $p = 20$, $q = 5$). Figure B.1.1 shows that along the diagonals, where ρ is constant, the ROC curves are similar indicating that as long as

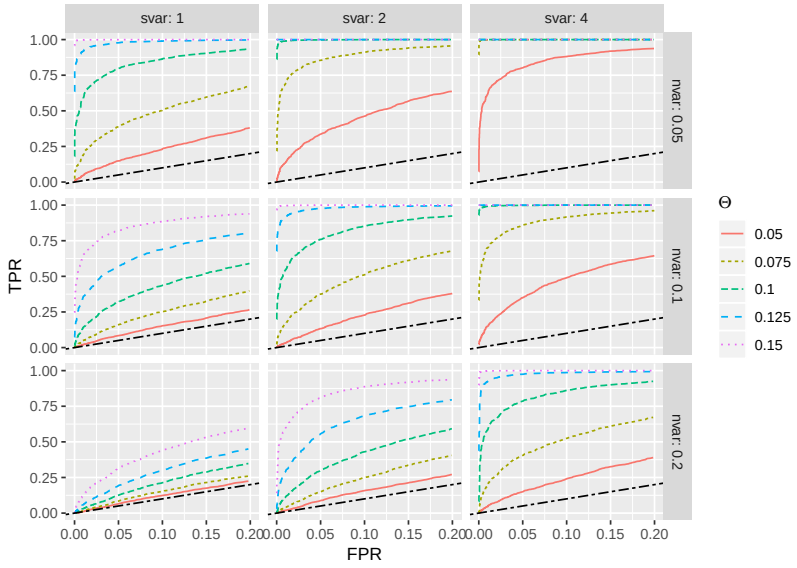


Figure B.1.1: ROC curves for varying signal to noise ratios within each subspace.

the ratio is the same we obtain similar power. Furthermore, as ρ increases (figures going from left to right or bottom to top) then the test statistic has more power as expected.

Next, we show the ROC curves for varying p and q. As the change size, Δ, is dependent upon the size of q it is harder to compare these scenarios. The remaining parameters are the same as in Section 4.3.2 ($n = 200$, $\sigma_s^2 = 1$, $\sigma_\epsilon^2 = 0.05$). Figure B.1.2 shows that if q is constant then we marginally lose power as p increases. On the other hand, if p is kept constant, as q increases we gain power, however, this could be linked to the increasing change size as q increases. A thorough theoretical investigation of the effect of p and q on the test statistic may recover an underlying relationship between the two, however, this is beyond the scope of this paper.

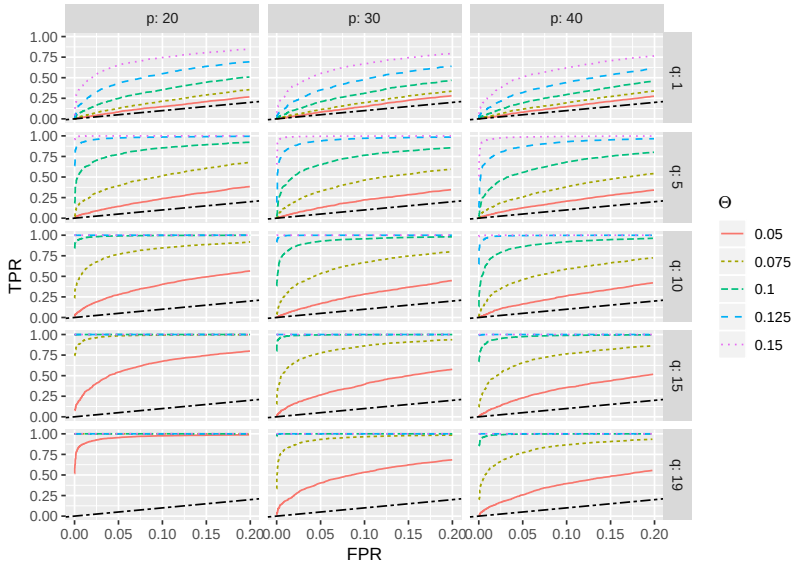


Figure B.1.2: ROC curves for scenarios with varying p and q.

B.2 Additional Simulation Results: Permutation Test

In this appendix we show extra simulation scenarios, accompanying those in Section 4.3.3, illustrating the effectiveness of the permutation test for determining significant changepoints.

First, we consider the effectiveness of the permutation test at controlling the FPR for different scenario setups. We vary p and q and show the FPR for varying n and number of permutations P, with the rest of the parameters the same as in Section 4.3.3 ($\sigma_s^2 = 1$, $\sigma_\epsilon^2 = 0.05$). Figure B.2.1 shows that for all the scenarios the FPR is being controlled close to the desired FPR of 0.05.

Next, we show the permutation test still maintains substantial power for detecting

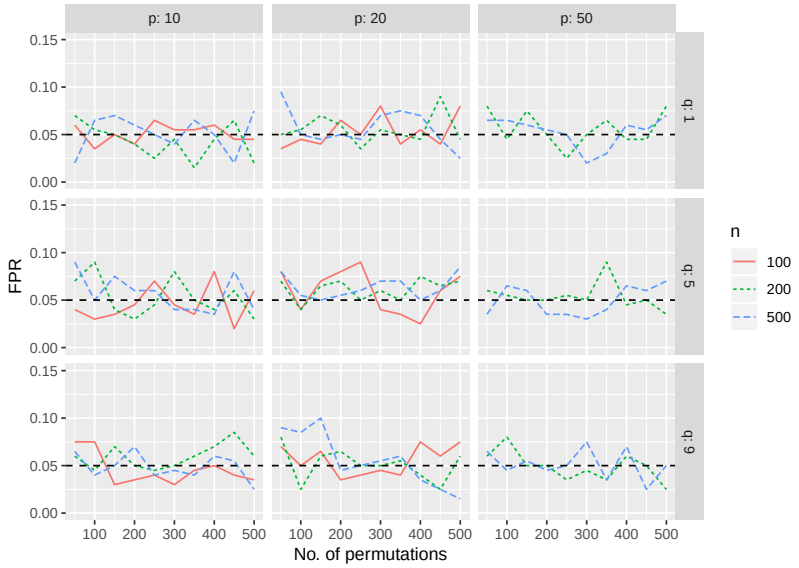


Figure B.2.1: FPR for data containing no changes under multiple different scenarios

changes when they exist. Again we vary p and q while allowing n to vary and set the number of permutations as $P = 200$ as Section 4.3.3 and Figure B.2.1 indicate is appropriate. The rest of the parameters are the same as in Section 4.3.3 ($\sigma_s^2 = 1$, $\sigma_\epsilon^2 = 0.05$, $\alpha = 0.05$). Again we deem a changepoint to be correct if it is within 20 time points on either side of the true change location. Figure B.2.2 shows that as q increases up to $p/2$ then the power for detecting a change decreases, however, looking at the scenario where $p = 10$ and $q = 9$, it seems that the power then increases as q goes from $p/2$ to p. This was verified in other scenarios as well. In all scenarios for $\Theta \geq 0.3$ the method has substantial power.

B.3 Application: Justification of Parameter Choices

In this appendix, we justify the parameter choices made in Section 4.5.

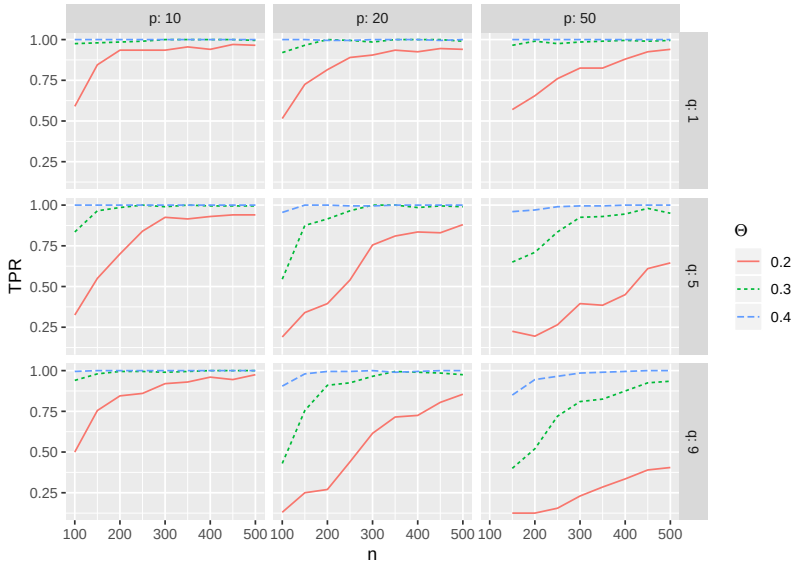


Figure B.2.2: TPR for data containing a changepoint under multiple scenarios.

In Section 4.5, we set the subspace dimension as $q = 9$ and concluded that $m = 6$ was an appropriate number of changepoints; here we justify these choices. First, we explore the choice of setting $q = 9$. To determine an appropriate subspace dimension we want to find the last 'large' eigenvalue, where afterwards all the remaining are 'small', and the number of 'large' eigenvalues is used as our subspace dimension q. To do this we made the assumption there was no changepoint within the first 400 time points and estimated the covariance of this data. Figure B.1(a) shows a scree plot of the ordered eigenvalues of this covariance matrix. There are a few points on this plot where we could argue the remaining eigenvalues are small, however, based upon our findings in Section 4.3.5 we prefer a larger subspace dimension. Hence, we choose the subspace dimension as $q = 9$ as the remaining eigenvalues are similar and small. We note Jiao et al. (2018) set the subspace size as $q = 5$, however, based upon our analysis this doesn't seem appropriate for this trial.

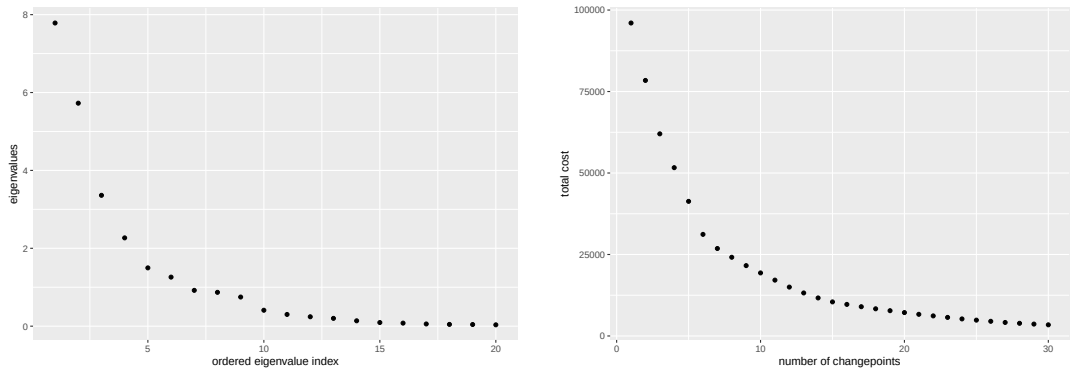


Figure B.3.1: (a) Scree plot showing the eigenvalues of the covariance matrix of the first 400 time points. (b) Total cost for an increasing number of changepoints.

To find an appropriate number of changepoints, we examined the total cost for the data defined as $\sum_{k=1}^{m+1} \mathcal{C}(y_{\tau_{k-1}+1:\tau k})$ for varying number of changepoints, m. We started with the most significant changepoint; calculated the total cost; split the data at this changepoint, and then kept adding the next most significant changepoint and repeating the process. Figure B.1(b) shows the total cost as extra changepoints are added. After adding 6 changepoints, the addition of more resulted in a marginal reduction in total cost. This indicates the most appropriate number of changepoints is $m = 6$.

Bibliography

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 19(6):716–723. 24

Anderson, T. W. (1963). Asymptotic Theory for Principal Component Analysis. *The Annals of Mathematical Statistics*, 34(1):122–148. 86

Andrews, D. W. K. (1991). Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation. *Econometrica*, 59(3):817–858. 99, 138

Aue, A., Dienes, C., Fremdt, S., and Steinebach, J. (2015). Reaction Times of Monitoring Schemes for ARMA Time Series. *Bernoulli*, 21(2):1238–1259. 113, 121, 131, 133

Aue, A., Hörmann, S., Horváth, L., and Reimherr, M. (2009a). Break Detection in the Covariance Structure of Multivariate Time Series Models. *The Annals of Statistics*, 37(6B):4046–4087. 36, 37, 81, 98, 99

Aue, A. and Horváth, L. (2004). Delay Time in Sequential Detection of Change. *Statistics & Probability Letters*, 67(3):221–231. 116

Aue, A., Horváth, L., and Reimherr, M. L. (2009b). Delay Times of Sequential Procedures for Multiple Time Series Regression Models. *Journal of Econometrics*, 149(2):174–190. 113

Auger, I. E. and Lawrence, C. E. (1989). Algorithms for the Optimal Identification of Segment Neighborhoods. *Bulletin of Mathematical Biology*, 51(1):39–54. 15, 104

Avanesov, V. and Buzun, N. (2018). Change-Point Detection in High-Dimensional Covariance Structure. *Electronic Journal of Statistics*, 12(2):3254–3294. 38, 81, 98, 99

Bai, J. (1994). Least Squares Estimation of a Shift in Linear Processes. *Journal of Time Series Analysis*, 15(5):453–472. 11

Bai, J. (1996). Testing for Parameter Constancy in Linear Regressions: An Empirical Distribution Function Approach. *Econometrica*, 64(3):597–622. 11

Bai, J. (2000). Vector Autoregressive Models with Structural Changes in Regression Coefficients and in Variance-Covariance Matrices. *CEMA Working Papers*. 11

Bai, J., Han, X., and Shi, Y. (2020). Estimation and Inference of Change Points in High-Dimensional Factor Models. *Journal of Econometrics*, 219(1):66–100. 157

Bai, J. and Perron, P. (1998). Estimating and Testing Linear Models with Multiple Structural Changes. *Econometrica*, 66(1):47–78. 3

Baranowski, R., Chen, Y., and Fryzlewicz, P. (2019). Narrowest-Over-Threshold Detection of Multiple Change Points and Change-Point-Like Features. *Journal of*

the Royal Statistical Society: Series B (Statistical Methodology), 81(3):649–672. 21, 25

Bardwell, L., Fearnhead, P., Eckley, I. A., Smith, S., and Spott, M. (2019). Most Recent Changepoint Detection in Panel Data. *Technometrics*, 61(1):88–98. 43

Barigozzi, M., Cho, H., and Fryzlewicz, P. (2018). Simultaneous Multiple Change-Point and Factor Analysis for High-Dimensional Time Series. *Journal of Econometrics*, 206(1):187–225. 157

Basseville, M. and Nikiforov, I. V. (1993). *Detection of Abrupt Changes: Theory and Application*. prentice Hall Englewood Cliffs. 112

Bemporad, A., Breschi, V., Piga, D., and Boyd, S. P. (2018). Fitting Jump Models. *Automatica*, 96:11–21. 30

Bleakley, K. and Vert, J.-P. (2011). The Group Fused LASSO for Multiple Change-Point Detection. *arXiv*, 1106.4199. 3, 50, 74

Blythe, D. A. J., von Bunau, P., Meinecke, F. C., and Muller, K.-R. (2012). Feature Extraction for Change-Point Detection Using Stationary Subspace Analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 23(4):631–643. 82

Bolton, A. D. and Heard, N. A. (2018). Malware Family Discovery Using Reversible Jump MCMC Sampling of Regimes. *Journal of the American Statistical Association*, 113(524):1490–1502. 40

Brodsky, E. and Darkhovsky, B. S. (2013). *Nonparametric Methods in Change Point Problems*. Springer Netherlands. 6, 50, 81

Cabrieto, J., Tuerlinckx, F., Kuppens, P., Grassmann, M., and Ceulemans, E. (2017). Detecting Correlation Changes in Multivariate Time Series: A Comparison of Four Non-Parametric Change Point Detection Methods. *Behavior Research Methods*, 49(3):988–1005. 40

Carroll, R., Lawson, A., and Zhao, S. (2019). A Data-Driven Approach for Estimating the Change-Points and Impact of Major Events on Disease Risk. *Spatial and Spatio-temporal Epidemiology*, 29:111–118. 3

Celisse, A., Marot, G., Pierre-Jean, M., and Rigaill, G. (2018). New Efficient Algorithms for Multiple Change-Point Detection With Reproducing Kernels. *Computational Statistics & Data Analysis*, 128:200–220. 13

Chakar, S., Lebarbier, E., Lévy-Leduc, C., and Robin, S. (2017). A Robust Approach for Estimating Change-Points in the Mean of an AR(1) Process. *Bernoulli*, 23(2):1408–1447. 11

Chapman, J.-L. and Killick, R. (2020). An Assessment of Practitioners Approaches to Forecasting in the Presence of Changepoints. *Quality and Reliability Engineering International*, 36(8):2676–2687. 111

Chen, K.-M., Cohen, A., and Sackrowitz, H. (2011). Consistent Multiple Testing for Change Points. *Journal of Multivariate Analysis*, 102(10):1339–1343. 20

Chen, L. (2015). Estimating the Common Break Date in Large Factor Models. *Economics Letters*, 131:70–74. 157

Chib, S. (1998). Estimation and Comparison of Multiple Change-Point Models. *Journal of Econometrics*, 86(2):221–241. 30

Cho, H. (2016). Change-Point Detection in Panel Data via Double CUSUM Statistic. *Electronic Journal of Statistics*, 10(2):2000–2038. 46

Cho, H. and Fryzlewicz, P. (2012). Multiscale and Multilevel Technique for Consistent Segmentation of Nonstationary Time Series. *Statistica Sinica*, 22(1):207—-229. 20

Cho, H. and Fryzlewicz, P. (2015). Multiple-Change-Point Detection for High Dimensional Time Series via Sparsified Binary Segmentation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2):475–507. 45

Cho, H. and Kirch, C. (2021). Bootstrap Confidence Intervals for Multiple Change Points Based on Moving Sum Procedures. *arXiv*, 2106.12844. 22, 38

Chu, C.-S. J., Stinchcombe, M., and White, H. (1996). Monitoring Structural Change. *Econometrica*, 64(5):1045—-1065. 112, 113, 115

Csörgö, M. and Horváth, L. (1988). Nonparametric Methods for Changepoint Problems. In *Handbook of Statistics*, volume 7, chapter 20, pages 403–425. 12

Csörgö, M. and Horváth, L. (1997). *Limit Theorems in Change-Point Analysis*. John Wiley & Sons. 6, 13, 26, 112

Dehning, J., Zierenberg, J., Spitzner, F. P., Wibral, M., Neto, J. P., Wilczek, M., and Priesemann, V. (2020). Inferring Change Points in the Spread of COVID-19 Reveals the Effectiveness of Interventions. *Science*, 369(eabb9789). 3

Dette, H., Pan, G., and Yang, Q. (2020). Estimating a Change Point in a Sequence of Very High-Dimensional Covariance Matrices. *Journal of the American Statistical Association*, pages 1–11. 37

Duan, J., Bai, J., and Han, X. (2021). Quasi-Maximum Likelihood Estimation of Break Point in High-Dimensional Factor Models. *arXiv*, 2102.12666. 157

Eckley, I. A., Fearnhead, P., and Killick, R. (2011). Analysis of Changepoint Models. In Barber, D., Cemgil, A. T., and Chiappa, S., editors, *Bayesian Time Series Models*, chapter 10, pages 205–224. Cambridge University Press. 6, 50, 68, 81

Eichinger, B. and Kirch, C. (2018). A MOSUM Procedure for the Estimation of Multiple Random Change Points. *Bernoulli*, 24(1):526–564. 22

Enikeeva, F. and Harchaoui, Z. (2019). High-Dimensional Change-Point Detection Under Sparse Alternatives. *The Annals of Statistics*, 47(4):2051–2079. 34, 47, 50, 81

Fearnhead, P. (2006). Exact and Efficient Bayesian Inference for Multiple Changepoint Problems. *Statistics and Computing*, 16(2):203–213. 29

Fearnhead, P., Maidstone, R., and Letchford, A. (2018). Detecting Changes in Slope With an L_0 Penalty. *Journal of Computational and Graphical Statistics*, 28(2):265–275. 57

Fildes, R. (1985). Quantitative Forecasting-The State of the Art: Econometric Models. *The Journal of the Operational Research Society*, 36(7):549—-580. 110

Fildes, R., Nikolopoulos, K., Crone, S. F., and Syntetos, A. A. (2008). Forecasting and Operational Research: A Review. *Journal of the Operational Research Society*, 59(9):1150–1172. 110

Fisch, A. T. M., Eckley, I. A., and Fearnhead, P. (2018). A Linear Time Method for the Detection of Point and Collective Anomalies. *arXiv*, 1806.01947. 58

Fisch, A. T. M., Eckley, I. A., and Fearnhead, P. (2019). Subset Multivariate Collective and Point Anomaly Detection. *arXiv*, 1909.01691. 43

Foret, P. (2019). *SP500R: Easy Loading of SP500 Stocks Data*. Github R package version 0.1.0. 77

Fremdt, S. (2014). Asymptotic Distribution of the Delay Time in Page's Sequential Procedure. *Journal of Statistical Planning and Inference*, 145:74–91. 118, 119, 120

Fremdt, S. (2015). Page's Sequential Procedure for Change-Point Detection in Time Series Regression. *Statistics*, 49(1):128–155. 118

Frick, K., Munk, A., and Sieling, H. (2014). Multiscale Change Point Inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3):495–580. 29

Fryzlewicz, P. (2014). Wild Binary Segmentation for Multiple Change-Point Detection. *The Annals of Statistics*, 42(6):2243–2281. 20, 21, 25, 51, 63, 105

Gallagher, C., Lund, R., and Robbins, M. (2013). Changepoint Detection in Climite Time Series with Long-Term Trends. *Journal of Climate*, 26(14):4994–5006. 28

Gama, J., Medas, P., Castillo, G., and Rodrigues, P. (2004). Learning with drift detection. In Bazzan, A. L. C. and Labidi, S., editors, *Advances in Artificial Intelligence – SBIA 2004*, pages 286–295, Berlin, Heidelberg. Springer Berlin Heidelberg. 114

Gombay, E. and Serban, D. (2009). Monitoring Parameter Change in AR(p) Time Series Models. *Journal of Multivariate Analysis*, 100(4):715–725. 113

Groen, J. J. J., Kapetanios, G., and Price, S. (2013). Multivariate Methods for Monitoring Structural Change. *Journal of Applied Econometrics*, 28(2):250–274. 44

Grundy, T., Killick, R., and Mihaylov, G. (2020). High-Dimensional Changepoint Detection via a Geometrically Inspired Mapping. *Statistics and Computing*, 30(4):1155–1166. 81

Hallac, D., Nystrup, P., and Boyd, S. (2019). Greedy Gaussian Segmentation of Multivariate Time Series. *Advances in Data Analysis and Classification*, 13(3):727–751. 33

Harchaoui, Z. and Cappe, O. (2007). Retrospective Mutiple Change-Point Estimation With Kernels. In *2007 IEEE/SP 14th Workshop on Statistical Signal Processing*, pages 768–772. IEEE. 13

Harchaoui, Z. and Lévy-Leduc, C. (2010). Multiple Change-Point Estimation With

a Total Variation Penalty. *Journal of the American Statistical Association*, 105(492):1480–1493. 25

Haynes, K., Eckley, I. A., and Fearnhead, P. (2017a). Computationally Efficient Changepoint Detection for a Range of Penalties. *Journal of Computational and Graphical Statistics*, 26(1):134–143. 24, 26, 76, 77

Haynes, K., Fearnhead, P., and Eckley, I. A. (2017b). A Computationally Efficient Nonparametric Approach for Changepoint Detection. *Statistics and Computing*, 27(5):1293–1305. 11, 58, 61, 71, 76, 167

Haynes, K. and Killick, R. (2021). *changepoint.np: Methods for Nonparametric Changepoint Detection*. R package version 1.0.3. 76

Hinkley, D. V. (1971). Inference About the Change-Point From Cumulative Sum Tests. *Biometrika*, 58(3):509. 12

Horváth, L. and Hušková, M. (2012). Change-Point Detection in Panel Data. *Journal of Time Series Analysis*, 33(4):631–648. 35, 44, 47, 50

Horváth, L., Hušková, M., Kokoszka, P., and Steinebach, J. (2004). Monitoring Changes in Linear Models. *Journal of Statistical Planning and Inference*, 126(1):225–251. 113, 118, 121

Hotelling, H. (1933). Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology*, 24(6):417–441. 81

Hušková, M. and Slaby, A. (2001). Permutation Tests for Multiple Changes. *Kybernetika*, 37(5):605–622. 22

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., and Yasmeen, F. (2021). *forecast: Forecasting Functions for Time Series and Linear Models*. R package version 8.14. 138

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. 12, 121, 126, 137

Jackson, B., Scargle, J., Barnes, D., Arabhi, S., Alt, A., Gioumousis, P., Gwin, E., San, P., Tan, L., and Tun Tao Tsai (2005). An Algorithm for Optimal Partitioning of Data on an Interval. *IEEE Signal Processing Letters*, 12(2):105–108. 16, 23, 104

James, N. A. and Matteson, D. S. (2014). ecp : An R Package for Nonparametric Multiple Change Point Analysis of Multivariate Data. *Journal of Statistical Software*, 62(7):1–25. 63, 75

Jiao, S., Shen, T., Yu, Z., and Ombao, H. (2021). Change-point Detection Using Spectral PCA for Multivariate Time Series. *arXiv*, 2101.04334. 157

Jiao, Y., Chen, Y., and Gu, Y. (2018). Subspace Change-Point Detection: A New Model and Solution. *IEEE Journal of Selected Topics in Signal Processing*, 12(6):1224–1239. 82, 85, 91, 93, 97, 177

Jirak, M. (2015). Uniform Change Point Tests in High Dimension. *The Annals of Statistics*, 43(6):2451–2483. 45, 50, 81

Johnstone, I. M. (2001). On the Distribution of the Largest Eigenvalue in Principal Components Analysis. *The Annals of Statistics*, 29(2):295–327. 82

Jolliffe, I. (2002). *Principal Component Analysis*. Springer series in statistics. Springer, New York, 2nd ed. edition. 81, 86

Jones, R. H. and Dey, I. (1995). Determining One or More Change Points. *Chemistry and Physics of Lipids*, 76(1):1–6. 24

Jones, S. A., Joy, M. P., and Pearson, J. (2002). Forecasting Demand of Emergency Care. *Health Care Management Science*, 5(4):297–305. 110

Kawahara, Y., Yairi, T., and Machida, K. (2007). Change-Point Detection in Time-Series Data Based on Subspace Identification. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 559–564. IEEE. 82

Killick, R. and Eckley, I. A. (2014). changepoint: An R Package for Changepoint Analysis. *Journal of Statistical Software*, 58(3). 62

Killick, R., Eckley, I. A., Ewans, K., and Jonathan, P. (2010). Detection of Changes in Variance of Oceanographic Time-Series Using Changepoint Analysis. *Ocean Engineering*, 37(13):1120–1126. 3

Killick, R., Fearnhead, P., and Eckley, I. A. (2012). Optimal Detection of Change-

points With a Linear Computational Cost. *Journal of the American Statistical Association*, 107(500):1590–1598. 18, 20, 23, 57, 59, 104

Killick, R., Haynes, K., and Eckley, I. A. (2016). *changepoint: An R package for changepoint analysis*. R package version 2.2.2. 77

Ko, S. I. M., Chong, T. T. L., and Ghosh, P. (2015). Dirichlet Process Hidden Markov Multiple Change-Point Model. *Bayesian Analysis*, 10(2):275–296. 11, 30

Kovács, S., Li, H., Haubner, L., Munk, A., and Bühlmann, P. (2020). Optimistic Search Strategy: Change Point Detection for Large-Scale Data via Adaptive Logarithmic Queries. *arXiv*, 2010.10194. 105

Krasheninnikov, V. R., Klyachkin, V. N., and Kuvayskova, Y. E. (2018). Models Updating for Technical Objects State Forecasting. In *2018 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC)*, pages 1–4. IEEE. 111

Krishnaiah, P. and Miao, B. (1988). Review About Estimation of Change Points. In *Handbook of Statistics*, volume 7, chapter 19, pages 375–402. Elsevier. 11

Lavielle, M. and Moulines, E. (2000). Least-Squares Estimation of an Unknown Number of Shifts in a Time Series. *Journal of Time Series Analysis*, 21(1):33–59. 11

Lavielle, M. and Teyssière, G. (2006). Detection of Multiple Change-Points in Multivariate Time Series. *Lithuanian Mathematical Journal*, 46(3):287–306. 33, 36

Lebarbier, E. (2005). Detecting Multiple Change-Points in the Mean of Gaussian Process by Model Selection. *Signal Processing*, 85(4):717–736. 25

Lévy-Leduc, C. and Roueff, F. (2009). Detection and Localization of Change-Points in High-Dimensional Network Traffic Data. *The Annals of Applied Statistics*, 3(2):637–662. 13

Ljung, L. and Söderström, T. (1983). *Theory and practice of recursive identification*, volume 4 of *The MIT press series in signal processing, optimization, and control*. MIT Press, United States. 4

Lung-Yut-Fong, A., Lévy-Leduc, C., and Cappé, O. (2011). Distributed Detection/Localization of Change-Points in High-Dimensional Network Traffic Data. *Statistics and Computing 2011 22:2*, 22(2):485–496. 13

Luong, T. M., Rozenholc, Y., and Nuel, G. (2013). Fast Estimation of Posterior Probabilities in Change-Point Analysis Through a Constrained Hidden Markov Model. *Computational Statistics & Data Analysis*, 68:129–140. 30

Lyapunov, A. M. (1901). Une Proposition Générale du Calcul des Probabilités. *Comptes rendus hebdomadaires de l'Académie des Sciences de Paris*. 161

Maboudou, E. and Hawkins, D. M. (2009). Fitting Multiple Change-Point Models to a Multivariate Gaussian Model. In *Proceedings of the Second International Workshop in Sequential Methodologies (IWSM 2009)*, pages 1–5. 33

Maboudou-Tchao, E. M. and Hawkins, D. M. (2013). Detection of Multiple Change-

Points in Multivariate Data. *Journal of Applied Statistics*, 40(9):1979–1995. 42, 45, 50, 80

Mahalanobis, P. C. (1936). On the Generalized Distance in Statistics. National Institute of Science of India. 61

Maidstone, R., Hocking, T., Rigaill, G., and Fearnhead, P. (2017). On Optimal Multiple Changepoint Algorithms for Large Data. *Statistics and Computing*, 27(2):519–533. 16, 18

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. 13, 14, 28, 40, 50, 51, 62, 63, 75, 89

Modisett, M. C. and Maboudou-Tchao, E. M. (2010). Significantly Lower Estimates of Volatility Arise From the Use of Open-High-Low-Close Price Data. *North American Actuarial Journal*, 14(1):68–85. 3, 49

Ng, C. T., Lee, W., and Lee, Y. (2018). Change-Point Estimators With True Identification Property. *Bernoulli*, 24(1):616–660. 25

Nugent, C. (2018). *SP 500 Stock Data*. Kaggle dataset version 4. 77

Nystrup, P., Kolm, P. N., and Lindström, E. (2021). Feature Selection in Jump Models. *Expert Systems with Applications*, 184:115558. 30

Nystrup, P., Lindström, E., and Madsen, H. (2020). Learning Hidden Markov Models

With Persistent States by Penalizing Jumps. *Expert Systems with Applications*, 150:113307. 30

Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004). Circular Binary Segmentation for the Analysis of Array-Based DNA Copy Number Data. *Biostatistics*, 5(4):557–572. 21, 35

Ordu, M., Demir, E., and Tofallis, C. (2020). A Decision Support System for Demand and Capacity Modelling of an Accident and Emergency Department. *Health Systems*, 9(1):31–56. 110

Padmore, J. (1993). The Analysis of Stratigraphic Data with Particular Reference to Zonation Problems. In Opitz, O., Lausen, B., and Klar, R., editors, *Infomation and Classification*, pages 490–499. Springer, Berlin, Heidelberg. 21

Page and E. S. (1955). A Test for a Change in a Parameter Occurring at an Unknown Point. *Biometrika*, 42(3-4):523–527. 11

Page, E. S. (1954). Continuous Inspection Schemes. *Biometrika*, 41(1/2):100. 4, 6, 12, 50, 81, 112, 115

Parsons, L., Haque, E., and Liu, H. (2004). Subspace Clustering for High Dimensional Data. *ACM SIGKDD Explorations Newsletter*, 6(1):90–105. 81

Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572. 81

Pein, F., Sieling, H., and Munk, A. (2017). Heterogeneous Change Point Inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1207–1227. 11, 29

Peluso, S., Chib, S., and Mira, A. (2019). Semiparametric Multivariate and Multiple Change-Point Modeling. *Bayesian Analysis*, 14(3):727–751. 40

Peng, T., Leckie, C., and Ramamohanarao, K. (2004). Proactively Detecting Distributed Denial of Service Attacks Using Source IP Address Monitoring. In Mitrou, N., Kontovasilis, K., Rouskas, G. N., Iliadis, I., and Merakos, L., editors, *Networking 2004*, pages 771–782, Berlin, Heidelberg. Springer Berlin Heidelberg. 3, 80

Pickering, B. (2016). *Changepoint Detection for Acoustic Sensing Signals*. PhD thesis, Lancaster University. 43

R Core Team (2019). *R: A Language and Enviroment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 62

Reeves, J., Chen, J., Wang, X. L., Lund, R., and Lu, Q. (2007). A Review and Comparison of Changepoint Detection Techniques for Climate Data. *Journal of Applied Meteorology and Climatology*, 46:900–915. 24

Robbins, M., Gallagher, C., Lund, R., and Aue, A. (2011). Mean Shift Testing in Correlated Data. *Journal of Time Series Analysis*, 32(5):498–511. 12, 114

Ross, G. J., Adams, N. M., Tasoulis, D. K., and Hand, D. J. (2012). Exponentially

weighted moving average charts for detecting concept drift. *Pattern Recognition Letters*, 33(2):191–198. 114

Ryan, S. and Killick, R. (2021). Detecting Changes in Covariance via Random Matrix Theory. *arXiv*, 2108.07340. 39

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461 – 464. 24

Scott, A. A. J. and Knott, M. (1974). A Cluster Analysis Method for Grouping Means in the Analysis of Variance. *Biometrics*, 30(3):507–512. 20, 51, 63, 104

Shen, D., Shen, H., and Marron, J. S. (2016). A General Framework for Consistency of Principal Component Analysis. *Journal of Machine Learning Research*, 17(150):1–34. 86

Shi, X., Gallagher, C., Lund, R., and Killick, R. (2021). A Comparison of Single and Multiple Changepoint Techniques for Time Series Data. *arXiv*, 2101.01960. 105

Siegmund, D., Yakir, B., and Zhang, N. R. (2011). Detecting Simultaneous Variant Intervals in Aligned Sequences. *The Annals of Applied Statistics*, 5(2A):645–668. 44

Srivastava, M. S. and Worsley, K. J. (1986). Likelihood Ratio Tests for a Change in the Multivariate Normal Mean. *Journal of the American Statistical Association*, 81(393):199. 32

Steward, R. M., Rigdon, S. E., and Pan, R. (2016). A Bayesian Approach to Diagnostics for Multivariate Control Charts. *Journal of Quality Technology*, 48(4):303–325. 40

Svetunkov, I. (2021). smooth: Forecasting Using State Space Models. R package version 3.1.0. 138

Syntetos, A. A., Babai, Z., Boylan, J. E., Kolassa, S., and Nikolopoulos, K. (2016). Supply Chain Forecasting: Theory, Practice, Their Gap and the Future. *European Journal of Operational Research*, 252(1):1–26. 110

Taib, A., Killick, R., Hussain, K., Patel, H., and Obeidallah, M. R. (2021). Is There Seasonal Variation in Gallstone Related Admissions in England? *HPB*. 149, 151

Tartakovsky, A., Nikiforov, I. V., and Basseville, M. (2014). *Sequential Analysis: Hypothesis Testing and Changepoint Detection*. CRC Press. 34, 112, 117

Terrera, G. M., van den Hout, A., and Matthews, F. E. (2011). Random Change Point Models: Investigating Cognitive Decline in the Presence of Missing Data. *Journal of Applied Statistics*, 38(4):705–716. 50

Tibshirani, R. and Wang, P. (2008). Spatial Smoothing and Hot Spot Detection for CGH Data Using the Fused LASSO. *Biostatistics*, 9(1):18–29. 25

Tickle, S. O., Eckley, I. A., and Fearnhead, P. (2021). A Computationally Efficient, High-Dimensional Multiple Changepoint Procedure With Application to Global Terrorism Incidence. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 00. 43

Tickle, S. O., Eckley, I. A., Fearnhead, P., and Haynes, K. (2020). Parallelization of a Common Changepoint Detection Method. *Journal of Computational and Graphical Statistics*, 29(1):149–161. 28, 58

Truong, C., Oudre, L., and Vayatis, N. (2020). Selective Review of Offline Change Point Detection Methods. *Signal Processing*, 167:107299. 6, 50, 81

Venkatraman, E. S. (1992). *Consistency Results in Multiple Change-Point Problems*. PhD thesis, Stanford University. 20

Venkatraman, E. S. and Olshen, A. B. (2007). A Faster Circular Binary Segmentation Algorithm for the Analysis of Array CGH Data. *Bioinformatics*, 23(6):657–663. 21

Vidal, R., Ma, Y., and Sastry, S. (2005). Generalized Principal Component Analysis (GPCA). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1945–1959. 83

Vostrikova, L. (1981). Detecting 'Disorder' in Multidimensional Random Processes. *Soviet Math Dokl*, 24:55–59. 51, 63, 104

Wang, D., Yu, Y., and Rinaldo, A. (2021). Optimal Covariance Change Point Localization in High Dimensions. *Bernoulli*, 27(1):554–575. 38, 81, 98

Wang, S. and Reynolds, M. R. (2013). A GLR Control Chart for Monitoring the Mean Vector of a Multivariate Normal Process. *Journal of Quality Technology*, 45(1):18–33. 34

Wang, T. and Samworth, R. (2016). *InspectChangepoint: High-Dimensional Changepoint Estimation via Sparse Projection*. R package version 1.0.1. 62

Wang, T. and Samworth, R. J. (2018). High Dimensional Change Point Estimation via Sparse Projection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):57–83. 28, 46, 50, 51, 62, 63, 75, 76, 81

Wickens, T. D. (1995). *The Geometry of Multivariate Statistics*. Lawrence Erlbaum Associates, Incorporated. 61

Xie, L., Xie, Y., and Moustakides, G. V. (2020). Sequential Subspace Change Point Detection. *Sequential Analysis*, 39(3):307–335. 82, 83

Yao, Y.-C. (1988). Estimating the Number of Change-Points via Schwarz' Criterion. *Statistics & Probability Letters*, 6(3):181–189. 24

Zamba, K. D. and Hawkins, D. M. (2006). A Multivariate Change-Point Model for Statistical Process Control. *Technometrics*, 48(4):539–549. 34

Zamba, K. D. and Hawkins, D. M. (2009). A Multivariate Change-Point Model for Change in Mean Vector and/or Covariance Structure. *Journal of Quality Technology*, 41(3):285–303. 34

Zeileis, A., Köll, S., and Graham, N. (2020). Various Versatile Variances: An Object-Oriented Implementation of Clustered Covariances in R. *Journal of Statistical Software*, 95(1). 138

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. 24, 62

Zhang, N. R., Siegmund, D. O., Ji, H., and Li, J. Z. (2010). Detecting Simultaneous Changepoints in Multiple Sequences. *Biometrika*, 97(3):631–645. 35, 44, 50, 80

Zou, C., Yin, G., Feng, L., and Wang, Z. (2014). Nonparametric Maximum Likelihood Approach to Multiple Change-Point Problems. *The Annals of Statistics*, 42(3):970–1002. 11, 58