Detecting Changepoints in Multivariate Data

Sean Ryan, B.Sc.(Hons.), M.Res

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

July 2020

STOR-i
excellence with impact

Abstract

In this thesis, we propose new methodology for detecting changepoints in multivariate data, focusing on the setting where the number of variables and the length of the data can be very large.

We begin by considering the problem of detecting changepoints where only a subset of the variables are affected by the change. Previous work demonstrated that the changepoint locations and affected variables can be simultaneously estimated by solving a discrete optimisation problem. We propose two new methods PSMOP (Pruned Subset Multivariate Optimal Partitioning) and SPOT (Subset Partitioning Optimal Time) for solving this problem. PSMOP uses novel search space reduction techniques to efficiently compute an exact solution for data of moderate size. SPOT is an approximate method, which gives near optimal solutions at a very low computational cost, and can be applied to very large datasets. We use this new methodology to study changes in sales data due to the effect of promotions.

We then examine the problem of detecting changes in the covariance structure of high dimensional data. Using results from Random Matrix Theory, we introduce a novel test statistic for detecting such changes. Importantly, under the null hypothesis of no change, the distribution of this test statistic is independent of the underlying covariance matrix. We utilise this test statistic to study changes in the amount of water on the surface of a plot of soil.

Acknowledgements

Firstly, I would like to thank all the staff at the STOR-i Center for Doctoral Training for providing a stimulating and supportive environment. In particular, I would like to thank Jonathan Tawn for his tireless efforts to support me, and every other student who has been a part of STOR-i.

I am very grateful for the financial support provided by the EPSRC and Tesco Plc. I would especially like to thank Trevor Sidery for his substantial commitment as my industrial supervisor as well as the Tesco team for making me feel so welcome on my visits.

I am extremely grateful to have been supervised by Rebecca Killick, who has worked so hard to guide me through the PhD process. I will always appreciate her support, mentorship and efforts to provide me with greater opportunities. I would also like to thank Prof. David Matteson of Cornell University who kindly took me on as a (virtual) visiting student in my final year.

My time in Lancaster would not have been the same without my cohort; Euan, Edwin, Hankui, Georgia, Rob and Zak. Their friendship has been one of the great joys of the PhD experience, which I will always treasure. I would also like to thank Emily, Jake, Kathryn, Lucy, Rob and Sam. I have greatly benefitted from their collective wisdom and experience.

Finally I would like to thank my parents and my family for their endless love and support.

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.

Sean Ryan

A version of Chapters 4 and 6 have been submitted for publication.

Chapter 3 is a collaboration wth my supervisor Rebecca Killick and her former student Ben Pickering.

Chapter 4 is a collaboration with my supervisor Rebecca Killick. We are grateful to Simon Mabon for guidance on the Syrian Civil War dataset.

Chapter 5 is a collaboration with my supervisor Rebecca Killick and my industrial supervisor Trevor Sidery.

Chapter 6 is a collaboration with my supervisor Rebecca Killick. We are grateful to Michael James and John Quinton for providing the soil data.

R code implementing the simulations included in this thesis is available upon request.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Due to advancements in technology, data of increasing complexity and size is being collected. Typically, this data is collected over periods of time and the behaviour of such data can change dramatically. If our statistical methodology does not take account of these changes, our capacity to model, understand and forecast the data will be significantly hampered. As a result there is substantial interest in developing new statistical methods that can capture and model data in a dynamic setting.

One approach to studying data which changes over time, is to assume that the data only changes at a small set of points, known as changepoints. This approach provides a natural way to extend standard models to the dynamic setting, and in many applications the changepoints themselves are interesting for practitioners. However while significant work has been completed on estimating changepoints for a single variable, less attention has been paid to the case where we have multiple variables.

In this thesis, we develop methodology for detecting changepoints in multivariate datasets, where there are potentially a very large number of variables under observation. We start in Chapter 2 by reviewing the literature on multivariate changepoints, focusing on the offline, frequentist setting which forms the basis of this thesis.

In Chapter 3 we consider the problem of detecting so called subset multivariate changepoints, where a change only affects a subset of the variables under observation. Previous work on this problem proposed a dynamic program, SMOP (Subset Multivariate Optimal Partitioning) which can simultaneously estimate the locations

of any changepoints and the set of variables affected by the change. However the computational cost of this procedure is substantial and it is infeasible for even small datasets. Therefore we propose a new dynamic program PSMOP (Pruned Subset Multivariate Optimal Partitioning), which utilises a number of novel search space reduction techniques to compute the same segmentation as SMOP at a substantially reduced computational cost.

Although it is considerably faster than its predecessor, the PSMOP procedure is still infeasible for datasets of moderate and large scale. Therefore in Chapter 4 we propose an approximate dynamic program, SPOT (Subset Partitioning Optimal Time). The computational cost of this procedure is, under mild conditions on the number of changes, linear in the dimension and length of the data and thus it can be applied to extremely large datasets. Furthermore we demonstrate that the loss of accuracy due to the approximation is very small in practice. In Chapter 5 we utilise the SPOT method to study changes in an industrial application.

A limitation of the subset multivariate approach is that it does not consider how the variables under observation relate to each other. In particular, it is not possible to detect changes in the relationships between variables. Therefore in Chapter 6, we examine the problem of detecting changes in the covariance structure of large data and, propose a new test statistic for detecting such changes in high dimensional data. The primary advantage of this method is that under the null hypothesis of no change, the distribution of the test statistic does not depend on the true covariance. We utilise this method to study changes in amount of water on the surface of soil.

We conclude the thesis with a discussion of the main contributions of this work and finally discuss a number of possible extensions to this research in Chapter 7.

Chapter 2

Literature Review

In this chapter, we review existing methodology for detecting changepoints in multivariate data. In particular, we focus on frequentist approaches to the offline multivariate changepoint problem. Changepoint detection has been a key area of research within the statistical literature for decades, having been first applied to quality control problems (Page, 1954). While much of the focus within this literature has been on the univariate changepoint problem, there has been a dramatic increase in interest in multivariate changepoint detection in recent years.

The changepoint literature can be separated into two distinct settings, the offline setting where all of the data is obtained prior to any analysis, and the online setting, where new data is observed over time. While there are clear connections between these settings, the primary issues considered in the two literatures are different. For example, in the offline setting we often need to identify multiple changepoints in the data. This is not typically the case in the online setting where the data generating process stops if a change occurs. Similarly, there is a particular focus in the online setting on detecting change as quickly as possible. This consideration is irrelevant in the offline setting. In this work, we focus exclusively on the offline setting. Readers interested in online setting should see Tartakovsky et al., 2014 for a thorough review.

The primary goal of this chapter is to provide a thorough review of existing methodology for detecting changepoints in multivariate data streams. As a necessary precursory step, we discuss some important contributions in the univariate change-

point literature. Note the goal of this discussion is not to give a complete review of univariate methods. Instead we focus on a small number of contributions which are particularly relevant to multivariate changepoint detection or this thesis specifically. In particular we review Binary Segmentation procedures (Section 2.2.1) and optimisation based search methods (Section 2.2.2). The latter discussion is of particular importance to Chapters 3 and 4, which builds on this body of literature. We then discuss recent advances in the multivariate changepoint problem. The multivariate changepoint literature considers a number of different types of changepoint problems and therefore we separate the literature by problem type. We consider the following types of changepoint problems; changes in mean (Section 2.3.1), changes in covariance (Section 2.3.2), changes in functional data (Section 2.3.3), nonparametric changes (Section 2.3.4), changes in vector autoregressive models (Section 2.3.5) and changes in network models (Section 2.3.6). We also highlight some advances in detecting changepoints in more specific data structures (Section 2.3.7).

2.1 Changepoint Model

Let $\{\boldsymbol{X}_i\}_{i=1}^n$ be a sequence of p dimensional random variables. Then a changepoint model for this sequence is given by

$$\begin{aligned} \boldsymbol{X}_t \sim F_k \text{ for } \tau_{k-1} < t \leq \tau_k \\ F_k \neq F_{k+1} \text{ for } 1 \leq k \leq m \\ 0 = \tau_0 < \tau_1 < \cdots < \tau_m < \tau_{m+1} = n \end{aligned} \tag{2.1.1}$$

where each $\boldsymbol{F}_k$ is a p dimensional data generating process. The goal in any changepoint analysis is to estimate the number of changepoints m and the locations of the changepoints, $\boldsymbol{\tau} := (\tau_1, \ldots, \tau_m)$. Changepoint models can be applied to univariate data ($p = 1$) and multivariate data ($p > 1$). The changepoint model above is general and works in the literature typically place some assumptions on the sequence of data generating processes $\{F_k\}_{k=1}^m$. For example, many authors consider the setting where each F_k belongs to the same family of distributions but differ in expectation. This is

known as the change in mean problem.

2.2 Univariate Multiple Changepoint Detection

We begin our discussion of univariate multiple changepoint detection methods by considering an important special case of (2.1.1), the At Most One Change (AMOC) setting where $m \leq 1$. The AMOC setting is the simplest version of the changepoint problem and a number of authors have developed methods for this problem under different assumptions. Methods include test statistics based on the likelihood ratio (J. Chen and Gupta, 1997; Hinkley, 1970) as well as test statistics based on normalized cumulative sums of functions of the data, also known as CUSUM statistics (Inclan and Tiao, 1994; Page, 1954). Significant theoretical work has been done analysing the behaviour of univariate changepoint tests in the AMOC setting. However this research is too broad to be covered in any detail here. Interested readers should refer to Csorgo and Horváth, 1997 for a thorough review.

There are two components to the AMOC setting, determining whether or not a change has occurred and if so identifying the location of the change. Given an appropriate likelihood function $\ell(\cdot)$ or cusum test statistic $T(\cdot)$, we can detect a single change in the univariate sequence $\{X_i\}_{i=1}^n$ by calculating

$$\max_{1\leq t\leq n} \ell(X_{1:t}) + \ell(X_{(t+1):n}) - \ell(X_{1:n}) \text{ or } \max_{1\leq t\leq n} T(X_{1:t}) - T(X_{(t+1):n}). \tag{2.2.1}$$

If this value exceeds a predefined threhold, then we say a change has occurred and an estimator for the location of the change is given by the optimiser of (2.2.1). While it is valuable to be able to detect changes in the AMOC setting, there are many settings where we need to be able to detect multiple changepoints. The remainder of this chapter considers methods for identifying multiple changepoints given a method for identifying a single change.

2.2.1 Binary Segmentation Methods

Scott and Knott, 1974 introduced the Binary Segmentation procedure, which extends tests for a single change to the multiple changepoint setting. The procedure first searches for a change in the entire dataset by applying a test such as in (2.2.1). If a change is detected then the test is applied separately to the data to the left and right of the change. This process continues recursively until no more changes are detected. A number of authors have studied the theoretical properties of the procedure (K. Chen et al., 2011; Fryzlewicz, 2014; Venkatraman, 1993). The Binary Segmentation procedure can be used with any test for a single changepoint and is computationally efficient with cost $\mathcal{O}(K(n)n\log n)$, where $K(n)$ is the cost of computing the likelihood function $\ell(\cdot)$ or cusum statistic $T(\cdot)$. For many common hypothesis tests (e.g. likelihood ratio test for a change in mean) these statistics can be computed based on summary statistics of the data and thus have $\mathcal{O}(1)$ cost. There are important limitations to the Binary Segmentation procedure. Firstly, if two changes move in opposite directions, they can mask each other and fail to be detected. Secondly as Binary Segmentation is a conditional search approach, if an early change is misspecified all future changepoint locations may also end up being misspecified. Despite these limitations, Binary Segmentation works well in practice and has been widely used in applications (Hernandez-Lopez and Rivera, 2014; Mahmoud et al., 2007).

In recent years, there have been a number of adaptations to the Binary Segmentation procedure that aim to address these issues. Olshen et al., 2004 proposed the Circular Binary Segmentation method, which addresses the issue of masking via a hypothesis test that fits two changepoints rather than one. Fryzlewicz, 2014 proposed the Wild Binary Segmentation method, which randomly samples M intervals and searches for a single changepoint over each interval. If a change τ is detected in the random interval (s, e), than the procedure is run again on the the intervals (s, τ) and $(\tau + 1, e)$. The procedure terminates if no more changes are detected. Fryzlewicz, 2020 introduced a more computationally efficient adaptation of the Wild Binary Segmentation procedure called Wild Binary Segmentation 2. This variant searches for a single change over a set of randomly drawn intervals. It then ranks any detected

changepoints by the size of test statistic. The largest candidate is added to the set of detected changes and the procedure is applied to the data to the left and right of this candidate. Again the procedure terminates when no more changes are detected. Note that since the intervals are drawn at random, there is no guarantee that the method will locate the same set of changepoints if the procedure is rerun on the data. Kovács et al., 2020 address this concern with the Seeded Binary Segmentation variant, which draws the intervals in a deterministic fashion. Finally we note that although Wild Binary Segmentation incorporates randomization, the purpose of this randomization is not uncertainty quantification and it would be incorrect to think of the method as a bootstrap type procedure.

Although the Binary Segmentation procedure was developed for the univariate setting, it is trivial to extend the procedure to the multivariate setting, and a number of authors have applied the technique to various multivariate changepoint problems, such as changes in covariance structure (Aue, Hörmann, et al., 2009; D. Wang, Yu, and Rinaldo, 2017) and changes in network structures (D. Wang, Yu, and Rinaldo, 2018). However there is an important class of multivariate changepoint problems where current Binary Segmentation approaches may be inappropriate. In particular, there is a growing interest in the literature on changepoint problems where not every variable is affected by a changepoint. While it is possible to conceive a Binary Segmentation type procedure for this setting, to our knowledge such a method is not widely available. Furthermore as we discuss in Section 2.4, there are significant advantages in jointly estimating the changepoints and the set of affected variables.

2.2.2 Multiple Changepoint Detection via Optimisation

The different Binary Segmentation procedures identify changepoints one by one with each subsequent changepoint conditional on the previously detected changes. However there are also methods that jointly estimate all the changepoint locations by solving an optimisation problem. Auger and Lawrence, 1989 examine the problem of detecting multiple changepoints where the number of changepoints m is known a priori, sometimes referred to as the constrained minimisation problem. Their method,

Segment Neighbourhood, selects the m changepoints which solve the following optimisation problem,

$$\underset{(\tau_1,\ldots,\tau_m)}{\arg\min} \sum_{k=0}^{m} \mathcal{C}(X_{(\tau_k+1):\tau_{k+1}}). \tag{2.2.2}$$

where $\mathcal{C}$ is a cost function measuring goodness of fit and as before $\tau_0 := 0$ and $\tau_{m+1} := n$. This optimisation problem can be solved using a dynamic program which has computational cost $\mathcal{O}(K(n)mn^2)$, where $K(n)$ is the cost of evaluating $\mathcal{C}$. Note that as with binary segmentation, many commonly used cost functions have computational cost $\mathcal{O}(1)$. Maidstone et al., 2017 and Rigaill, 2010, 2015 introduce dynamic programs which under certain conditions can solve the constrained minimisation problem at a substantially lower computational cost than the Segment Neighbourhood procedure. Of course in practice it is unlikely that the true number of changes is known a priori, thus practitioners typically solve (2.2.2) for multiple values of m and choose the value of m that minimises some criteria such as a penalised likelihood.

Jackson et al., 2005; Yao, 1988 consider the setting where the number of changepoints is unknown and jointly estimate the number and locations of changepoints by solving the following optimisation problem,

$$\underset{m,(\tau_1,\ldots,\tau_m)}{\arg\min} \sum_{k=0}^{m} \mathcal{C}(X_{(\tau_k+1):\tau_{k+1}}) + \beta f(m), \tag{2.2.3}$$

where β is a penalty to prevent overfitting of changepoints, and $f(m) = m$. This optimisation problem can be solved exactly via a dynamic program with computational cost $\mathcal{O}(K(n)n^2)$. We refer to (2.2.3) as a penalised cost function. Davis, Lee, et al., 2006 propose to detect changepoints via the principle of Minimum Description Length which can be formulated as a special case of (2.2.3). The authors introduce a genetic algorithm which provides accurate (although potentially suboptimal) solutions to the resulting optimisation problem. There has been significant work in developing techniques that reduce the computational cost of solving (2.2.3). If the cost function $\mathcal{C}$ is convex with respect to the data, the Pruned Exact Linear Time (PELT) method introduced by Killick, Fearnhead, et al., 2012 and the Functional Pruned Optimal Partitioning (FPOP) method of Maidstone et al., 2017 can solve (2.2.3) in linear time under mild conditons on the spread of the changepoints within the data. Fearnhead

and Rigaill, 2019 introduce the Robust Functional Pruned Optimal Partitioning (R-FPOP) which efficiently solves (2.2.3) for cost functions that are robust to outliers. Finally, a number of authors have studied the theoretical properties of optimisation based estimators (Tickle et al., 2020; Yao, 1988).

The penalised cost function approach has a number of advantageous properties. Firstly Killick, Fearnhead, et al., 2012 show that many hypothesis tests (such as all likelihood ratio tests) can be reformulated as penalised cost function problems, with the β penalty replacing the threshold. Under this framework Binary Segmentation procedures can be thought of as heuristic methods that produce sub optimal solutions, whereas exact methods such as Optimal Partitioning and PELT always produce the best possible solution. Secondly, since it utilises a generic cost function, it can be applied to a wide range of problems. In particular, optimisation based changepoint methods can be easily extended to the multivariate setting by utilising a multivariate cost function such as a multivariate likelihood function. However this approach has the same limitation in the multivariate setting as the Binary Segmentation methods; it is only appropriate if every variable is affected by each change and thus, not suitable for many of the applications we consider in this work. Finally we note that the penalised cost function approach is not necessarily applicable in all settings and there are certain test statistics that can not be formulated in the penalised cost function framework in (2.2.3), for example test statistics which are based on maximising the distance between segments.

2.3 Multivariate Changepoint Methods

This section reviews the literature for identifying changepoints in multivariate data. A naive implementation would be to consider each series independently to identify changepoints but this is an inefficient use of available information and, would likely lead to changepoints being missed and a larger error in the changepoints locations. We do not consider such an approach further here and instead describe methodology which explicitly considers the multivariate nature of the problem.

2.3.1 Changes in Mean

We begin by considering the literature on detecting a common change in mean across multiple series. Formally, we have the following model,

$$X_{t,j} = \mu_j + \delta_j I(t > \tau) + \epsilon_{t,j}$$

where $\mathbb{E}(\epsilon_{t,j}) = 0$, $\delta_j \in \mathbb{R}$ is the size of the change in variable j and $\tau \in \mathbb{Z}$ is the location of the changepoint. The goal of this literature is to exploit the common location of the changepoint, in order to detect changes more accurately than is possible in the univariate setting. For example, if $|\delta_j|$ is small for each j, then it will be difficult to detect a change by looking at each series individually. However, if the sum, $\sum_{j=1}^{p} |\delta_j|$ is large than we should be able to detect the changepoint by aggregating information across the series. Thus, there are two key questions in this literature, how can we efficiently combine information across different series, and to what extent does this improve changepoint estimation.

Much of the work in this area focuses on first applying a univariate changepoint test to each series, and then aggregating this information. The majority of authors consider aggregating the univariate CUSUM test originally derived by Page, 1954, however we note that a least squares approach has also been considered. The CUSUM test statistic for data $\{\boldsymbol{X}_i\}_{1 \leq i \leq b}$ is defined as,

$$T(t, j) := \sqrt{\frac{t(n-t)}{n}} \left(\frac{1}{n-t} \sum_{r=t+1}^{n} X_{j,r} - \frac{1}{t} \sum_{r=1}^{t} X_{j,r} \right) \tag{2.3.1}$$

The value $T(t, j)$ is the likelihood ratio test statistic for a change in the mean occuring at t. Note for the purposes of aggregation, authors utilise $T^2(t, j)$ to avoid positive and negative changes cancelling each other.

There are a range of different approaches for aggregating information across different series. The best approach depends heavily on the application. Throughout the rest of this subsection, we discuss papers that have examined this issue, with a focus on the settings where these methods are most appropriate. Note, some of these approaches incorporate extra algorithmic steps such as post processing or only rejecting the null hypothesis if the threshold is exceeded at multiple consecutive points. We do

not consider these concepts in this discussion, as they can be generalised to other test statistics and make it more difficult to compare methods. Similarly, although these methods can be extended to the multiple changepoint setting via the binary segmentation heuristic, we focus on the single changepoint case for simplicity. Finally, the majority of methods we discuss assume cross-sectional independence, i.e. that $\epsilon_{t,j}$ are independent over j. Throughout this subsection, we also assume cross-sectional independence unless otherwise specified.

Zhang et al., 2010 average over the square CUSUM test statistic,

$$\max_{1\leq t\leq n}\sum_{j=1}^{p}T^2(t,j).$$

J. Bai, 2010; Horváth and Hušková, 2012 average over a slight variant of the CUSUM test statistic which incorporates a normalization constant and different scaling,

$$\max_{1\leq t\leq n}\frac{1}{\sqrt{p}}\frac{t(n-t)}{n^2}\sum_{j=1}^{p}\left\{T^2(t,j)-1\right\}.$$

Furthermore, Jirak, 2012 consider the case of averaging with both cross-sectional and temporal dependence. Their method first uses an estimate of the long run covariance to whiten the CUSUM test statistics. They then take a pointwise average of the normalised test statistics. Note that estimating the long run covariance becomes difficult as p grows. Thus, this approach is only possible if p is small. J. Li et al., 2019 study a CUSUM type statistic with a bias term, which takes account of spatial and temporal dependence.

Taking an average is appropriate if each δ_j is small, but $\sum_{j=1}^{p}|\delta_j|$ grows quickly with p. For example, J. Bai, 2010; Horváth and Hušková, 2012 demonstrate that their method will consistently detect changepoints, if the sum of δ_j^2 diverges faster than $\sqrt{p}$. Furthermore, since this approach averages over a set of changepoint estimators, the resulting changepoint estimator should be more accurate. An important criticism of these approaches is raised by T. Wang and Samworth, 2018. An unweighted average is inefficient, if the δ_j values are not of similar size. T. Wang and Samworth, 2018 argue that a better approach would be to take a weighted average, with weights proportional

to the size of δ_j. However, this approach has not been analysed directly, as T. Wang and Samworth, 2018 also incorporate sparsity constraints in their model.

There are many applications where the assumption that every series under observation undergoes a change is unrealistic. Cho and Fryzlewicz, 2015 incorporate sparsity into a test statistic, by averaging over values of the CUSUM which exceed some threshold i.e.

$$\max_{1\leq t\leq n}\sum_{j=1}^{p}|T(t,j)|I\left(T(t,j)>\pi_n\right),$$

where π_n is a user specified threshold and I is an indicator function. Note, we say that a change has occurred if this value is greater than zero. Theoretical results demonstrate that this method can consistently estimate changepoints, even in the presence of temporal dependence. However we note that while this approach does utilise a sparsified test statistic, it does not determine whether variables are affected by a change and the method does not report the set of affected variables.

Many authors consider sparse alternatives directly by adding the following sparsity assumption to (2.3.1),

$$\sum_{j=1}^{p}I(|\delta_j|>0)=k,$$

if the number of non zero elements, k is known, and

$$\sum_{j=1}^{p}I(|\delta_j|>0)\leq k \text{ otherwise.}$$

Note that this sparsity constraint introduces a new component to the problem, distinguishing the series that change from those that do not. We can use the CUSUM value to distinguish these sets. The magnitude of the CUSUM indicates how likely it is that a change has occurred, and series that are affected by a change are more likely to have large CUSUM values. Thus, it should be possible to partition the series into two groups at each time point, based on the magnitude of the CUSUM values. The group with larger CUSUM values will have been affected by the change.

Enikeeva and Harchaoui, 2019 study the following test statistic,

$$\max_{1\leq t\leq n}\max_{1\leq l\leq p}\left(\log\frac{knp}{\alpha}\binom{p}{l}\right)^{-1}\sum_{j=1}^{l}\left\{T(t,\pi(t,j))^2-1\right\},$$

where k is a constant, α is a significance level and $\pi(j)$ denotes the label of the series with the jth largest CUSUM value at time t. This test statistic takes into account the multiple different possible combinations of series through the combinatoric term. Cho, 2016 utilise the following test statistic,

$$\max_{1\leq t\leq n}\max_{1\leq m\leq p}\left\{\frac{m(2n-m)}{2n}\right\}^{\xi}\frac{1}{m}\sum_{j=1}^{m}\left(|T(t,\pi(t,j))|-\frac{1}{2n-m}\sum_{j=m+1}^{n}|T(t,\pi(t,j))|\right)$$

where $\xi\in[0,1]$. This is equivalent to fitting an elbow plot to the ordered CUSUM values at each time point. These approaches are appropriate when we can easily separate the set of change sizes, $\{|\delta_j|\}_{j=1}^{p}$, into two groups and a large number of these δ_j values are zero. Note however that if the ordered sequence of $|\delta_j|$ values decays smoothly to zero, these methods will struggle even in the presence of true sparsity, as a clear seperation point will not exist. Furthermore, the accuracy of these methods depends heavily on the number of non zero δ_j values. As this number increases, the methods underperform compared to a simple average.

T. Wang and Samworth, 2018 consider a weighted average of CUSUM values, which does not suffer this limitation. Their approach can be broken into two steps. The weights are estimated as the leading sparse principal component, v of the CUSUM matrix T where $[T]_{i,j}=T(i,j)$. Then an estimator for the change is given by,

$$\max_{1\leq t\leq n}\left(\sum_{j=1}^{p}vT(t,j)\right)^{2}.$$

Note that the weight vector v is an estimator for the vector $\boldsymbol{\delta}:=(\delta_j)_{j=1}^{p}$. Thus this approach is equivalent to weighting by the size of the change in each series. As a result, this method is applicable in settings where previously discussed methods struggle, such as the case when some of the affected series experience much larger changes than the others. However, estimating v requires estimating an extra k parameters. Thus, if each affected series experiences a similar sized change or the change is not truly sparse, then this approach may add complexity without improving statistical efficiency. We note that in some applications there can be a mix of sparse and dense changes, and the Inspect procedure may perform worse than other methods on these problems. The Inspect method has been implemented in the R package InspectChangepoint (T.

Wang and Samworth, 2016). Due to the fact that it is a state of the art competitor method, in Chapters 3 and 4 we compare our contributions with the Inspect method.

There are certain settings where taking an average over test statistic values may be inappropriate. For example, in certain applications small changes which occur across multiple series may be unimportant, and the primary interest would be a large break in a single series. Jirak et al., 2015 take the maximum over the set of CUSUM statistics at each time point, which would be more appropriate than an average in these settings. Note that we would not expect the resulting changepoint location estimates to be more accurate than the equivalent estimates from applying a univariate CUSUM test to the series with the largest change. This is in contrast with averaging, where we would expect some improvement. The problem discussed in Jirak et al., 2015 can be described as detecting a statistically significant change. However there is another way of framing this. Dette and Gösmann, 2018 consider the problem of detecting relevant changes in mean, that is changes in mean which exceed some prespecified level, $\Delta\mu$. The advantage of this approach is that it allows the practitioner to define a significant change, which is useful if small changes are not important. The authors study the maximum of the CUSUM statistic at each time point, after applying a correction for $\Delta\mu$. There are clearly deep links between these two approaches. In particular, one can map the desired minimum size of change, $\Delta\mu$, to a significance level and vice versa.

Incorporating dependence between series in a CUSUM style statistic is difficult as, the distribution of the resulting test statistic will depend on dependence structure. As a result, the practitioner has to estimate this structure to use these methods, which is challenging in the presence of changepoints. R. Wang et al., 2019 study the following U statistic based process instead,

$$D(\tau; l, k) := \sum_{l \leq j_1 \neq j_3 \leq \tau} \sum_{\tau < j_2 \neq j_4 \leq k} (\boldsymbol{X}_{j_1} - \boldsymbol{X}_{j_3})^T (\boldsymbol{X}_{j_2} - \boldsymbol{X}_{j_4}).$$

The advantage of this approach is that, the test statistic can be normalised without estimating the dependency structure, by dividing by linear combinations of $D^2(\tau; \cdot, \cdot)$ calculated on different subsets of the data. This approach is valuable if there is some

cross-sectional dependence in the data. However, we note that in their simulation studies, the method only marginally outperforms T. Wang and Samworth, 2018 in a setting where it is favoured. Therefore, it is questionable whether this approach is effective in practice.

2.3.2 Changes in Covariance Structure

We now move on to the problem of detecting changes in the covariance structure of multivariate time series. The literature on this problem has grown substantially in recent years. An important distinction between the different approaches is how they incorporate the structure of the underlying covariance matrix. We begin our discussion by considering models which do not assume any structure. Formally, we study the mean zero vectors $\boldsymbol{X}_t$ such that,

$$\boldsymbol{\Sigma}_1^* = \mathbb{E}(\boldsymbol{X}_1\boldsymbol{X}_1^T) = \cdots = \mathbb{E}(\boldsymbol{X}_\tau\boldsymbol{X}_\tau^T) \neq \mathbb{E}(\boldsymbol{X}_{\tau+1}\boldsymbol{X}_{\tau+1}^T) = \cdots = \mathbb{E}(\boldsymbol{X}_n\boldsymbol{X}_n^T) = \boldsymbol{\Sigma}_2^*$$
$$\text{where } \|\Sigma_2^* - \Sigma_1^*\| = \delta > 0$$

where as before, $\tau \in \mathbb{Z}$ is the location of the changepoint. As in our previous discussion, we assume that the vectors $\{\boldsymbol{X}_t\}_{t=1}^n$ are I.I.D, unless otherwise stated and restrict our attention to the single changepoint setting.

Given the amount of work focused on the problem of detecting changes in mean, a natural approach to this problem is to look for changes in the mean of the vectorized matrix $X_iX_i^T$. In particular, Cho and Fryzlewicz, 2015; R. Wang et al., 2019 feature results, which show that their methods for changes in mean can also detect changes in second order structure. This approach does not exploit the relationships between the entries of $X_iX_i^T$, and thus may lose power in settings where these relationships are stronger than the change.

A number of authors examine the problem of detecting changes in covariance directly via the Covariance CUSUM,

$$T(t) := \alpha_{t,n}\left(\frac{1}{n-t}\sum_{r=t+1}^{n}\boldsymbol{X}_r\boldsymbol{X}_r^T - \frac{1}{t}\sum_{r=1}^{t}\boldsymbol{X}_r\boldsymbol{X}_r^T\right) \tag{2.3.2}$$

where $\alpha_{t,n}$ is an appropriate weight function. Aue, Hörmann, et al., 2009 study the following statistic,

$$\max_{1\leq t\leq n} vech(T(t))^T \hat{\Sigma}^{-\frac{1}{2}} vech(T(t))$$

where $vech(X)$ is the $p(p+1)/2$ dimensional vectorization of the matrix X and $\hat{\Sigma}$ is a plug in estimator for the long run covariance of $vech(T(t))$. The $\hat{\Sigma}$ term accounts for cross sectional and temporal dependence between the entries of $X_i X_i^T$. Note estimating $\hat{\Sigma}$ can become very difficult as p increases. D. Wang, Yu, and Rinaldo, 2017 consider the following statistic,

$$\max_{p\log n<t\leq n-p\log n} \|T(t)\|_{op}$$

where $\|X\|_{op}$ is the largest principal component of X. This approach is shown to be minimax optimal if the vectors are independent and sub-Gaussian. However, the method requires bounds on the variance of each $\boldsymbol{X}_i$. If these are not known a priori then an upper bound must be estimated from the data, which requires knowledge of the unknown covariance for each segment. This approach can be understood as a projection method, where the data is projected along the first principal component of $T(t)$. These methods are state of the art competitors for the method we develop in Chapter 6, and we compare our proposed approach with these methods via a simulation study where the methods were implemented in the R programming language.

Dette, Pan, et al., 2018 study the problem of detecting a change in the covariance of very large covariance matrices. They study a similar test statistic to Aue, Hörmann, et al., 2009, however they incorporate a sparsification step similar to that used in Cho and Fryzlewicz, 2015. If the threhold is sufficiently large, then the method can be used to study very high dimensional time series. Steland, 2020 study bilinear forms of the covariance cusum, i.e. quadratic forms, $v^T T(t) w$ where v, w are non random vectors chosen by the user. This approach is useful in settings where, there is some a priori knowledge of the structure of the covariance, such as a block structure. Note that this approach allows for temporal dependence. Avanesov and Buzun, 2018 study changes in the inverse covariance (precision) of high dimensional time series. Their test statistic measures differences between debiased estimates of the inverse covari-

ance matrix. The authors provide a bootstrap procedure for selecting the threshold, however this threhold requires knowledge of the underlying initial covariance matrix. Note that while a change in the inverse covariance matrix is equivalent to a change in the covariance matrix, there are many applications where the primary parameter of interest is the precision matrix and assumptions can be made about the matrix. In such settings, a method that focuses on changes in the precision matrix may be preferable. We have implemented a version of this method in the R programming language and this approach is also included in the simulation study in Chapter 6.

The methods described in this section all require an estimate of an initial covariance or autocovariance in order to calculate the test statistic or set an appropriate threshold for detecting changepoints. However to accurately estimate the initial covariance we need to know where the changepoints are. As a result, it can be difficult to correctly specify the appropriate threshold for changes in covariance. Furthermore if the initial covariance is estimated from a heterogenous sample, the power of the method may suffer. In Chapter 6 we propose a new method for detecting changes in covariance structure which does not require any knownledge of the underlying covariance structure.

2.3.3 Changes in Functional Data

There has been significant interest within the literature in detecting changes in functional data. In functional data analysis, each vector $\boldsymbol{X}_i$ is assumed to be a discrete realisation of a continuous function. Formally, we have that

$$\begin{aligned} &\boldsymbol{X}_t = \boldsymbol{\mu}_t + \boldsymbol{\epsilon}_t \text{ where} \\ &\mu_t = && \mu_k^* \in \mathcal{L}^2(\mathcal{I}),\ \epsilon_t \sim F_k \\ &\text{for } \tau_{k-1} + 1 \leq t \leq \tau_k,\ 1 \leq k \leq m \end{aligned}$$

$\mathcal{I}$ is some compact set, and F_j is distribution over $\mathcal{L}^2(\mathcal{I})$. In other words, μ_t and ϵ_t are square integrable real valued functions from $\mathcal{I}$ to the reals. Each distribution F_k has covariance function

$$K_k(r, s) = \mathbb{E}(\epsilon_{\tau_k}(r)\epsilon_{\tau_k}(s)), r, s \in \mathcal{I}.$$

If $\mu_k^* \neq \mu_{k+1}*$ then we say there is a change in mean, while if $K_k(r,s) \neq K_{k+1}(r,s)$ we say there is a change in covariance.

It is common to assume that the covariance function K_k has some low dimensional representation. This low dimensional representation can then be estimated using functional principal component analysis (Ramsay, 2004). Aue, Gabrys, et al., 2009; Berkes et al., 2009 estimate changes in mean of indepedent functional observations, by projecting a CUSUM style statistic onto sample functional principal components. Aston and Kirch, 2012a extend this approach to the setting where there is temporal dependence. Aston and Kirch, 2012b demonstrate that this method can be used to identify non-stationary fMRI data. The dimension reduction approach is appropriate if the data can be accurately described by a low dimensional representation. However there is also interest in so called fully functional data which does not admit such a representation and thus dimension reduction techniques perform poorly. Aue, Rice, et al., 2018 propose a CUSUM style estimator for changes in mean in this setting, which does not use dimension reduction. Note the question of whether or not to use a dimension reduction technique here depends on the data. If the change occurs in the direction of the primary principal components (or we are only interested in changes in these directions) then a method which uses a dimension reduction technique is preferable. However if the change occurs in the direction of the subspace orthogonal to the principal components, then the use of dimension reduction techniques may make the change harder to find.

There has been growing interest in detecting changes in the covariance function $K(r,s)$. Jarušková, 2013 propose a two sample test for detecting a difference in covariance operator which they extend to the changepoint setting. Dimension reduction techniques are also used for this problem. Stoehr et al., 2020 first perform dimension reduction and then utilise the estimator proposed by Aue, Hörmann, et al., 2009 to detect changes in covariance. Dette and Kutta, 2019 propose a self-normalised two sample test statistic to detect differences in the eigensystem of the covariance function K. Finally, Aue, Rice, et al., 2020 study changes in the spectrum function and trace of the covariance function.

2.3.4 Nonparametric Changepoints

So far we have focused on methods which attempt to identify changes in the moments of a distribution, where the distribution is either known or satisfies some strong assumptions. However this ignores problems where the distribution of the data changes or, the data does not satisfy the required assumptions. Thus there is a growing interest in nonparametric methods, which do not place assumptions on the type of change and minimal assumptions on the data. Formally, we study the original changepoint model in (2.1.1) without extra assumptions which we repeat here for convenience,

$$\begin{aligned} \boldsymbol{X}_t \sim F_k \text{ for } \tau_{k-1} < t \leq \tau_k \\ F_k \neq F_{k+1} \text{ for } 1 \leq k \leq m \\ 0 = \tau_0 < \tau_1 < \cdots < \tau_m < \tau_{m+1} = n \end{aligned} \tag{2.3.3}$$

Matteson and James, 2014 study a test statistic based on energy distances which measure the distance or divergence between random variables. In particular given two samples $\boldsymbol{X}$ and $\boldsymbol{Y}$ of length n and m respectively, the authors demonstrate that the following two sample test statistic,

$$\zeta(\boldsymbol{X}, \boldsymbol{Y}, \alpha) = \frac{2}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} |\boldsymbol{X}_i - \boldsymbol{Y}_j| - \binom{n}{2}^{-\frac{1}{2}} \sum_{1 \leq i \leq k \leq n} |\boldsymbol{X}_i - \boldsymbol{X}_k| - \binom{m}{2}^{-\frac{1}{2}} \sum_{1 \leq j \leq l \leq n} |\boldsymbol{Y}_j - \boldsymbol{Y}_l|$$

can be used to consistently detect changes in the distribution of data. The authors develop a bootstrap procedure for selecting the significance threshold which can be applied in the multiple changepoint setting. This method, E-Divisive, is implemented in the ECP R package (James and Matteson, 2015) and has computational cost $\mathcal{O}(n^2)$ (as it requires computing every possible pairwise distance). We compare our contributions with this method in Chapters 3 and 4. Finally we note that although the energy statistic ζ is based on euclidean distances, in theory it could be extended to consider other metrics and with the proper choice can be applied to a wide range of data types such as functional data and compositional data.

A number of authors have have studied Kernel based changepoint estimators. Inspired by clustering methods, these methods utilise a kernel transform to map a

change in distribution to a change in expectation in the kernel space. One advantage of the kernel approach is that we can analyse any type of data, so long as a suitable kernel exists. S. Li et al., 2015 study kernel M-statistics for which the tail probability, and thus the threshold for determining a change, can be fully characterized. Arlot et al., 2019; Garreau, Arlot, et al., 2018 study the statistical properties of kernel based changepoint estimators. They prove that such estimators are consistent if the kernel of the data satisfies some assumptions. Importantly, these assumptions are not placed on the data itself, thus we can detect changes in any type of data, so long as a suitable kernel function exists. Finally, J. Li, 2020 study the properties of a kernel based estimator in high dimensions, although they describe their procedure as being based on interpoint distances. Note, although their approach is applicable in high dimensions, they place stricter assumptions on the kernel matrix. Finally, we note that the energy statistic approach described above can also be expressed as a kernel method.

Kernel changepoint estimators have two important limitations. Firstly, the practitioner must choose an appropriate kernel family and set the hyperparameters correctly. Different kernels will favour certain types of changes over others and, the power of the method may decrease substantially if the hyperparameters are not suitable. Thus the choice of kernel may have significant impact on results. Secondly, the computational cost of estimating the kernel is $\mathcal{O}(n^2)$, which may be prohibitively large for longer datasets. Celisse et al., 2018; Truong et al., 2019 introduce computationally efficient implementations of these test statistics, for large datasets with multiple changepoints.

Lung-Yut-Fong et al., 2015 study a rank based statistic for estimating changepoints. In particular, at each time point they measure the distance between segments using a Wilcoxon-Mann-Whitney type test for each variate independently. The authors then aggregate this information using a normalised sum of squares, where the normalisation term is an estimator of the covariance. Note there is a clear connection here with the aggregation techniques discussed for the change in mean problem. This procedure is consistent, assuming some conditions on the gradient of the true distribution. Note, the effectiveness of this procedure depends on accurate estimation

of the nuisance covariance parameter, which may be difficult in the multiple changepoint setting. Brault et al., 2018 extend this approach, to study changes in the block structure of high dimensional symmetric matrices.

There is growing interest in graph based test statistics. H. Chen and Zhang, 2015 detect changes by first constructing a graph from the data. Then at each time point, they count the number of edges between data to the left and right of the candidate change. If this number is large then the two segments are more likely to have the same distribution and vice versa. Chu and H. Chen, 2019 argue that this approach can be ineffective for certain types of changes, and introduce new statistics that are more appropriate for these settings. Similar to kernel based methods, the assumptions required for a graph based test statistic to be consistent, depends on how the graph is constructed rather than the statistical properties of the data. However, the power of the method also depends on how the graph is constructed and whether this construction illuminates the type of change.

Dubey and Müller, 2019 study changes in the Fréchet mean and variance of data observed in some metric space. Fréchet mean and variance generalise the concepts of location and scale to objects in a metric space. The proposed estimator selects changepoints, by maximising the distance in sample Fréchet mean and variance estimates. Although this method assumes knowledge of the type of change, it can be considered nonparametric as the data is observed in a generic metric space. The method is consistent with assumptions on the metric space. However, the performance of the method depends on the choice of metric. Padilla, Yu, D. Wang, et al., 2019 propose a CUSUM type estimator for estimating changes in distribution of real valued data. Their method measures the distance between probability distribution functions, of data to the left and right of each candidate change using Kernel Density Estimators. This approach is consistent assuming the true density functions are uniformly Lipschitz.

2.3.5 Changes in Vector Autoregressive Models

There is a growing literature focused on changes in vector autoregressive (VAR) models. VAR models are widely used in multivariate time series analysis, with applications in biology (Fujita et al., 2007) and finance (Fan et al., 2011). Formally, we consider the following model,

$$\boldsymbol{Y_t} = \sum_{l=1}^{q_k} \boldsymbol{A}_l^k \boldsymbol{Y}_{t-l} + \boldsymbol{\epsilon}_t \text{ for } \tau_k < t \leq \tau_{k+1}$$

$$\{A_l^k\}_{l=1}^{q_k} \neq \{A_l^{k+1}\}_{l=1}^{q_{k+1}} \text{ for } 1 \leq k \leq m$$

where $q_t \in \mathbb{R}$ is the order of the model, $\boldsymbol{A}_l^k$ is a $p \times p$ matrix and the error terms $\boldsymbol{\epsilon}_t$ are IID normal with covariance Σ_t. Depending on the work, the covariance and order terms may be assumed to be stationary or piecewise stationary. Furthermore, there is some disagreement in the literature, about whether data immediately after a change should be affected by data before the change or not. If the order terms q_t are small, this issue is unlikely to substantially alter the analysis. However for series with large order terms, this problem may need to be addressed directly.

It is possible to detect a change in the VAR parameters by examining changes in the covariance of $\boldsymbol{Y}_t$ or changes in expectation of the parameters $\boldsymbol{A}_t$ and as such, previously discussed methods may also be appropriate for this problem. However, there are also methods that tackle this problem directly. Davis, Lee, et al., 2006 propose a consistent changepoint estimator, based on the principle of Minimum Description Length (MDL). Their estimator allows for changes in the covariance and order terms. The proposed estimator is the solution to a computationally intractable optimisation problem. Therefore the authors utilise a genetic algorithm to optimise the function and estimate changepoints. Kirch et al., 2015 estimate the changepoint in two steps. Firstly they jointly estimate the autoregressive parameters and changepoint locations by solving a regularized regression problem for the entire dataset. This regularized regression problem incorporates two penalties, one for controlling the number of changepoints and another for controlling the sparsity of the VAR model. Due to the fact that this estimator consistently overestimates the number of changes, they

then study a CUSUM style statistic of the fitted residuals. The method assumes that the order and covariance terms are constant across segments. Note this second stage introduces more hyperparameters. This approach is computationally efficient if the hyperparameters are fixed and is suitable for high dimensional VAR problems. However in practice the hyperparameters must be tuned to a given dataset which is a costly and non trivial exercise, particularly since there are two sets of interconnected hyperparameters which must be trained.

The number of parameters required in a VAR model is quadratic in p. A number of authors have addressed this problem by inducing sparsity in the model with a LASSO type penalty (Davis, Zang, et al., 2016; Nicholson et al., 2017). Recently, some authors have applied this approach to the change in autoregression problem. Safikhani and Shojaie, 2020 utilise a two step procedure. They first jointly estimate the parameters of the model and the changepoints by optimising a penalised cost function. This procedure consistently overestimates the true number of changepoints. Therefore, the second step reduces the set of estimated changes by choosing the subset which optimises an information criterion. D. Wang, Yu, Rinaldo, and Willett, 2019 estimate changepoint locations by minimising the penalised cost function (2.2.3), where $\mathcal{C}$ is a likelihood function which uses a LASSO estimator for the autoregressive terms. Note that unlike the previous approach, this method directly penalises the number of changepoints and does not assume any prior beliefs about the sparsity of the changes. However this approach does have a number of limitations. Firstly whereas the previous approach solves a single (large) convex optimisation problem, this method must solve $\mathcal{O}(n)$ convex optimisation subproblems on average and $\mathcal{O}(n^2)$ in the worst case. This may be a computationally intensive process. Secondly the authors do not include any simulation results and thus it is unclear whether this approach works in practice. Finally, as with all regularization methods, the penalty term must be tuned for the problem which is a computationally intensive process.

2.3.6 Changes in Network Structures

A number of authors have studied the problem of detecting changes in a sequence of networks. Barnett and Onnela, 2016 detect changes in correlation networks by maximising the ℓ_2 distance between sample covariances. This procedure uses a bootstrap procedure to test for a significant change. The authors extend the method to the multiple changepoint case via the binary segmentation procedure. This extension is unsatisfactory as in the multiple changepoint setting, the bootstrap samples will have different distributions leading to different thresholds for significance. Furthermore the authors do not address the question of sparsity when studying the changepoints, which is important in this case as correlation networks are typically sparse.

D. Wang, Yu, and Rinaldo, 2018 utilise a CUSUM style statistic to identify changes in independent Bernoulli networks. The authors argue these networks structures are very general including the stochastic block model and random dot product models as special cases. While this is true, it ignores the fact that these models require stronger assumptions about the data to accurately represent the data and, without them the the variance of the estimates of the network structure will overwhelm the signal. As such while the approach is very general, it is unlikely to be useful in practice. Padilla, Yu, and Priebe, 2019 propose a two step estimator, for detecting changes in a sequence of independent random dot product graphs. The authors first estimate the latent coordinates of each graph and then use a nonparametric CUSUM style test to detect a change. By utilising more realistic assumptions, this approach should be applicable to a greater range of real datasets than the previous general approach. However we do note that the authors assume that the dimension of the latent space is fixed and known which is unlikely to be true.

Cribben, Haraldsdottir, et al., 2012; Cribben, Wager, et al., 2013 study changes in functional connectivity networks of fMRI data. The authors detect changepoints by minimising a BIC type penalty which uses a multivariate normal log likelihood with a Graphical LASSO based estimator for the precision matrix. Then conditional on the changepoints the segment networks can be estimated. The authors detect multiple changepoints via the binary segmentation procedure and use a bootstrap procedure

to test for significance. While this approach has merit, it suffers from some important limitations. Each segment has a mean and covariance parameter which implies that changes in the mean of the time series or scale of the covariance will be reported as changepoints. In other words, the model can report a change where the structure of the network is constant. Furthermore the regularization parameter requires tuning. Londschien et al., 2019 consider a similar approach for detecting changes in graphical models with missing data. In particular, they consider a penalised cost function approach with a cost function based on the LASSO penalised likelihood function. The authors consider a number of data imputation strategies for estimating the covariance of the full data, which can then be used to calculate the penalised likelihood.

Many network models assume that the graph can be represented as a point in a Euclidean space, however there is significant interest in non-Euclidean based representations (Bronstein et al., 2017). Grattarola et al., 2019 study changes in networks by first embedding each graph on a constant curvature Riemannian manifold via an adversarial autoencoder. The authors then test the resulting sequence for a single change. Note this procedure requires a large sample of homogenous data on which to train the autoencoder, which is not typically available.

Gibberd and Nelson, 2014, 2017 study changes in the dependency structure of Gaussian Graphical Models via group LASSO penalties. Both methods jointly estimate a sequence of n inverse covariance (precision) matrices $\{\hat{\Theta}_t\}$ by optimizing a penalised cost function. The penalised cost function incorporates two types of penalty, a shrinkage term which penalises non zero entries in each precision matrix and a smoothness term which penalises non zero differences between the same entry in consecutive precision matrices i.e. $|\Theta_{i,j}^{t+1} - \Theta_{i,j}^{t}|$. This penalty structure produces a sequence of sparse precision matrices in the sense that many entries $\Theta_{i,j}^{t} = 0$. Furthermore the sequence $\{\Theta_{i,j}^{t}\}_{t=1}^{n}$ exhibits a piecewise constant structure. Note however that this approach does not penalise the number of changepoints and in theory there can be a change at each time point. These methods are particularly valuable in high dimensional settings where there is true sparsity in the covariance structure and a small number of entries change at each time point. However the method harshly pe-

nalises changepoints where the majority of entries change. Furthermore the penalty terms need to be tuned for the method to work, which can be difficult in practice.

2.3.7 Changes in Other Data Structures

So far in our discussion, we have considered changes in the mean, changes in covariance, changes in Vector Autoregressive models, changes in functional data and nonparametric changepoints. We now discuss some interesting works in the literature that do not fit neatly into these categories. As we have already seen, hypothesis tests based on likelihood functions are widely used to detect changepoints in a wide range of models. However there are a range of applications where the full likelihood is computationally intractable, due to a high dimensional integral term. In such situations, it can be useful to work instead with the composite likelihood function, which combines likelihoods calculated on subsets of the data. Ma and Yau, 2016 combine the penalised cost function approach with a cost function based on the pairwise likelihood, an example of a composite likelihood. Although this approach is necessarily less efficient than a full likelihood approach, the authors demonstrate that it can be used to consistently detect changepoints and outperforms nonparametric methods when correctly specified. Zhao et al., 2019 adapt this approach to detect changes in spatio-temporal processes. Prabuchandran et al., 2019 consider the problem of detecting changes in compositional data where each element is a probability mass function. They propose a penalised cost function approach with a cost function based on the parametric Dirichlet likelihood function. We note the authors only consider the single changepoint setting in this work, however it should be possible to extend this approach to the multiple changepoint setting via a dynamic program or binary segmentation.

2.4 Multivariate Changepoint Detection via Optimisation

In the previous section, we saw that there is significant interest in methods that can detect so called sparse changepoints, changes where only a subset of the variables under consideration are affected by the change. However in Section 2.2.2 we saw that the primary search methods for detecting multiple changepoints, namely the various binary segmentation procedures and the penalised cost function approach, do not allow for such sparse changes and assume that every variable is affected by a change. Pickering, 2016 consider the problem of simultaneously estimating multiple sparse changepoint locations and the set of variables affected by each change. In particular, the author proposes to estimate changepoints by solving an optimisation problem with a dual penalty cost function.

Let m^j be the number of changepoints and $\boldsymbol{\tau}^j := \{\tau_0^j, \tau_1^j, \ldots, \tau_{m^j}^j, \tau_{m^j+1}^j\}$ be the set of changepoints that affect variable j, where $\tau_0^j = 0$ and $\tau_{m^j+1} = n$. Furthermore, let $\boldsymbol{m} := \{m^1, \ldots, m^p\}$ and $\mathcal{T} := \{\boldsymbol{\tau}^j\}_{j=1,\ldots,p}$. Then the optimal subset multivariate segmentation for the dataset, $\boldsymbol{X} := \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$, is given by the solution to the following optimisation problem,

$$\min_{\boldsymbol{m},\mathcal{T}} \sum_{j=1}^{p} \sum_{k=1}^{m^j+1} \left(\mathcal{C}^j \left(X^j_{(\tau_k^j+1):(\tau_{k+1}^j)} \right) + \alpha \right) + \beta\psi(\mathcal{T}) \tag{2.4.1}$$

where p is the length of the vectors $\boldsymbol{X}_i$, $\mathcal{C}^j$ is a cost function measuring goodness of fit for variable j, β penalises the number of changepoints, α penalizes each series affected by the change, and ψ is a function which counts the number of unique elements in a set. Note under this framework changepoints are shared across multiple variables via the β penalty, however not every variable is affected by each change due to the addition of the α penalty. Furthermore the multivariate penalised cost function approach can be thought of as a special case of this model (for example by setting $\beta = 0$). Note, as it is foundational to the ideas developed in Chapters 3 and 4, an equivalent definition of (2.4.1) (and necessary associated terms) is repeated in Sections 3.2 and 4.2.1.

We can identify sparse changepoints by selecting the model that minimises the

dual penalty cost function. This optimisation problem can be solved via a dynamic program introduced by Pickering, 2016, Subset Multivariate Optimal Partitioning (SMOP). The reported computational cost of this procedure is $\mathcal{O}(K(n)pn^{2p})$ which is prohibitively expensive for even small datasets. In Chapter 3 we propose a number of techniques which substantially reduce the computational cost of this procedure. Subsequently, in Chapter 4 we introduce a computationally efficient approximate method which can be applied to very large datasets. Although this procedure is not exact, we show that it always performs at least as well as the single penalty framework with a similar computational cost.

Chapter 3

Exact Subset Multivariate Changepoints

As we have discussed in the literature review, there is increasing interest in and demand for methods that can detect changepoints in multivariate datasets. Consider the copy number variation dataset included in the ecp R package (James and Matteson, 2015). This dataset contains information related to 43 individuals with bladder tumours. Changes which occur across multiple individuals, may be linked to the presence of tumours and are of significant scientific interest. When looking to detect these changes, it is important to combine information across series or risk missing important changes due to lack of power. Matteson and James, 2014; T. Wang and Samworth, 2018 use multivariate methods to detect changepoints in this setting. In Figure 3.0.1, we can see the resulting segmentations for the first individual using a univariate and multivariate method, where the multivariate method assumes every variable is affected by the change. Looking at the results, it appears as if the multivariate approach overfits changepoints, particularly when compared with the univariate approach. However this is not actually the case. The multivariate methods are detecting true changes in the whole data, however the first individual is not affected by these changes. To accurately represent the data shown in Figure 3.0.1, we need a method that can detect multivariate changepoints and identify whether or not each variable changes at that changepoint. For this example in particular, we would like to identify the changes

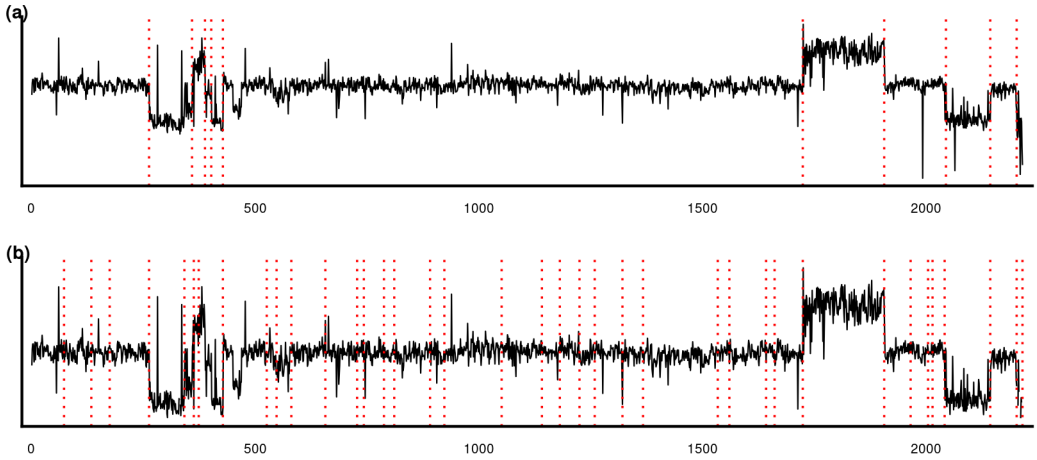

Figure 3.0.1: (a) Segmentation using a univariate method for the first individual. (b) Segmentation using a multivariate method for the same individual. Since the first individual is not affected by many of the changes the multivariate segmentation appears to overfit the data.

that affect the majority of individuals as these are more likely related to the disease in question. This leads us to the idea of subset multivariate changepoints, changepoints in multivariate data where only a subset of the channels are affected by a change.

In this chapter, we consider the problem of detecting subset multivariate changepoints via the penalised cost function framework introduced by Pickering, 2016. This framework formulates the problem of locating subset multivariate changepoints as a discrete optimisation problem, which can be solved using a dynamic program. However the computational cost of this dynamic program is extremely high and scales poorly. As a result, the proposed approach is infeasible for even small datasets. Therefore we propose a new preprocessing algorithm which significantly reduces the cost of optimising this penalised cost function. We propose a simple set of rules which identify sets of suboptimal solutions within the search space for the dynamic program. We then use a dynamic program to identify the optimal solution within the reduced search space significantly reducing the computational cost. This chapter is structured as follows. In Section 3.1, we review an important method for detecting changepoints, which has inspired the work that follows. This approach has significant limitations

in the multivariate setting, namely that it assumes every variable is affected by the change. Therefore in Section 3.2, we discuss a related method, Subset Multivariate Optimal Partitioning (SMOP), which does not make this assumption. The computational cost of SMOP is prohibitive, even for small datasets. Therefore in Section 3.3, we introduce novel techniques for reducing the computational cost and propose a new more efficient algorithm. In Section 3.4, we analyse the performance of our method on a range of simulated datasets and demonstrate that our proposed approach can detect subset multivariate changepoints at a significantly reduced computational cost. In Section 3.5, we use the new method to identify changes in growth rates of confirmed Covid-19 cases in Great Britain. Finally in Section 3.6, we review the contributions we have made in this chapter and discuss some remaining limitations.

3.1 Single Penalty Framework

We now review one of the most popular methods in changepoint analysis, the single penalty cost function. There are two reasons why this review will be useful. Firstly it allows us to address some limitations of this approach, which are relevant in the multivariate setting. Secondly, throughout this chapter we draw inspiration from ideas from the single penalty cost function literature, and as such understanding these ideas in the single penalty setting provides intuition for later discussions.

Throughout this section we will consider data $X = \{X_l\}_{l=1,\ldots,n}$, where each X_l can be scalar or vector valued. We also introduce a cost function $\mathcal{C}(\{X_l\}_{l=s+1,\ldots,t})$ which measures goodness of fit. For simplicity of notation we define

$$\mathcal{C}(s,t) := \mathcal{C}(\{X_l\}_{l=s+1,\ldots t}). \tag{3.1.1}$$

A typical choice for the cost function $\mathcal{C}$ is twice the negative log likelihood of an appropriate model for the data, X. A standard approach to segmenting the data, X, is to solve the following optimisation problem,

$$\min_{\boldsymbol{\tau},m} \sum_{k=1}^{m+1} \mathcal{C}(\tau_{k-1},\tau_k) + m\beta, \tag{3.1.2}$$

where β is a penalization term set by the user to prevent overfitting of changepoints, $\boldsymbol{\tau} := \{\tau_0, \tau_1, \ldots, \tau_m, \tau_{m+1}\}$ is the vector of changepoint locations, $\tau_0 = 0$, $\tau_{m+1} = n$ and m is the number of changepoints. In other words the problem of detecting changes in the dataset, X, can be formulated as a discrete optimisation problem.

The optimisation of (3.1.2) has been addressed by a number of authors and the optimal solution can be determined via a dynamic program (Jackson et al., 2005). The key idea is to construct a recursion by conditioning on the location of the last changepoint prior to time n. Let $F(t)$ denote the cost of the optimal segmentation of the data $\{X_l\}_{l=1,\ldots,t}$. If we knew that t was the optimal last changepoint prior to the time point T, then we could calculate $F(T)$ as follows,

$$F(T) = F(t) + \mathcal{C}(t, T) + \beta.$$

We do not know which time point t is the optimal prior changepoint so we search over all prior values, $\Lambda_T := \{t \in \mathbb{Z} : 0 \leq t < T\}$. Then we can calculate $F(T)$ by solving the following recursion,

$$F(T) = \min_{t \in \Lambda_T} \left\{ \underbrace{F(t)}_{\text{Optimal Cost up to time } t} + \underbrace{\mathcal{C}(t, T)}_{\text{Cost of data after } t} + \underbrace{\beta}_{\text{Cost of extra change at time } t} \right\}. \tag{3.1.3}$$

In other words, identifying the optimal segmentation of X up to time n is equivalent to identifying the most recent change prior to n. The cost of calculating $F(n)$ if we know $F(t)$ for all $t < n$ is thus an order n calculation. We can identify the optimal segmentation by calculating $F(t)$ for all $t \in \Lambda_n$ in order. The computational complexity of this calculation is $\mathcal{O}(K(n)n^2)$, where $K(n)$ is the computational complexity of evaluating $\mathcal{C}$. Note that a number of commonly used cost functions (e.g. likelihood for a change in mean) can be evaluated using summary statistics and as a result have computational complexity $\mathcal{O}(1)$.

There has been significant work examining how the computational cost of this dynamic program can be reduced. This is achieved by reducing the number of candidate last changepoints that need to be considered i.e. reducing the size of the set Λ_t for each $t \leq n$. In particular, Killick, Fearnhead, et al., 2012 and Maidstone et al., 2017

introduce simple time dependent conditions on each time point t, which if satisfied at some time $s > t$, imply that t can not be the optimal prior changepoint for any $T \geq s$, and as a result t can be removed from each set Λ_T reducing the computational cost of the dynamic program. These simple conditions can, under certain conditions, reduce the computational cost of a dynamic program from quadratic in the length of the data to linear. In Section 3.3, we examine how these ideas can be extended to the dual penalty framework.

The single penalty cost function provides accurate segmentations for univariate data, with little computational cost. However in the multivariate setting it has a significant limitation; it assumes every variable is affected by the changepoint. This produces segmentations such as the one illustrated in Figure 3.1.1 (a). The single penalty cost function partitions each variable at time $t = 125$. However we can see that variable 2 is unaffected by the change. Allowing every variable to be affected by each change, whether 1 or all variables are affected, will automatically produce a better model fit. Therefore in order to detect subset multivariate changes, we must also penalise the number of variables affected by each change. In the next section, we discuss the dual penalty penalised cost function introduced by Pickering, 2016 for detecting subset multivariate changepoints and how this function can be optimised via a dynamic program.

3.2 Dual Penalty Framework

Throughout this section we will consider data $\boldsymbol{X} = \{\boldsymbol{X}_l\}_{l=1,\ldots,n}$, where each $\boldsymbol{X}_l$ is vector of length p. For each variable $1 \leq j \leq p$, we have a corresponding cost function $\mathcal{C}^j(\{X_l^j\}_{l=s,\ldots,t})$ which measures the goodness of fit in variable j. Note that these cost functions can differ across variables. Again for simplicity of notation we define

$$\mathcal{C}^j(s,t) := \mathcal{C}^j(\{X_l^j\}_{l=s+1,\ldots t}).$$

Pickering, 2016 propose to detect subset multivariate changepoints by solving a generalization of the discrete optimisation problem in (3.1.2). Let m^j be the number

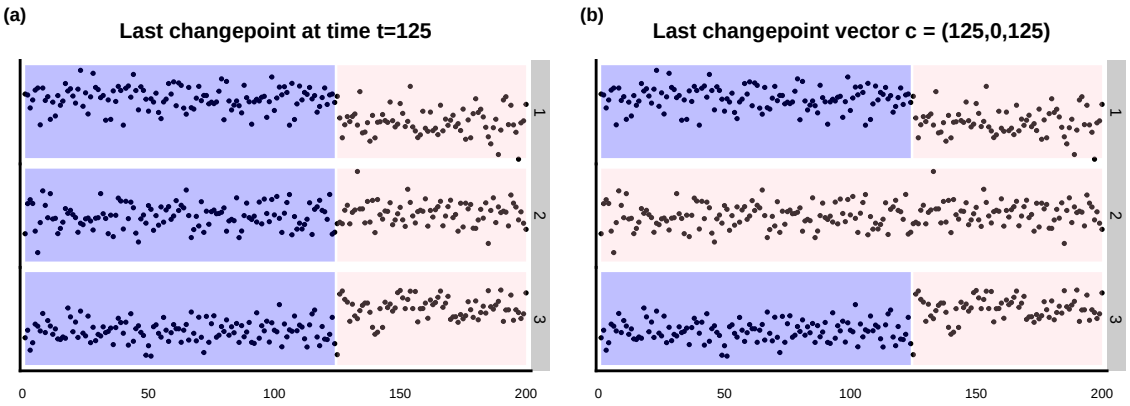

Figure 3.1.1: (a) Multivariate segmentation using the single penalty cost function. Note that the change must affect every variable. (b) Multivariate segmentation using the dual penalty cost function.

of changepoints and $\boldsymbol{\tau}^j := \{\tau_0^j, \tau_1^j, \ldots, \tau_{m^j}^j, \tau_{m^j+1}^j\}$ be the set of changepoints that affect variable j, where $\tau_0^j = 0$ and $\tau_{m^j+1} = n$. Furthermore, let $\boldsymbol{m} := \{m^1, \ldots, m^p\}$ and $\mathcal{T} := \{\boldsymbol{\tau}^j\}_{j=1,\ldots,p}$. Then the optimal subset multivariate segmentation for the dataset, $\boldsymbol{X}$, is given by the solution to the following optimisation problem,

$$\min_{\boldsymbol{m},\mathcal{T}} \sum_{j=1}^{p} \sum_{k=1}^{m^j+1} \left(\mathcal{C}^j\left(\tau_k^j, \tau_{k+1}^j\right) + \alpha\right) + \beta\psi(\mathcal{T}) \tag{3.2.1}$$

where β penalises the number of changepoints, α penalizes each series affected by the change, and ψ is a function which counts the number of unique elements in a set. This cost function penalises the number of changepoints through the β parameter and the number of variables affected by a change through α.

In this chapter, we focus on solving the optimisation problem and do not consider appropriate values for the hyperparameters α and β, however we do note a number of interesting features of the penalty structure. Firstly the cost of having an additional variable be affected by a change is indepedent of the total number of variables affected by the change. This allows us to recover the set of affected variables for both dense and sparse changes. However in some settings this penalty structure may be suboptimal. If all the changes are dense and p is large, it may be better to have the cost of each additional affected variable decrease as the total number of affected variables increase,

so that the cost of adding an 11th variable is larger than the cost of adding a 12th variable.

Secondly this framework includes a number of interesting special cases. If $\beta = 0$, there will be no incentive to fit common changepoint locations across the different series, and the model is thus equivalent to estimating changepoint locations independently for each series. Similarly if $\alpha = 0$, then the model is equivalent to the single penalty framework with $\mathcal{C}(s,t) := \sum_{j=1}^{p} \mathcal{C}^j(s,t)$. With the selection of an appropriate cost function, it is possible to detect changes in a wide range of datasets. Furthermore we note that different variables can use different cost functions meaning that this approach can also be applied to multivariate datasets of mixed type and distribution.

Similar to the single penalty setting, we can detect subset multivariate changepoints in $\boldsymbol{X}$ by identifying the segmentation that minimises (3.2.1). Furthermore we can identify the optimal segmentation with respect to (3.2.1) via a dynamic program. However unlike the single penalty cost framework, it is not possible to condition on a single time point. Instead we construct a recursion by conditioning on the location of the most recent change in each variable. Suppose we wished to calculate the cost of the optimal subset multivariate segmentation of data $\{\boldsymbol{X}_l\}_{l=1,\ldots,n}$, which we denote by $F(\boldsymbol{c}_n)$ where $\boldsymbol{c}_n := (n, \ldots, n)$. The vector $\boldsymbol{c}_n$ denotes the last point for which the likelihood is calculated at for each variable. Furthermore suppose we knew a priori that the optimal last changepoint in each variable j prior to time n was $c^j_\star$. Let $\boldsymbol{c}_\star := (c^1_\star, \ldots, c^p_\star)$. Then the cost of the optimal subset multivariate segmentation can be computed as,

$$F(\boldsymbol{c}_n) = F(\boldsymbol{c}_\star) + \sum_{j=1}^{p} \left[I\left(c^j_\star \neq c^j_n\right) \left(\mathcal{C}^j\left(c^j_\star, c^j_n\right) + \alpha \right) \right] + m(\boldsymbol{c}_\star, \boldsymbol{c}_n)\beta$$

where $m(\boldsymbol{c}_\star, \boldsymbol{c}_n)$ is the number of changepoints between $\boldsymbol{c}_\star$ and $\boldsymbol{c}_n$ (including the changes at $\boldsymbol{c}_\star$ but not the changes at $\boldsymbol{c}_n$).

It is useful here to compare the segmentation above with the single penalty segmentation by looking at Figure 3.1.1. Both equations split the data into two sections, a left section with known cost and a right section with a single segment for each variable. However under the single penalty framework, the end point for the left section is

the same in each variable, whereas the dual penalty framework allows for different end points for each variable. These end points are encoded through the vector $\boldsymbol{c}_\star$ which we call a changepoint vector. For clarity, we formally define a changepoint vector as well as a partial ordering on the set of changepoint vectors.

Definition 3.2.1 (Changepoint Vectors)**.** *We say that $\boldsymbol{c} := (c^1, \ldots, c^p)$ is a changepoint vector with respect to the data $\boldsymbol{X}$ if $c^j \in \mathbb{Z}$ and $0 \leq c^j \leq n$ for all $1 \leq j \leq p$. Furthermore for any two changepoint vectors $\boldsymbol{c}_a$ and $\boldsymbol{c}_b$, we have a well defined partial ordering $\prec$ such that,*

$$\boldsymbol{c}_a \prec \boldsymbol{c}_b \iff c_a^j \leq c_b^j \ \forall \ 1 \leq j \leq p \text{ and } c_a^j < u_b \, \forall 1 \leq j \leq p \text{ where } u_b = \max_{1 \leq j \leq p} c_b^j.$$

If $\boldsymbol{c}_a \prec \boldsymbol{c}_b$, we say that $\boldsymbol{c}_a$ is prior to $\boldsymbol{c}_b$.

In practice we do not know the optimal prior changepoint vector $\boldsymbol{c}_\star$, and instead must search over all possible changepoint vectors prior to $\boldsymbol{c}_n$. This leads to the following recursive formula for calculating the optimal subset multivariate segmentation.

Theorem 3.2.2 (Pickering, 2016)**.** *For any given changepoint vector $\boldsymbol{c}_f$, let $\Lambda_{\boldsymbol{c}_f} := \{\boldsymbol{c} \prec \boldsymbol{c}_f\}$. Then we have that*

$$F(\boldsymbol{c}_f) = \min_{\boldsymbol{c} \in \Lambda_{\boldsymbol{c}_f}} \Bigg[\underbrace{F(\boldsymbol{c})}_{\text{Optimal Cost up to } \boldsymbol{c}} + \underbrace{\sum_{j=1}^{p} I(c^j \neq c_f^j) \mathcal{C}^j \left(c^j, c_f^j\right)}_{\text{Cost of data after } \boldsymbol{c}} + \sum_{j=1}^{p} \underbrace{I(c^j \neq c_f^j)\alpha}_{\text{Cost of variable } j \text{ being affected by a change}} + \underbrace{m(\boldsymbol{c}, \boldsymbol{c}_f)\beta}_{\text{Cost of any new changepoints}} \Bigg]. \tag{3.2.2}$$

We denote the optimal changepoint vector prior to $\boldsymbol{c}_f$ (i.e. the optimizer of (3.2.2)*) as $\ell(\boldsymbol{c}_f)$.*

The problem of calculating the optimal subset multivariate segmentation is equivalent to solving (3.2.2) for $F(\boldsymbol{c}_n)$ where $c_n^j = n$ for all $1 \leq j \leq p$. We therefore need to first compute $F(\boldsymbol{c})$ for all $\boldsymbol{c} \in \Lambda_{\boldsymbol{c}_n}$. To construct a dynamic program, we need to generate the changepoint vectors in an ordering such that if $\boldsymbol{c}_a$ is generated after $\boldsymbol{c}_b$, then $\boldsymbol{c}_a \not\prec \boldsymbol{c}_b$. Such an ordering is given by the following result.

Proposition 3.2.3. *Let*

$$a_\tau^j := \{\boldsymbol{c} | \max_{1\le k\le p} c^k = \tau \text{ and } c^j = \tau\} \text{ and } A_\tau := \bigcup_{1\le j\le p} a_\tau^j. \tag{3.2.3}$$

Then for changepoint vectors $\boldsymbol{c}, \boldsymbol{c}'$*,*

$$\boldsymbol{c}, \boldsymbol{c}' \in A_\tau \implies \boldsymbol{c} \not\prec \boldsymbol{c}' \text{ and } \boldsymbol{c}' \not\prec \boldsymbol{c}, \tag{3.2.4}$$

$$\boldsymbol{c}' \in A_\tau \text{ and } \boldsymbol{c} \prec \boldsymbol{c}' \implies \boldsymbol{c} \in A_t \text{ for some } t < \tau. \tag{3.2.5}$$

Proof. Proof in Appendix, Section A.2. □

The set A_τ is the set of all changepoint vectors that have a change at τ, but do not have a change after τ. Equation (3.2.4) states that no element of A_τ is prior to another element of A_τ, while equation (3.2.5) states that any changepoint vector prior to $\boldsymbol{c} \in A_\tau$ must be contained in one of the sets $\{A_t\}_{t=1,\dots,\tau-1}$. Thus in order to solve the recursion for some changepoint vector $\boldsymbol{c}_f$, we must solve the recursion for all

$$\boldsymbol{c} \in A_t | \boldsymbol{c} \prec \boldsymbol{c}_f \text{ for } 1 \le t \le u \text{ where } u := \max_{1\le j\le p} c_f^j.$$

Combining this list with the recursive formula defined in (3.2.2) produces a dynamic program, Subset Multivariate Optimal Partitioning (SMOP), for computing the optimal subset multivariate segmentation which is described in Algorithm 1. Let

$$\Omega_n := \{A_\tau\}_{1\le\tau\le n},$$

a complete set of appropriately ordered changepoint vectors. The computational cost of this algorithm is a function of two components, the number of changepoint vectors $\boldsymbol{c}_f \in \Omega_n$ and the size of each set $\Lambda_{\boldsymbol{c}_f}$. For any $\boldsymbol{c}_f \in \Omega_n$, the set of changepoint vectors prior to $\boldsymbol{c}_f$ is equal to a modified cross product of time points prior to each c_f^j and the computational cost of solving (3.2.2) grows rapidly with n. For example, assuming we knew a priori $F(\boldsymbol{c})$ for all $\boldsymbol{c} \in \Lambda_{\boldsymbol{c}_n}$, the computational cost of solving the recursion for $\boldsymbol{c}_n$ is $\mathcal{O}((n-1)^p)$. Furthermore, the number of changepoint vectors in Ω_n also grows rapidly with n. As a result, this method is computationally infeasible for even small datasets and, an efficient implementation of the algorithm requires over 3.5 hours to

compute the optimal subset multivariate segmentation for data of length $n = 100$ with $p = 3$ variables compared with less than a second for the single penalty approach.

In the next section, we introduce a preprocessing algorithm that substantially reduces the computational cost of calculating the optimal subset multivariate segmentation. This procedure reduces the computational cost in two ways. Firstly, it reduces the number of changepoint vectors in Ω_n for which we must solve (3.2.2). Secondly, it reduces the number of changepoint vectors in each $\Lambda_{\boldsymbol{c}}$, reducing the cost of solving (3.2.2) for each $\boldsymbol{c}$. We achieve this computational improvement by extending search space reduction methods from the single penalty setting to the dual penalty setting, as well as introducing novel conditions which exploit the structure of the dual penalty setting.

3.3 Search Space Reduction

In the single penalty setting, a number of authors have explored how the computational cost of dynamic programs can be reduced. Typically this involves defining necessary conditions under which a given candidate change, t, may be an optimal solution to (3.1.2). If the condition is tight then a large number of points will not satisfy it and thus can be excluded, resulting in a significant reduction in the computational cost. In this section we demonstrate how a similar approach can significantly reduce the computational complexity of the subset multivariate algorithm discussed in the previous section.

A natural extension to the dual penalty setting would be to construct conditions which remove changepoint vectors. However checking whether each individual changepoint vector satisfies a given rule would have cost comparable to that of the original algorithm, negating any benefit (Pickering, 2016). Therefore we develop conditions which indicate whether or not complete sets of changepoint vectors are suboptimal. By focusing on sets of changepoint vectors, we substantially reduce the number of conditions that must be checked and thus the computational cost of checking them. As a result, this approach can be used to substantially reduce the computational cost

Algorithm 1: Subset Multivariate Optimal Partitioning (SMOP)

Input : Data $\boldsymbol{X}$ of length n, dimension p, Cost functions $\{\mathcal{C}^j\}$, Penalties α, β

for $1 \leq \tau \leq n$ **do**

  $A_\tau := \underset{1 \leq j \leq p}{\cup} a_\tau^j$;

end

$F((0, \ldots, 0)) = 0$;

for $1 \leq \tau \leq n$ **do**

  for $\boldsymbol{c}_c \in A_\tau$ **do**

    $F(\boldsymbol{c}_c) = \min_{c \prec \boldsymbol{c}_c} F(\boldsymbol{c}) + \sum_{j=1}^{p} I(c^j \neq c_c^j) \left(\mathcal{C}^j \left(c^j, c_c^j\right) + \alpha\right) + m(\boldsymbol{c}, \boldsymbol{c}_c)$;

    $\ell(\boldsymbol{c}_c) = \arg\min_{c \prec \boldsymbol{c}_c} F(\boldsymbol{c}) + \sum_{j=1}^{p} I(c^j \neq c_c^j) \left(\mathcal{C}^j (c^j, c_c^j) + \alpha\right) + m(\boldsymbol{c}, \boldsymbol{c}_c)$;

    $O(\boldsymbol{c}_c) = \ell(\boldsymbol{c}_c) \cup O(\ell(\boldsymbol{c}_c))$;

  end

end

$F(\boldsymbol{c}_n) = \min_{c \prec \boldsymbol{c}_n} F(\boldsymbol{c}) + \sum_{j=1}^{p} I(c^j \neq c_n^j) \left(\mathcal{C}^j \left(c^j, c_n^j\right) + \alpha\right) + m(\boldsymbol{c}, \boldsymbol{c}_n)$;

$\ell(\boldsymbol{c}_n) = \arg\min_{c \prec \boldsymbol{c}_n} F(\boldsymbol{c}) + \sum_{j=1}^{p} I(c^j \neq c_n^j) \left(\mathcal{C}^j (c^j, c_n^j) + \alpha\right) + m(\boldsymbol{c}, \boldsymbol{c}_n)$;

$O(\boldsymbol{c}_n) = \ell(\boldsymbol{c}_n) \cup O(\ell(\boldsymbol{c}_n))$;

Output: Set of optimal changepoint vectors $O(\boldsymbol{c}_n)$

of computing the optimal segmentation.

We construct two types of conditions which are described in Section 3.3.1 and Section 3.3.2. Throughout this section we assume that each cost function is convex with respect to data $\boldsymbol{X}_{t:v}$, i.e. each $\mathcal{C}^j$ satisfies the following equation,

$$\mathcal{C}^j(t,s) + \mathcal{C}^j(s,v) \leq \mathcal{C}^j(t,v). \tag{3.3.1}$$

This is a common assumption in the univariate setting (Killick, Fearnhead, et al., 2012; Maidstone et al., 2017). An algorithm for testing these conditions and solving the reduced optimisation problem is discussed in Section 3.3.3.

3.3.1 Pruning Rule

The recursion in (3.2.2) seeks to identify the most recent changepoint in each variable. If there is significant evidence of a changepoint at time t in variable j, then we should be able to ignore changepoint vectors with changes prior to t in variable j as these would be suboptimal. If the series changes frequently then this computational saving may be substantial. In the single penalty literature, this idea is referred to as pruning. We extend this idea to the dual penalty setting via the following proposition.

Proposition 3.3.1. *Let P_t^j denote the following set of changepoint vectors,*

$$P_t^j := \{\boldsymbol{c} | c^j = t\}. \tag{3.3.2}$$

for some $1 \leq j \leq p$. Suppose $\mathcal{C}^l$ satisfies (3.3.1) *for $1 \leq l \leq p$, and for some triple $0 \leq t < s < v$ we have that,*

$$\mathcal{C}^j(t,v) - \mathcal{C}^j(t,s) - \mathcal{C}^j(s,v) > \beta + \alpha. \tag{3.3.3}$$

Then if $\boldsymbol{c}_f$ is a changepoint vector such that $c_f^j = v$ and $\boldsymbol{c}_p$ is changepoint vector,

$$\boldsymbol{c}_p \in P_t^j \implies \boldsymbol{c}_p \neq \ell(\boldsymbol{c}_f).$$

Proof. Proof in Appendix, Section A.3. □

If the conditions in Proposition 3.3.1 are satisfied by any $0 \leq t < s < v$, then changepoint vectors in P_t^j can be safely excluded when solving the recursion for any changepoint vector with a change at v in variable j, reducing the cost of solving the recursion. In the univariate setting, pruning conditions allow you to prune the candidate for the end point (in this case v) as well as all future end points (values greater than v). Proposition 3.3.1 does not allow for this type of pruning, however we can achieve this type of pruning via a more stringent condition.

Proposition 3.3.2. *Let P_t^j be defined as in* (3.3.2). *Suppose $\mathcal{C}^l$ satisfies* (3.3.1) *for $1 \leq l \leq p$ and, for some triple $0 \leq t < s < v$ we have that,*

$$\mathcal{C}^j(t, v) - \mathcal{C}^j(t, s) - \mathcal{C}^j(s, v) > 2\beta + 2\alpha. \tag{3.3.4}$$

Then if $\boldsymbol{c}_f$ is a changepoint vector such that $c_f^j \geq v$

$$\boldsymbol{c}_p \in P_t^j \implies \boldsymbol{c}_p \neq \ell(\boldsymbol{c}_f).$$

Proof. Proof in Appendix, Section A.3. □

If the conditions in Proposition 3.3.2 are satisfied for some $0 \leq t < s < v$, then changepoint vectors in P_t^j can be safely excluded when solving the recursion for any changepoint vector with a change after v in variable j. Thus the savings obtained from Proposition 3.3.2 are potentially much greater than 3.3.1. The previous two results address the issue of reducing the cost of solving the recursion for a given changepoint vector. The following result demonstrates that the conditions in Proposition 3.3.2 can be used to reduce the number of changepoint vectors for which we must solve the recursion.

Corollary 3.3.3. *Suppose $\mathcal{C}^j$ satisfies* (3.3.1) *for $1 \leq j \leq p$ andthat the condition* (3.3.4) *holds for some t, s, v, $\boldsymbol{c} \in P_t^j$ and $c^k \geq v$ for some $k \neq j$. Then there does not exist a changepoint vector $\boldsymbol{c}_f$ such that $\boldsymbol{c} = \ell(\boldsymbol{c}_f)$.*

Proof. Proof in appendix, Section A.3. □

Corollary 3.3.2 states that if the condition (3.3.4) holds, then we can significantly reduce the number of changepoint vectors for which we must solve (3.2.2), by excluding

any changepoint vector which satisfies Corollary 3.3.2 from all future partial orderings. The computational gains from this type of pruning can be considerable.

The worst case complexity of checking conditions (3.3.3) and (3.3.4) is $\mathcal{O}(pn^3)$. However because of the forward looking nature of Proposition 3.3.2 the cost of checking the conditions is likely to be smaller in practice. In particular, if condition (3.3.4) is satisfied for a given $t < s < v$ and variable j, the set P_t^j can be pruned for all future values v. Thus, under the common assumption of a linear increasing number of changepoints (Killick, Fearnhead, et al., 2012), the expected cost would be linear in the length of the data. Furthermore as we shall see in the simulation study, the computational benefits of pruning via these conditions are substantial.

The pseudocode for a procudure which checks these conditions is described in Algorithm 2. The outputs of this procedure are the following sets,

$$SL := \{SL_{v,j}\}_{1\leq v\leq n, 1\leq j\leq p} \text{ and } CL := \{CL_{v,j}\}_{1\leq v\leq n, 1\leq j\leq p},$$

where

$$\begin{aligned} SL_{v,j} &:= \left\{t | \mathcal{C}^j(t,v) - \mathcal{C}^j(t,s) - \mathcal{C}^j(s,v) < \beta + \alpha \text{ for } t < s < v\right\}, \\ CL_{v,j} &:= \left\{t | \mathcal{C}^j(t,v) - \mathcal{C}^j(t,s) - \mathcal{C}^j(s,v) < 2\beta + 2\alpha \text{ for } t < s < v\right\}. \end{aligned} \tag{3.3.5}$$

In other words, each $SL_{v,j}$ gives the candidates which do not satisfy (3.3.3) in variable j given an end point v. These sets are used to generate the candidate prior changepoint vectors for a given changepoint vector when solving (3.2.2). Similarly, the set $CL_{v,j}$ indicate which time points in variable j have not been pruned at time v. These sets are used to generate the set of changepoint vectors for which we must solve (3.2.2).

3.3.2 Selection Rule

The pruning rules address datasets which change frequently, however we also want computational savings when changes are infrequent. One approach is to consider how much a changepoint improves the model fit. For a changepoint vector to be optimal, it must improve the model fit by more than the minimum penalty for a change. If the improvement does not exceed this threshold, then we can safely exclude them. We

Algorithm 2: Prune: Prune Changepoint Vectors

Input : Data $\boldsymbol{X}$ of length n and dimension p, Cost functions $\{\mathcal{C}^j\}$, Penalties α, β

for $1 \leq j \leq p$ **do**

  $CL_{0,j} = 0; CL_{1,j} = 0$; $SL_{0,j} = 0; SL_{1,j} = 0$;

  for $1 < v < n$ **do**

    $CL_{v,j} := (CL_{v-1,j}, v)$; $SL_{v,j} := CL_{v,j}$;

    for $t \in CL_{v,j}$ **do**

      $D = \max_{t<s<v} C^j(t,v) - C^j(t,s) - C^j(s,v)$;

      if $D > 2\beta + 2\alpha$ **then**

        $CL_{v,j} = CL_{v,j} \setminus \{t\}$;

      else if $D > \beta + \alpha$ **then**

        $SL_{v,j} = SL_{v,j} \setminus \{t\}$;

      end

    end

  end

end

Output: Checklist CL, Selectionlist SL

can build a search space reduction rule from this intuition, however first, for every candidate t and variable j, we need an upper bound on how much including a change at t in variable j improves the model fit. For a given set of points $t < s < v$, this upper bound can be computer as follows,

$$\pi^j_{s,v} := \max_{t \in SL_{s,j}} \left(\mathcal{C}^j(t, v) - \mathcal{C}^j(t, s) - \mathcal{C}^j(s, v) \right).$$

The value $\pi^j_{s,v}$ can be thought of as a bound on the marginal gain from having a segment from s to v in variable j. Note that we are exploiting the fact that some candidates have already been pruned via the set $SL_{s,j}$ to get a tighter bound. The following proposition states that if $\pi^j_{s,v}$ does not exceed α (the minimum cost of having a change affect variable j at s), then changepoint vectors in P^j_s can be safely excluded when solving the recursion for any changepoint vector in P^j_v.

Proposition 3.3.4. *Suppose $\mathcal{C}^j$ satisfies* (3.3.1) *for $1 \leq j \leq p$ and, that for some $s < v$ and $1 \leq j \leq p$ we have that*

$$\pi^j_{s,v} < \alpha.$$

Then if $\boldsymbol{c}_c$ and $\boldsymbol{c}_f$ are changepoint vectors such that $c^j_c = s$ and $c^j_f = v$,

$$\boldsymbol{c}_c \neq \ell(\boldsymbol{c}_f).$$

Proof. Proof in appendix, Section A.4. □

Most candidate changes will violate the constraint $\pi^j_{s,v} < \alpha$. Thus Proposition 3.3.4 is unlikely to substantially reduce the computational cost by itself. However we can combine this with another rule to produce a much more effective subset reduction strategy. For a changepoint vector to be optimal, the improvement in model fit across all affected variables must exceed the minimum penalty for a changepoint (in this case the β penalty). If the improvement does not exceed this threshold then we can safely exclude them. We compute an upper bound on the profit from having a changepoint at time s across all variables as follows,

$$\Pi_s := \sum_{j=1}^{p} \max_{\{v \geq s | \pi^j_{s,v} > 0\}} \left(\pi^j_{s,v} - \alpha \right) I(\pi^j_{s,v} > \alpha). \tag{3.3.6}$$

where I is an indicator function. The following proposition describes how, we can use the bound Π_s to significantly reduce the number of changepoint vectors, for which we must solve (3.2.2).

Proposition 3.3.5. *Suppose $\mathcal{C}^j$ satisfies (3.3.1) for $1 \leq j \leq p$ and for some $s < n$ we have that*

$$\Pi_s \leq \beta \tag{3.3.7}$$

Then if $\boldsymbol{c}_c$ is a changepoint vector such that $\boldsymbol{c}_c \in P_s^j$ for some $1 \leq j \leq p$, there does not exist a changepoint vector $\boldsymbol{c}_f$ such that

$$\boldsymbol{c}_p = \ell(\boldsymbol{c}_f).$$

Proof. Proof in appendix, Section A.4. □

Proposition 3.3.5 states that if Π_s does not exceed β, then we do not need to solve (3.2.2) for any changepoint vector in the set $\cup_{1\leq j\leq p} P_s^j$, substantially reducing the cost of computing the optimal segmentation.

The worst case computational cost of calculating $\pi_{s,v}^j$ is again $\mathcal{O}(pn^3)$. However in practice this computation is likely to be much smaller as the procedure benefits twice from the pruning described in the previous section. Firstly the cost of calculating $\pi_{s,v}^j$ depends on $SL_{v,j}$ which is reduced by the pruning. Secondly we only need to calculate $\pi_{s,v}^j$ for each $s \in SL_{v,j}$ reducing the cost further.

Using the Propositions 3.3.4 and 3.3.5, we can reduce the size of each $SL_{v,j}$ and $CL_{v,j}$ as follows,

$$SL_{v,j} = \{s \in SL_{v,j} | \pi_{s,v}^j > \alpha \text{ and } \Pi_s > \beta\} \text{ and } CL_{v,j} = \{s \in CL_{v,j} | \Pi_s > \beta\}.$$

As a final point we note that the bound $\pi_{s,v}^j$ is defined as a minimum of a function over the set $SL_{v,j}$. However we have just seen that the set $SL_{v,j}$ may be reduced by applying the results from the previous section. Thus we can now recompute $\pi_{s,v}^j$ (and by extension Π_s) to get a tighter bound and further reduce the size of the sets SL and CL. Therefore rather than testing the conditions described above just once, the

conditions can be applied a fixed number of times or until the following condition is satisfied

$$\Pi_s > \beta \text{ for all } s \in \bigcup_{1\leq t\leq n} \bigcup_{1\leq j\leq p} SL_{t,j}.$$

Note checking this condition is trivial. Pseudocode for a procedure which iteratively checks whether the conditions in Propositions 3.3.4 and 3.3.5 are violated is given in Algorithm 3.

3.3.3 Implementation

We now describe how we use the sets SL and CL to calculate the optimal subset multivariate segmentation. The set SL is used to generate optimal prior changepoint vectors for a given changepoint vector. If we wish to solve the recursion in (3.2.2) for the changepoint vector $\boldsymbol{c}$, we can replace the set of prior changepoint vectors $\Lambda_{\boldsymbol{c}}$ with the smaller set,

$$\Lambda'_{\boldsymbol{c}} = \bigcup_{1\leq j\leq p} \bigcup_{t\in SL_{c^j,j}} P_t^j.$$

The set CL is used to generate the ordered list of changepoint vectors for which we must solve the recursion. Let

$$L = \cup_{t=1}^{n} \cup_{j=1}^{p} CL_{t,j} \text{ and } L^j = \cup_{1\leq t\leq n} CL_{t,j}. \tag{3.3.8}$$

Then an ordered set of changepoint vectors can be constructed as follows,

$$b_\tau^j := \{\boldsymbol{c} | c^j = \tau \text{ and } c^k \in CL_{\tau,k} \text{ for } k \neq j\} \text{ and } B_\tau := \bigcup_{j|\tau\in L^j} b_\tau^j. \tag{3.3.9}$$

Thus to solve the recursion in (3.2.2) for $\boldsymbol{c}_n$, we must solve the recursion for all $\boldsymbol{c} \in W_n$, where

$$W_n := \{B_\tau\}_{\tau\in L}$$

is a reduced set of appropriately ordered changepoint vectors. We refer to this procedure as Pruned Subset Multivariate Optimal Partitioning (PSMOP) and pseudocode is given in Algorithm 4.

Algorithm 3: Select: Select profitable changepoint vectors

Input : Data $\boldsymbol{X}$ of length n and dimension p, Cost functions $\{\mathcal{C}^j\}$, Penalties α, β, Checklist CL, Selectionlist SL

while *Improvement* > 0 **do**

  Improvement $= 0$;

  for $0 \leq j \leq p$ **do**

    for $1 < v < n$ **do**

      for $s \in SL_{v,j}$ **do**

        $\pi^j_{s,v} = \max_{t \in SL_{s,j}} C^j(t,v) - C^j(t,s) - C^j(s,v) - \alpha$;

        if $\pi^j_{s,v} < 0$ **then**

          $SL_{v,j} = SL_{v,j} \setminus \{s\}$;

        end

      end

    end

  end

  for $1 \leq s \leq n$ **do**

    $\Pi_s = \sum_{j=1}^{p} \max_{s<v} \pi^j_{s,v}$;

    if $\Pi_s > \beta$ **then**

      Improvement = Improvement + 1 ;

      $SL_{v,j} = SL_{v,j} \setminus \{s\}$ for $1 \leq j \leq p, v > s$;

    end

  end

end

Output: Checklist CL, Selectionlist SL

Algorithm 4: Pruned Subset Multivariate Optimal Partitioning (PSMOP)

Input : Data $\boldsymbol{X}$ of length n and dimension p, Cost functions $\{\mathcal{C}^j\}$, Penalties α, β

(CL, SL) = Prune($\boldsymbol{X}, n, p, \alpha, \beta$);

(CL, SL) = Select($\boldsymbol{X}, n, p, \alpha, \beta, CL, SL$);

$L_j = \cup_{t=1}^{n} CL_{t,j}$ for $1 \leq j \leq p$;

$L = \cup_{j=1}^{p} L_j$;

$A_0 = (0, \ldots, 0)$;

for $\tau \in L$ **do**

$B_\tau := \underset{j|\tau \in L_j}{\cup} b_\tau^j$ Defined in equation (3.3.9)

end

$F((0, \ldots, 0)) = 0$;

for $1 \leq \tau < n$ **do**

for $\boldsymbol{c}_c \in B_\tau$ **do**

$\Lambda = \underset{1 \leq j \leq p}{\cup} \underset{t \in SL_{c^j,j}}{\cup} P_t^j$ Defined in equation (3.3.2)

$F(\boldsymbol{c}_c) = \min_{\boldsymbol{c} \in \Lambda} F(\boldsymbol{c}) + \sum_{j=1}^{p} I(c^j \neq c_c^j) \left(\mathcal{C}^j\left(c^j, c_c^j\right) + \alpha\right) + m(\boldsymbol{c}, \boldsymbol{c}_c)$;

$\ell(\boldsymbol{c}_c) = \arg\min_{\boldsymbol{c} \in \Lambda} F(\boldsymbol{c}) + \sum_{j=1}^{p} I(c^j \neq c_c^j) \left(\mathcal{C}^j(c^j, c_c^j) + \alpha\right) + m(\boldsymbol{c}, \boldsymbol{c}_c)$;

$O(\boldsymbol{c}_c) = \ell(\boldsymbol{c}_c) \cup O(\ell(\boldsymbol{c}_c))$;

end

end

$F(\boldsymbol{c}_n) = \min_{\boldsymbol{c} \prec \boldsymbol{c}_n} F(\boldsymbol{c}) + \sum_{j=1}^{p} I(c^j \neq c_n^j) \left(\mathcal{C}^j\left(c^j, c_n^j\right) + \alpha\right) + m(\boldsymbol{c}, \boldsymbol{c}_n)$;

$\ell(\boldsymbol{c}_n) = \arg\min_{c \prec \boldsymbol{c}_n} F(\boldsymbol{c}) + \sum_{j=1}^{p} I(c^j \neq c_n^j) \left(\mathcal{C}^j(c^j, c_n^j) + \alpha\right) + m(\boldsymbol{c}, \boldsymbol{c}_n)$;

$O(\boldsymbol{c}_n) = \ell(\boldsymbol{c}_n) \cup O(\ell(\boldsymbol{c}_n))$;

Output: Set of optimal changepoint vectors $O(\boldsymbol{c}_n)$

3.4 Simulations

In this section, we study the effectiveness of the dual penalty framework on a range of simulated data sets. We begin by measuring the computational savings achieved via the pruning and selection procedures introduced in Section 3.3, and then examine whether the dual penalty changepoint estimator accurately locates changepoints and affected subsets. To facilitate this analysis we now define a number of error metrics, which we use throughout the section.

Firstly throughout we use $\boldsymbol{\tau} := \{\tau_1, \ldots, \tau_m\}$ and $\hat{\boldsymbol{\tau}} := \{\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{m}}\}$ to denote the set of true changepoints and the set of estimated changepoints respectively. Note $\hat{m}$ need not equal m. A common approach for evaluating changepoint methods is to examine true and false discovery rates. We say that the changepoint estimate τ_i has been detected if

$$\min_{1 \leq j \leq \hat{m}} |\hat{\tau}_i - \tau_j| \leq h.$$

Throughout this section we set $h = 10$, although it should be noted that in reality the desired accuracy would be application specific. We denote the set of correctly estimated changes by $\boldsymbol{\tau}_c$. Then we define the true discovery rate (TDR) and false discovery rate (FDR) as follows,

$$TDR := \frac{|\boldsymbol{\tau}_c|}{|\boldsymbol{\tau}|}, \; FDR := \frac{|\hat{\boldsymbol{\tau}}| - |\boldsymbol{\tau}_c|}{|\hat{\boldsymbol{\tau}}|}.$$

The TDR is the proportion of the correctly estimated true changes, while the FDR is the proportion of estimated changes that correctly estimate a true change. An important concern is whether or not the segmentation allows us to accurately estimate the model parameters. Therefore we also report the Mean Square Error (MSE) for all variables,

$$MSE := \frac{1}{n} \sum_{j=1}^{p} \sum_{i=1}^{n} \|\boldsymbol{\theta}_i^j - \hat{\boldsymbol{\theta}}_i^j\|_2^2.$$

Note we use this metric to compare fully multivariate methods with the dual penalty approach. This measure will favour the dual penalty approach if it correctly estimates all the subsets, as the subset segmentation does not overfit the data. However if the

dual penalty approach does not correctly estimate all the subsets, this metric may prefer a fully multivariate approach which only needs to detect changepoints.

3.4.1 Computational Savings from Preprocessing

The search space reduction techniques introduced in Section 3.3 should reduce the computational cost of solving the subset multivariate optimization problem exactly. However we have yet to quantify how large these savings are in practice. Quantifying the savings is important for two reasons. Firstly, we need to demonstrate that the computational savings from the preprocessing algorithm exceed the cost of running the preprocessing step. Secondly, we would like to understand how the computational cost of the algorithm scales with respect to the dimension of the data in practice. Finally, it is important to acknowledge that the computational cost of both SMOP and PSMOP are very large in practice (even thought PSMOP is much quicker). In particular, both algorithms scale poorly as the dimension increases. As a result, for all these simulations we only consider $p = 3$ to ensure computational feasibility.

Comparison of pruned and unpruned computational cost

We generated 100 datasets of length $n = 100$ and dimension $p = 3$. The data is normally distributed with unit variance and changes in mean. There are 3 segments with lengths $(33, 33, 34)$. The segment parameters for each variable respectively are as follows,

$$(0, \delta, \delta), (0, 0, -\delta), (0, \delta, 0)$$

where $\delta = 1.25$. We use a cost function based on the likelihood for normal data with unit variance i.e.

$$\mathcal{C}^j(s, t) := \sum_{i=s+1}^{t} x_i^2 - \frac{\left(\sum_{i=s+1}^{t} x_i\right)^2}{t - s}.$$

For each dataset we computed the optimal subset multivariate segmentation with and without the preprocessing algorithm and recorded the time taken to solve the optimisation problem for each case as well as the the computation time for the preprocessing

algorithm. The results of this analysis are shown in Table 3.4.1. Looking at the table, it is immediately clear that the preprocessing algorithm substantially reduces the cost of solving the optimisation problem. In fact, on average the unpruned algorithm takes almost 500 times longer to run, demonstrating that the preprocessing algorithm substantially reduces the cost of computing the optimal subset partition.

Scaling with respect to size of change

Pruning rules remove candidate changepoints if they identify a candidate change (s in Proposition 3.3.1) that dramatically improves the model fit. However if the change is small, the improvement may not exceed the pruning threshold and the method will not prune any changes. Note this impacts the selection rules as well, through the marginal profit term. This intuition indicates that the computational cost of our approach may depend on the size of the change. To investigate this we repeated the above simulation and varied the δ parameter. In particular, we generated 100 datasets as above for each $\delta = \{.5, .75, 1, 1.25, 1.5, 1.75, 2\}$. We computed the optimal subset segmentation using the PSMOP algorithm and recorded the time taken (including preprocessing). The results are shown in Figure 3.4.1(a). We can see that as the size of the change increases, the computational cost drops substantially.

Scaling with respect to the dimension of the data

We are interested in how the method scales with respect to the length of the data n, particularly when the number of changepoints increases linearly with n and when the number of changepoints is fixed (the best and worst case scenarios for pruning in the univariate setting). We consider the increasing case first. We generated 100 datasets for each $n = \{100, 200, 300, 400, 500, 600, 700, 800\}$ and $p = 3$. The sequence from the first experiment is repeated $n/100$ times to obtain a linearly increasing number of changepoints. We ran the preprocessing algorithm on each dataset and calculated the number of times we need to solve the recursion (3.2.2) in order to compute the optimal subset segmentation. The results are shown in Figure 3.4.1(b). We can see that in this scenario the number of recursions increases as a linear function of the data. This

Method	Minimum Time (s)	Mean Time (s)	Max Time(s)
SMOP	11224.2	14130.61	22530.24
Preprocessing	9.33	22.44	38.7
Dynamic Program	.038	8.45	53.51
PSMOP	9.41	30.89	73.91

Table 3.4.1: Computational runtime for the pruned and non pruned dynamic program with $n = 100$ and $p = 3$.

matches results for the single penalty framework, which state that the computational cost of the dynamic program is $\mathcal{O}(n)$ when the number of changepoints grows linearly with n.

For the scenario with a fixed number of changes the segment parameters were

$$(0, \delta, \delta), (0, 0, -\delta), (0, \delta, 0),$$

with lengths $(33n/100, 34n/100, 33n/100)$. The results of this analysis can be seen in Figure 3.4.1(c). The required number of recursions remains roughly constant as n grows indicating that the selection rule is successfully restricting the set of feasible changepoints to those close to the true change. Note that because there are fewer changepoints the preprocessing procedure takes much longer to run in the second scenario as less pruning occurs. It is useful here to compare the results for the two settings. The number of recursions required when the number of changes is constant is much smaller (by a factor of 10) than the number of recursions required when the number of changes grows with n. This is the reverse of what we typically find in the single penalty setting and is due to the fact that the selection rules significantly reduce the search space. In particular, the selection rules remove all the points outside a narrow window around the change. However since each change gets a window, if there are a lot of changes than the procedure will be slower.

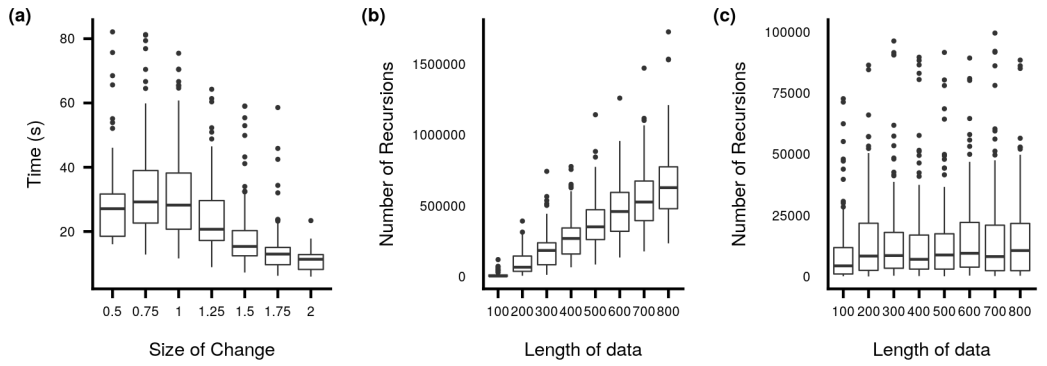

Figure 3.4.1: (a) Boxplot of time to run PSMOP (including the preprocessing step) vs the size of the change. (b) Boxplot of number of recursion solves vs n where the number of changes increases with n. (c) Same where the number of changes is fixed. Note that unlike in the single penalty setting the dynamic program is faster when there are fewer changes. This is due to the selection procedure discussed in Section 3.3.2

3.4.2 Performance of Dual Penalty framework

We now meaure how capable our proposed approach is at detecting subset multivariate changepoints. We consider two changepoint problems that have received significantly less attention than the problem of detecting changes in mean of normal data. In particular, we examine changes in variance of normal data and changes in rate of Poisson data. One advantage of the cost function approach is that we do not need to transform the data before applying the method. Instead we can define cost functions based on the likelihood for these data. In theory, this should increase our ability to detect changes.

Changes in variance of normally distributed data

We generated 100 datasets of length $n = 400$ and dimension $p = 3$. The data is normally distributed with changes in variance. We use a cost function based on the likelihood for normal data with zero mean and unknown variance i.e.

$$\mathcal{C}^j(s,t) := (t-s)\left(\log \hat{\sigma}^2_{s,t} - \log(t-s) + 1\right) \text{ where } \hat{\sigma}^2_{s,t} := \sum_{i=s+1}^{t} \left(x_i - \sum_{i=s+1}^{t} x_i \right)^2 .$$

It is possible to treat this changepoint problem as a change in mean problem by first squaring the data. However, this approach is likely to be ineffecient, as the data will feature nuisance changes in variance which make it more difficult to select a threshold and increase the probability of getting a false positive. There are 5 segments with lengths $(25, 50, 150, 75, 100)$. The segment parameters for each variable respectively are as follows,

$$(1, 2.2, 2.2, 0.5, 1.3), (1, .4, 1, 1, 1.5, 1.5), (1, 1, 2, 1.5, 0.8).$$

An example dataset is shown in Figure 3.4.2. Note there is no correlation between the three variables. Examining the parameters we can see that each dataset has 4 changepoints with $\boldsymbol{\tau} = \{25, 75, 225, 300\}$ and $\boldsymbol{\tau}^1 = \{25, 255, 300\}, \boldsymbol{\tau}^2 = \{25, 75, 225\}, \boldsymbol{\tau}^3 = \{75, 225, 300\}$. Examining Figure 3.4.2 we can see that there is a mix of small and large changes, and short and long segments. We apply the dual penalty estimator to each of the datasets. For comparison purposes we also applied the univariate PELT method to each variable separately. For the dual penalty estimator, we used a BIC type penalty that adds $\log n$ for each extra parameter. This is equivalent to setting $\alpha = \log n$ and $\beta = \log n$.

A histogram of estimated changepoint locations for the PELT (red) method and the dual penalty approach (green) are shown in Figure 3.4.2 (b). Looking at the plot, we can see that the dual penalty approach consistently does a better job detecting changepoints across the replications. This is particularly noticeable for the change at $t = 225$. The univariate approach struggles to detect the subtle changes in the second and third variables, whereas the dual penalty approach is able to detect the changes by combining information across the different variables. This improvement in performance can also be in the performance metrics for this dataset shown in Table 3.4.2. We can see that there is a statistically significant difference between the methods for the TPR and MSE error metrics for the second and third variables, with the dual penalty approach performing the best. The dual penalty also reports a statistically lower FPR for the second variable. In summary, the dual penalty approach successfully shares information across variables improving the accuracy of

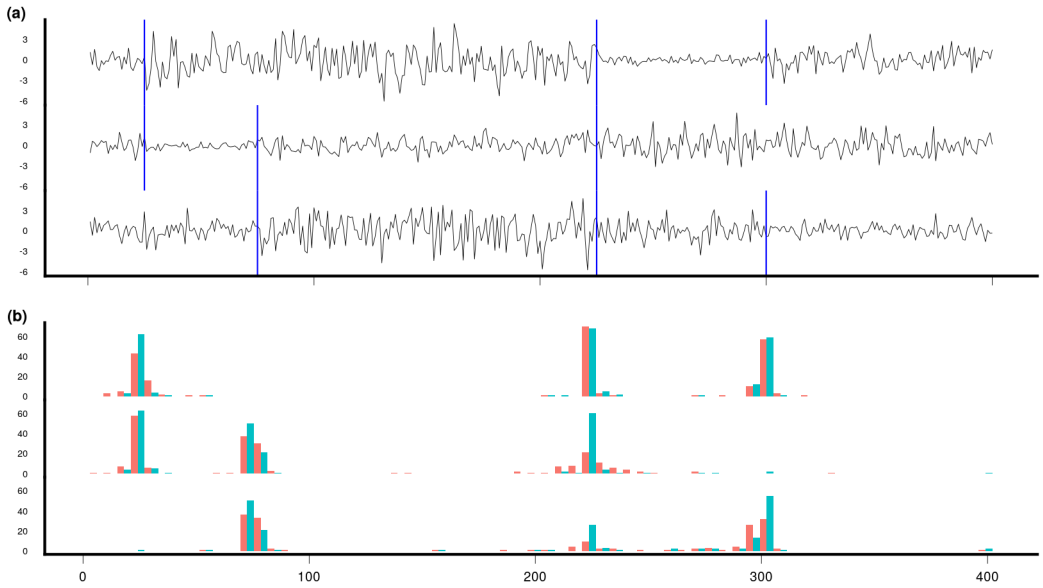

Figure 3.4.2: (a) Multivariate normally distributed data with four changes in variance. (b) Estimated subset multivariate segmentation via the dual penalty approach (green) and estimated univariate segmentation from the PELT algorithm (red).

the method. This improvement in accuracy does not cause the FPR to increase and the dual penalty method accurately reports the subset structure of the changepoints.

Changes in rate of Poisson data

From the above results, we can see that the dual penalty approach successfully detects changes in variance of normal data. However it is worthwhile also studying the performance of the dual penalty approach on other datasets, as a significant advantage of the dual penalty framework is the flexibility of the method. One important example from the literature is the problem of detecting changes in count data (Franke et al., 2012), which can be modelled as changes in the mean level of Poisson data.

We generated 100 datasets of length $n = 300$ and dimension $p = 3$. Examining Figure 3.4.3, we can see that there is a mix of small and large changes, and short and long segments. We use a cost function based on the likelihood for Poisson data with

Method	Variable	TPR	FPR	MSE
PELT	1	(0.93, 0.98)	(0.02 , 0.08)	(10.05, 14.45)
Dual Penalty	1	(0.98, 1.00)	(0.01 , 0.04)	(7.65 , 12.10)
PELT	2	(0.79, 0.88)*	(0.10 , 0.19)*	(6.71 , 9.55)*
Dual Penalty	2	**(0.94, 0.99)***	**(0.02 , 0.07)***	**(4.02 , 6.15)***
PELT	3	(0.63, 0.72)*	(0.07 , 0.17)	(16.28, 19.25)*
Dual Penalty	3	**(0.73, 0.82)***	(0.05 , 0.14)	**(11.69, 15.20)***

Table 3.4.2: 95% confidence intervals for mean errors for the dual penalty approach and PELT on normally distributed data with changes in variance. A statistically significant difference is indicated by $*$ and the best value is in bold.

unknown rate i.e.

$$\mathcal{C}^j(s,t) := f_{s,t}\left(\log\left(t-s\right) - \log f_{s,t} + 1\right) \text{ where } f_{s,t} := \sum_{i=s+1}^{t} x_i.$$

As in the previous example, it is possible to treat this is a change in mean problem by transforming the data. In particular, the Anscombe transform should produce approximately normal data with unit variance provided that the rate is sufficiently large. However a cost function based on the likelihood function of the data should provide greater statistical accuracy. There are 4 segments with lengths $(50, 125, 50, 75)$. The segment rates for each variable respectively are as follows,

$$(3.5, 5, 5, 7), (10, 10, 8.5, 6.5), (1.5, 3, 5, 5).$$

Each dataset has 3 changepoints with $\boldsymbol{\tau} = \{50, 175, 225\}$ and $\boldsymbol{\tau}^1 = \{50, 225\}, \boldsymbol{\tau}^2 = \{175, 225\}, \boldsymbol{\tau}^3 = \{50, 175\}$. We compare the dual penalty approach with the Inspect method (T. Wang and Samworth, 2018) and the E.divisive method (Matteson and James, 2014). Note the Inspect method assumes the data is Gaussian, while the E.divisive method is non-parametric. There is not to our knowledge a parametric method for detecting changes in multivariate count data, thus we argue that these are a reasonable comparison. We use the same default penalties as before with $\alpha =$

$\beta = \log n$. Furthermore we set the cost function equal to twice the negative Poisson likelihood.

A histogram of estimated changepoint locations for the dual penalty approach (green) is shown in Figure 3.4.3 (b). Looking at the plot, we can see that the dual penalty approach detects changepoints and subsets reliably. Looking at the second change (which is smaller in magnitude than the others), we can see that the variance of the changepoint location depends on the size of the change as expected. However we note that the method struggles to detect the first change in the second variable. We can see the performance metrics for our method and the competitor methods are shown in Table 3.4.3. Both our method and Inspect are able to detect the changepoint locations as noted by the TPR, while the E.divisive method performs worse (although this difference is not significant). However our approach reports a statistically smaller FPR than the other approaches. For the Inspect method, the larger FPR is unsurprising as it does not take account of the fact that the data has a Poisson distribution. The larger FPR reported by the E.divisive method is more surprising as the dataset satisfies the assumptions of this approach. Although our approach does a better job in detecting changepoints, it reports the largest MSE. This is due to the fact that it regularly misses a change for the second variable. In summary, the dual penalty approach with default penalties can detect changepoints and subsets in count data, without an increase in false positives that can affect other change in mean methods.

	Method	TPR	FPR	MSE
1	Dual Penalty	(.87,.9)	**(.04,.07)***	(66.36,79.2)*
2	Inspect	(.87,.9)	(.08,.12)*	**(49.24,54.52)***
3	E.divisive	(.8, .88)	(.11, .19)*	(52.75, 64.88)*

Table 3.4.3: 95% confidence intervals for mean errors for the dual penalty approach and Inspect and E.divisive on Poisson data with changes in rate. A statistically significant difference is indicated by $*$ and the best value is in bold.

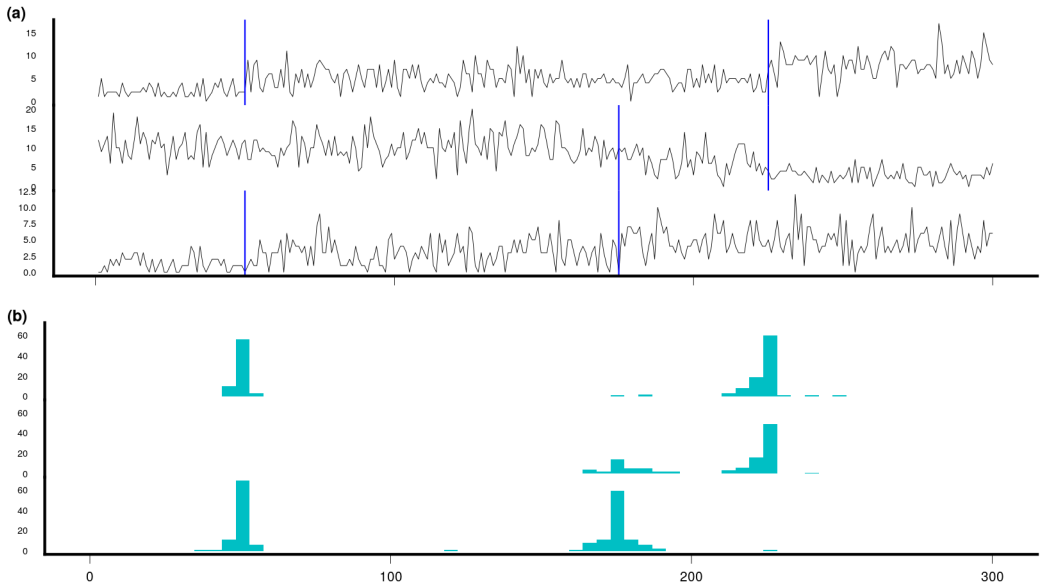

Figure 3.4.3: (a) Multivariate count data with three sparse changes in mean. (b) Estimated subset multivariate segmentation via the dual penalty approach.

3.5 Application to Covid 19 data in the UK

The current Covid-19 pandemic has presented a number of cascading social, political, economic and humanitarian crises. As a result, there has been significant interest in measuring changes in the spread of the disease. Increases in the spread of the disease present a serious concern for policymakers, while decreases can reflect effective policy interventions. Previous work utilised univariate changepoint models to study changes in the spread of Covid-19 within each of the 50 US states (Wagner et al., 2020). Given the similarity between some states, we would expect changes to occur at the same time across multiple locations, motivating a multivariate approach.

In this work, we study the daily case reports for the three constituent countries within Great Britain; Scotland, England and Wales which can be seen in Figure 3.5.1. Our goal is to detect changes in the doubling rate of the daily cases i.e. the slope of the $\log_2$ of daily cases. A significant challenge with this dataset is that the mean level of the data is time dependent, which means that there is nonstationary temporal behaviour within segments. Many multivariate methods do not allow for this type of behaviour and are limited to settings with either temporal independence

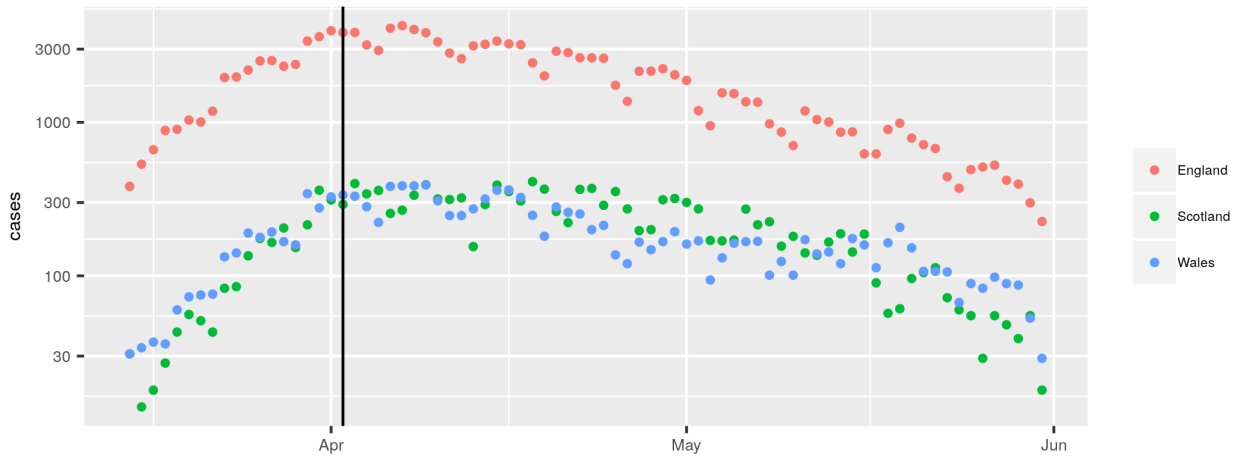

Figure 3.5.1: New cases of Covid-19 by day for England, Scotland and Wales on a log scale with detected changepoint.

or a stationary dependence in the noise. A significant advantage of the dual penalty framework is that it can be applied directly to this problem by selecting an appropriate cost function. Let

$$\log_2(X_t^j) = \theta_{k,0} + \theta_{k,1}t + \epsilon_t$$

where $\theta_{k,0}$ and $\theta_{k,1}$ are segment specific intercept and slope terms respectively and $\{\epsilon_t\}_{t=1}^n$ is a sequence of IID standard normal variables. Then we can let $\mathcal{C}^j$ be twice the negative log likelihood given by this model and we are interested in detecting changes in the intercept and slope term. We again use a BIC type penalty and set $\beta = \log n$ and $\alpha = 2\log n$ (since each segment has two parameters).

Our analysis finds a single changepoint which affects each country on April 2nd. This is 10 days after the goverment announced the stay at home period on March 23rd and indicates that there is a 10 day delay period between the public health intervention and a response in the data. This is similar to results from previous work which found an 11-12 day delay between a public intervention and outcome in Covid-19 case data (Wagner et al., 2020). Finally we note that this approach could be used to detect future changes in the number of cases, such as a future outbreak as well as differing impacts in the countries due to different local approaches.

3.6 Conclusions

In this chapter we have discussed the subset multivariate segmentation model for detecting multivariate changepoints, examined how subset multivariate segmentations can be estimated via a dynamic program and introduced a number of techniques which substantially reduce the computational cost of this dynamic program. In simulations, we demonstrated the advantages of this approach over current univariate and multivariate methods. Finally, we used this method to study changes in the growth rate of Covid 19 within Great Britain, demonstrating the value of a cost function based approach to multivariate segmentations.

While the proposed dual penalty approach has a number of advantages there is still a significant limitation; despite the savings achieved through the proposed preprocessing algorithm, the computational cost of the procedure scales poorly with both n and (especially) p. As a result, there are a number of applications where this approach may be useful but is infeasible due to the size of the data. To address such applications, in the next chapter we introduce a computationally efficient approximate algorithm for calculating subset multivariate segmentations based on this dual penalty framework. This algorithm uses an approximate cost function to produce a much simpler dynamic program, which can be solved in at worst quadratic time while still providing good solutions to the original dual penalty optimisation problem.

Chapter 4

Approximate Subset Multivariate Changepoints

4.1 Introduction

In the age of Big Data, datasets of increasing length, dimension and complexity are being collected. Often, the underlying distributional properties of these datasets can change over time. In order to accurately model these datasets it is necessary to take account of this heterogenity. One approach is to assume that the changes occur at a small number of time points known as changepoints. Changepoints are relevant in a wide range of applications including finance (J. Chen and Gupta, 1997), network traffic analysis (Kwon et al., 2006) and oceanography (Killick, Eckley, et al., 2010), and a significant literature has been developed on the problem of detecting and locating them. Much of this literature is focused on the univariate setting. A number of papers have examined this problem such as Killick, Fearnhead, et al., 2012, Frick et al., 2014 and Fryzlewicz, 2014.

In the univariate setting, a common approach to detecting changepoints is to define a cost function for a segmentation and then minimise a penalised version of this cost function. If, conditional on the locations of the changes, the costs of the segments are independent, then this optimisation can be solved exactly via dynamic programming with computational cost $\mathcal{O}(K(n)n^2)$ where n is the length of the data and $K(n)$ is the

cost of evaluating the cost function. Note many commonly used cost functions can be evaluated using summary statistics and thus $K(n) = \mathcal{O}(1)$. The computational cost of this dynamic program can be significantly reduced to $\mathcal{O}(n)$ (Killick, Fearnhead, et al., 2012; Maidstone et al., 2017) in certain settings. This approach is flexible since it utilises a generic cost function without placing assumptions on the underlying distribution of the data or the type of change. Due to the speed and flexibility of the method, it has become popular among practitioners.

The literature on detecting changepoints in multivariate time series has grown substantially in recent years. Multivariate datasets with changepoints have appeared in a wide range of applications including modelling fMRI scans in a dynamic setting (Cribben and Yu, 2017) and measuring the effect of blindness treatments on mice (Storchi et al., 2019). Subsequently there has also been greater interest in changepoint methods for multivariate time series. Methods have been developed for detecting changes in a range of different settings. These include changes in covariance structure (Aue, Hörmann, et al., 2009 and D. Wang, Yu, and Rinaldo, 2017), graphical models (Gibberd and Nelson, 2017) and network structure (D. Wang, Yu, and Rinaldo, 2018).

The multivariate nature of the problem brings with it additional challenges. Unlike the univariate case, it is not necessary for every time series under observation to be affected by a change. This makes it much more difficult to aggregate information across series. Methods that assume every variable changes will lose statistical power if only a subset of the variables are actually affected. Furthermore, depending on the application, it may be interesting to determine which series are (or are not) affected by a change.

A number of authors have considered the problem of detecting changepoints in multivariate time series where there is some dependence between series. Matteson and James, 2014 utilise a nonparametric energy statistic based on pairwise distances between points to detect changes in the underlying distribution of multivariate time series. Arlot et al., 2019 utilise cost functions based on semi-positive definite kernels to detect changepoints without assumptions on the distribution. This work has been extended by Celisse et al., 2018, who develop more computationally efficient approxi-

mate methods based on the kernel approach and by Garreau, Arlot, et al., 2018, who prove that the resulting method is consistent under certain regularity assumptions. H. Chen and Zhang, 2015 introduce a graph based approach for multivariate changepoint detection. Their method can be applied to any dataset where an appropriate similarity measure can be defined.

Another approach is to consider how to combine or aggregate information across multiple series in order to estimate changepoints. These methods do not allow for dependence between series. Zhang et al., 2010, Horváth and Hušková, 2012 and Enikeeva and Harchaoui, 2019 develop test statistics for detecting changes in the mean of normal data with homogenous variance. These methods aggregate information by taking a pointwise mean or max of univariate statistics. Cho and Fryzlewicz, 2015 and T. Wang and Samworth, 2018 consider how to combine information across series whilst taking account of the fact that not every series may be affected by the change. Cho and Fryzlewicz, 2015 only aggregates information across series if a univariate statistic exceeds a wavelet based threshold. T. Wang and Samworth, 2018 aggregate information using a weighted average. The weights are estimated by solving an optimisation problem with an ℓ_1 penalty. As a result the weights for unaffected series should be zero. In both approaches, the subset of affected series is not of interest nor output by software.

In this chapter, we consider how to simultaneously estimate multivariate changepoints and the set of variables affected by changes. Bardwell et al., 2018 study detecting changes in panel data where changepoints are allowed to affect only a subset of the data. This method detects multiple changepoints, but only outputs the most recent change in each series ignoring prior changes. This is inappropriate in our setting where we would like to locate all changes. As in the univariate setting, there is benefit to considering the optimisation problem. Pickering, 2016 develop a penalised cost function framework that incorporates two penalties, one for introducing a changepoint and another for having a variable affected by the change. The changepoints and set of affected subsets can be estimated by optimising this function via a dynamic program. This penalised cost function framework is flexible, however due to the computational

cost of the optimisation, this approach is infeasible for datasets of even moderate size. An ideal method would have this flexibility, while still being feasible for large datasets.

In this paper we propose an approximate optimisation algorithm for the penalised optimisation problem introduced by Pickering, 2016. Much effort in the univariate setting has been devoted to approximating the optimisation step. We take an alternative approach and consider an approximation of the cost function for each segment. This is based on the idea of windowed cost functions, where the model parameters are estimated on a subset of the data. These cost functions provide a basis for an approximate dynamic program with worst case complexity $\mathcal{O}(pn^2)$, comparable to $\mathcal{O}(n^2)$ in the univariate setting. Furthermore, under mild conditions, the algorithm has computational cost which is $\mathcal{O}(np)$.

The paper is organised as follows. We begin in Section 4.2, by discussing the subset multivariate approach introduced by Pickering, 2016. In Section 4.3, we define windowed cost functions for detecting changepoints. We then introduce our new efficient approximate search method, SPOT. We demonstrate that this method always finds better solutions than comparable fully multivariate methods and discuss the computational cost. In Section 4.4, we demonstrate the accuracy and efficiency of our approach via simulations, Section 4.5 applies SPOT to real world datasets.

4.2 Multivariate Changepoint Model

As we use the same framework for our optimisation, this section discusses the subset multivariate changepoint model introduced in Pickering, 2016. In particular, we discuss how a multivariate penalised cost function framework can be used to locate subset multivariate changepoints, and examine how this function can be optimised using a dynamic program.

We begin by defining notation. Suppose we have ordered data $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n \in \mathbb{R}^p$. We use x_i^j to denote the jth element of the vector $\boldsymbol{X}_i$. Note, throughout we use the subscript to refer to time and the superscript to refer to the variate. We denote

the number of changepoints by m and their locations by $\boldsymbol{\tau} = \{\tau_0, \tau_1, \tau_2, \ldots, \tau_m, \tau_{m+1}\}$. We assume that each τ_k is integer valued with $\tau_k < \tau_l$ for $k < l$ and $\tau_0 = 0$ and $\tau_{m+1} = n$. For each changepoint τ_k, we associate a set $\mathcal{S}_k \subseteq \{1, \ldots, p\}$ and a vector $\boldsymbol{S}_k := \{S_k^1, \ldots, S_k^p\}$ where $S_k^j = 1$ if $j \in \mathcal{S}_k$ and $S_k^j = 0$ otherwise. If $j \in \mathcal{S}_k$ then we say that the variable j is affected by the change at τ_k. For notational simplicity we let $\mathcal{S}_0 = \mathcal{S}_m = \{1, \ldots p\}$. We are interested in accurately estimating m, $\boldsymbol{\tau}$ and $\boldsymbol{\mathcal{S}} := \{\mathcal{S}_1, \ldots, \mathcal{S}_m\}$.

We use $\mathcal{C}^j$ to refer to a cost function for the jth variable of a multivariate time series. Note, we allow cost functions to vary across variables. For simplicity of notation, given a dataset $\boldsymbol{X}$ we define

$$\mathcal{C}^j(s,t) := \mathcal{C}(x_{s+1}^j : x_t^j) := \mathcal{C}(\{x_{s+1}^j, \ldots, x_t^j\}).$$

Pickering, 2016 use changepoint vectors, to encode information about the affected subsets into their penalised cost function framework. Changepoint vectors are non-negative integer valued vectors. We denote the jth entry of the changepoint vector $\boldsymbol{c}$, by c^j. This entry gives the location of a changepoint that affects variable j. Given two changepoint vectors $\boldsymbol{c}_a$ and $\boldsymbol{c}_b$ we can define a strict partial ordering as follows,

$$\boldsymbol{c}_a \prec \boldsymbol{c}_b \iff c_a^j \leq c_b^j \ \forall \ 1 \leq j \leq p \text{ and } c_a^j < u_b \, \forall 1 \leq j \leq p \text{ where } u_b = \max_{1 \leq j \leq p} c_b^j$$

Then, an ordered collection of changepoint vectors describes the location of each of the changepoints, as well as the set of affected variables.

4.2.1 Penalised Cost Function

We now review some important background material for this chapter. These concepts have been covered in depth in the previous two chapters and but are shown here for clarity. Readers familiar with these concepts are encouraged to skip ahead to Section 4.3. A common approach to locating changepoints for univariate data $x_1, \ldots x_n$ is to minimise

$$\sum_{i=1}^{m+1} [\mathcal{C}(\tau_{i-1}, \tau_i)] + \beta m \tag{4.2.1}$$

where $\mathcal{C}$ is a cost function and β is a penalty to guard against overfitting. Pickering, 2016 extend this approach to the subset multivariate setting. Let $\boldsymbol{\tau}^j$ and m^j be the set and number of changepoints that affect variable j i.e.

$$\boldsymbol{\tau}^j = \bigcup_{\{0 \leq k \leq m+1 | S_k^j = 1\}} \tau_k \text{ and } m^j = \sum_{k=1}^{m} S_k^j.$$

Similarly, let $\boldsymbol{m} := \{m^1, \ldots, m^p\}$ and $\mathcal{T} := \{\boldsymbol{\tau}^1, \ldots, \boldsymbol{\tau}^p\}$. Then given data $\boldsymbol{X}$, the optimal subset multivariate segmentation is given by the solution to the following opimisation problem

$$\min_{\boldsymbol{m}, \mathcal{T}} \sum_{j=1}^{p} \sum_{k=0}^{m^j+1} \left(\mathcal{C}^j \left(\tau_k^j, \tau_{k+1}^j \right) + \alpha \right) + \beta \psi(\mathcal{T}) \tag{4.2.2}$$

where $\mathcal{C}^j$ is a cost function for variable j, α and β are penalty parameters to guard against overfitting, and ψ is a function which counts the number of unique elements in a set. This approach allows a different cost function for each variable, which implies that variables need not be identically distributed.

The novel addition here is the second penalty parameter, α. In this framework, each changepoint location incurs a fixed cost of β and with a further α cost incurred for each variate that is affected. In this work, we are not concerned with how these penalties should be set, focusing instead on how to optimise (4.2.2).

This approach is flexible, since any cost function that can be used in the univariate setting can also be used in this setting. However the value of $\mathcal{C}^j(s,t)$ is exclusively a function of the data $X^j_{(s+1):t}$ and as such the model does not allow for dependence between cost functions. Hence whilst we allow dependence structures within series (such as auto correlation), we do not allow dependence structures between series (such as cross correlation).

4.2.2 Subset Multivariate Optimal Partitioning

Pickering, 2016 develop a dynamic program, Subset Multivariate Optimal Partitioning (SMOP), using a recursion based on changepoint vectors that can be used to minimse (4.2.2). They demonstrate that the penalised cost of the segmentation of data up to

the changepoint vector $\boldsymbol{c}_u$, denoted by $F(\boldsymbol{c}_u)$, is given by,

$$\min_{\boldsymbol{c} \prec \boldsymbol{c}_u} F(\boldsymbol{c}) + \sum_{i=1}^{p} \left[I\left(c^j \neq c_u^j\right) \left(\mathcal{C}^j\left(c^j, c_u^j\right) + \alpha \right) \right] + m(\boldsymbol{c}, \boldsymbol{c}_u)\beta \tag{4.2.3}$$

where $m(\boldsymbol{c}, \boldsymbol{c}_u)$ is the number of unique positive elements of the vector $\boldsymbol{c}_u - \boldsymbol{c}$.

Intuitively, the cost of a changepoint vector can be expressed as the cost of a prior changepoint vector and the cost up to the new change. Whereas in the univariate case, we obtain a recursion by conditioning on the location of the most recent changepoint, we now condition on the locations of the most recent changepoint in each variable.

The problem of finding the optimal segmentation is, therefore, equivalent to finding $F(\boldsymbol{c}_n)$ by recursively calculating $F(\boldsymbol{c})$ for all $\boldsymbol{c} \prec \boldsymbol{c}_n$. These sets explode in size as n increases. Unfortunately this means that the dynamic program has prohibitively expensive computational cost which is infeasible for even small datasets.

Conditioning on changepoint vectors creates difficulties, since the set of changepoint vectors explodes as n and p increase. It would be more efficient to condition on the last changepoint location, as this is a much smaller set. However for a generic cost function this is impossible. Suppose the jth variable was not affected by a change at t. Then the cost of that segment would depend on data before and after t. In particular, since a generic cost function can use all the data for parameter estimation, the cost of the data up to time t would depend on data after this point. This means conditioning on a change at time t is meaningless since the cost of the data before t is constantly changing as we observe new data. In order to condition on the last changpoint location we therefore need to restrict our attention to cost functions that can be more easily partitioned.

4.3 Subset Partitioning Optimal Time (SPOT)

We saw in the previous section how the exact dynamic program was computationally infeasible due to the size of the set of possible changepoint vectors. In this section, we approximate the cost function to give a computationally feasible solution.

4.3.1 Windowed Cost Functions

Let $\mathcal{C}^j(\cdot, \cdot|\theta^j)$ denote a parametric cost function with parameters θ^j and $\theta^j(p, q)$ denote the parameter estimates using the data $x^j_{p:q}$. Then the windowed cost for a segment is given by

$$\hat{\mathcal{C}}^j(p, q, w) = \begin{cases} \mathcal{C}^j\left(p, q|\theta(p, p + w)\right) \text{for } q - p \geq w \\ \mathcal{C}^j\left(p, q|\theta(p, q)\right) \text{for } q - p < w \end{cases}$$

where $w > 0$ is a given window length. A windowed cost function estimates the parameters on a fixed window, rather then using the whole data. Note that if the length of a segment is less than w, the cost is the standard cost for this segment.

Unlike a generic cost function, a windowed cost function does not use all the data for parameter estimation. As a result, the cost function can be easily partitioned. Given any $q > p > w$ we have

$$\hat{\mathcal{C}}^j(t, t + q, w) = \mathcal{C}^j(t, t + p|\ \theta(t, t + w)) + \mathcal{C}^j(t + w, t + q|\ \theta(t, t + w)).$$

In other words, we can always split the windowed cost function of data $X^j_{t:q}$ into two terms, where the first term is independent of data after the split. Note, the first term on the right hand side is independent of data after $t + p$. Thus, after w points have been observed, the windowed cost of a segment can be partitioned into a left cost that does not change as new data is observed, and a right cost that does.

Restricting parameter estimation to a given window seems like a significant restriction, since it increases the variance for parameter estimators making it more difficult to detect changes. However, if this increase in variance is not too large, then the windowed cost functions may still be useful for changepoint estimation. Furthermore, we note that other authors within the literature use subsets of the data for changepoint detection (Eichinger and Kirch, 2018).

In order to explore the accuracy of the windowed approximation, we demonstrate that a classic result in univariate changepoint analysis from Yao, 1988 holds for windowed cost functions. Let X_i be a random variable and $\boldsymbol{\tau}$ be a vector of changepoint

locations with length m_0. We assume that

$$X_i \sim \mathcal{N}(\mu_s^0, \sigma^2) \text{ for } \tau_{s-1} < i \leq \tau_s,$$

for $s = 1, \ldots, m$ and, that we have some known upper bound on the number of changes m_U. For $1 \leq s \leq m_U$, we have a window length w_s such that $\tau_{i-1} + w_s < \tau_i \, \forall i = 1, \ldots, m_0$. Finally, let $\hat{m}$ be the solution of the following optimisation problem,

$$\hat{m} := \underset{1 \leq m \leq m_U}{\arg\min} \min_{|\boldsymbol{\tau}|=m} \sum_{s=1}^{m+1} \left[\bar{\mathcal{C}}(\tau_{s-1}, \tau_s | w_s)\right] + \beta m \tag{4.3.1}$$

$$\text{where } \bar{\mathcal{C}}(p, q) := \sum_{i=p+1}^{q} \left(\boldsymbol{X}_i - \theta(p, p+w)\right)^2, \; \theta(p, q) = \frac{1}{q-p} \sum_{i=p+1}^{q} \boldsymbol{X}_i.$$

The following result shows that $\hat{m}$ is a consistent estimator for the number of changes m.

Theorem 4.3.1. *Suppose we have $m_0 \leq m_U$ changepoints with mean levels $\mu_s^0 \neq \mu_{s+1}^0$ $(1 \leq s \leq m_0)$. Furthermore assume that $(\tau_s - \tau_{s-1})/n$ converge to q_s $(1 \leq s \leq m_0)$ and $w/n \to 1$ as $n \to \infty$ for some $0 < q_1 < \cdots < q_{m_0} < 1$. Then $\Pr(\hat{m} = m_0) \to 1$ as $n \to \infty$.*

Proof. Proof in Section 4.7. □

To demonstrate this result holds, we show that the error from using the windowed cost function $\bar{C}$ in (4.3.1) (as opposed to the true cost) is small with high probability. As a result, the windowed estimator for the number of changes is equal to the non windowed estimator with high probability. However the windowed cost function uses less data, and hence has larger variance. Due to this increased variance, in the finite sample setting there is a greater chance of both overfitting and missing changes. However as we shall see in the next section, the partitioning property can lead to significant improvements in the computational cost.

4.3.2 Multivariate Dynamic Program

We now show how windowed cost functions can be incorporated into the subset penalised cost function framework. An important consideration is how to choose the

window size w for the cost functions. In order to maximise the accuracy of our parameter estimates, we need to have the window as large as possible. Therefore for variable j, we choose $w_i^j = \tau_i - \tau_{i-1}$. If the next changepoint affects variable j then the cost of the segment will be the standard cost. However if variable j is not affected by the next changepoint then we will estimate the parameters θ_i^j using the window of data and in effect, introduce an artificial partition. Thus the optimal windowed subset segmentation is given by

$$\min_{\boldsymbol{\tau},\boldsymbol{\mathcal{S}}} \sum_{k=1}^{m+1} \left(\sum_{j=1}^{p} \Big[I\left(S_{k-1}^j = 1\right) \left(\mathcal{C}^j \left(\tau_{k-1}, \tau_k | \hat{\theta}_k^j \right) + \alpha \right) + I\left(S_{k-1}^j = 0\right) \left(\mathcal{C}^j \left(\tau_{k-1}, \tau_k | \hat{\theta}_{k-1}^j \right) \right) \Big] + \beta \right) \tag{4.3.2}$$

where

$$\hat{\theta}_k^j := \theta(\tau_{k-1}^j, \tau_k^j) I\left(S_k^j = 1\right) + \hat{\theta}_{k-1}^j I\left(S_k^j = 0\right).$$

We optimise (4.3.2) using an approximate recursion. We define $W(s)$ as the solution to the following recursion.

$$\begin{aligned} W(0) &:= -p\alpha - \beta, \\ W(s) &:= W(\tau(s)) + \hat{C}(\tau(s), s) + \beta \end{aligned}$$

where

$$\tau(s) := \underset{0 \le t < s}{\arg\min} \left\{ W(t) + \hat{C}(t, s) + \beta \right\}, \tag{4.3.3}$$

$$\begin{aligned} \hat{C}(t, s) := \sum_{j=1}^{p} & I(S^j(t, s) = 1) \left(\mathcal{C}^j \left(t, s | \theta(t, s) \right) + \alpha \right) \\ & + I(S^j(t, s) = 0) \left(\mathcal{C}^j \left(t, s | \hat{\theta}_t^j \right) \right), \end{aligned} \tag{4.3.4}$$

$$S^j(t, s) := I\left(\mathcal{C}^j \left(t, s | \theta(t, s) \right) + \alpha < \mathcal{C}^j \left(t, s | \hat{\theta}_t^j \right) \right), \tag{4.3.5}$$

$$\hat{\theta}_s^j := \theta(\tau(s), s) I\left(S^j(\tau(s), s) = 1\right) + \hat{\theta}_{\tau(s)}^j I\left(S^j(\tau(s), s) = 0\right). \tag{4.3.6}$$

The value $W(s)$ can be evaluated by solving (4.3.3) for $s = 1, \ldots, n$. The cost of solving this recursion is dependent on the cost of evaluating $\hat{C}(t, s)$, which is an $\mathcal{O}(K(n)p)$

calculation. Thus, the cost of solving the recursion for time s, is $\mathcal{O}(K(n)sp)$. Then the overall cost of finding $W(n)$ is $\mathcal{O}(K(n)pn^2)$.

There are aspects of this approach that are worth highlighting. Firstly the windowed cost function approach does not produce the same segmentation if the data is read backwards rather than forewards. While this is an undesirable property for a changepoint method, we note that if the approximation error is small then the difference between the segmentations should be marginal. Secondly, if two changes are close together, the approximation may break down as we will have a smaller window to evaluate the model parameters. Finally there are some limitations to the procedure for updating the segment parameters given in (4.3.6). Suppose we have three changepoints $\tau_k < \tau_{k+1} < \tau_{k+2}$ where τ_k affects variable j while the other two changepoints do not. Then our approach only uses the data between τ_k and τ_{k+1} to calculate the model parameters when evaluating $\hat{C}(t, s)$ for any $s, t > \tau_{k+2}$. If there is not much data between τ_k and τ_{k+1}, this approach will be inefficient and we may overfit changes as a result. Alternatively, we could update the model parameters and use the data between τ_k and τ_{k+2} to calculate the relevant model parameters. This approach should reduce the approximation error and may lead to a better segmentation. However there are other settings where this approach may be less effective.

The solution to the recursion above does not produce an exact minimizer of (4.3.2). However it is guaranteed to produce a solution at least as good, in terms of (4.2.2), as a comparable full multivariate segmentation. This result is stated formally below.

Theorem 4.3.2. *Let $\mathcal{C}(t, s) = \sum_{j=1}^{p} \mathcal{C}^j(s, t)$ and let $F(n)$ be the optimal solution to* (4.2.1) *for data of length n with penalty $p\alpha + \beta$. Similarly let $W(n)$ be defined as before. Then we have that*

$$W(n) \leq F(n).$$

Proof. Proof in Section 4.7. □

This result follows directly from the fact that this dynamic program contains the full set of possible fully multivariate results within its search space. Furthermore this

is a lower bound on the performance of the algorithm. In practice, as we demonstrate through simulations, this approach can produce significantly better solutions.

4.3.3 Pruning Step

We can solve the approximate recursion introduced above with a worst case computational complexity of $\mathcal{O}(pn^2)$. This can be improved further by reducing the search space in the optimisation problem (4.3.3). In particular, this optimisation problem is equivalent to identifying the optimal changepoint prior to s. Intuitively, if we have strong evidence that a change has occured at some time $t < s$, then it is unlikely that the optimal changepoint prior to s will occur before t. This is the intuition for pruning. Pruning is a technique used to speed up the computational cost of the dynamic programs used for detecting changepoints. The following theorem describes when a candidate prior change can be pruned.

Theorem 4.3.3. *Assume that there exists a constant K such that for all $t < s < T$,*

$$\mathcal{C}^j(t, s) + \mathcal{C}^j(s, T) + K \leq \mathcal{C}^j(t, T)$$

If the following inequality holds at a future time $T > s$

$$W(t) + \hat{C}(t, s) - p\alpha \geq W(s), \tag{4.3.7}$$

then t can never be the optimal last changepoint prior to T.

Proof. Proof in Section 4.7. □

We can explore the theoretical computational cost of SPOT using a similar framework to Killick, Fearnhead, et al., 2012. We restrict our attention to models where segment parameters are independent across segments and the cost function for a segment is negative the maximum log-likelihood values for the data in the segment. An underlying stochastic model for the data generating process is defined. The computational cost is the cost of analysing n data points generated by this process. Note that the dimension p is assumed to be fixed. Our result also assumes for $j = 1, \ldots, p$, that the parameters associated with a given segment are IID with density function

$\pi^j(\theta^j)$. Similarly for notational simplicity, we assume that, given the parameter θ^j, the data points within the segment are IID with density function $f^j(y|\theta)$. Finally we have that

$$\mathcal{C}^j(t,s) = -\max_{\theta^j} \sum_{i=t+1}^{s} \log f^j(X_i^j|\theta^j)$$

We also place assumptions on the changepoint locations, $\boldsymbol{\tau}$. For $s = 1, 2\ldots$ let $Q_s = \tau_s - \tau_{s-1}$. We assume the Q_s are IID copies of a random variable Q. Furthermore the Q_s are assumed to be independent of the parameters associated with a segment.

Theorem 4.3.4. *Define θ_* to be the value that maximises the expected log likelihood*

$$\boldsymbol{\theta_*} = \arg\max_{\boldsymbol{\theta}} \int\int f(\boldsymbol{X}|\boldsymbol{\theta}) f(\boldsymbol{X}|\boldsymbol{\theta}_0) dX \pi(\boldsymbol{\theta}_0) d\boldsymbol{\theta}_0.$$

Let $\boldsymbol{\theta}_i$ be the true parameter associated with the segment containing $\boldsymbol{X}_i$ and $\hat{\boldsymbol{\theta}}_n$ be the maximum likelihood estimate for $\boldsymbol{\theta}$ given data $\boldsymbol{X}_{1:n}$ and an assumption of a single segment,

$$\hat{\boldsymbol{\theta}}_n = \arg\max \sum_{i=1}^{n}\sum_{j=1}^{p} \log f^j(x_i^j|\theta^j) \text{ and } B_n = \sum_{i=1}^{n} \left[\log f(\boldsymbol{X}_i|\hat{\boldsymbol{\theta}}_n) - \log f(\boldsymbol{X}|\hat{\boldsymbol{\theta}}_*^j)\right]$$

Then if we have

(A1) $\mathbb{E}(B_n) = \mathcal{O}(n)$, $\mathbb{E}([B_n - \mathbb{E}(B_n)]^4) = \mathcal{O}(n^2)$ *(A2)* $\mathbb{E}([\log f(\boldsymbol{X}_i|\boldsymbol{\theta}_i) - \log f(\boldsymbol{X}|\boldsymbol{\theta}_*)]) < \infty$

(A3) $\mathbb{E}(\log f(\boldsymbol{X}_i|\boldsymbol{\theta}_i) - \log f(\boldsymbol{X}_i|\boldsymbol{\theta}_*)) > \dfrac{\beta + p\alpha}{\mathbb{E}(Q)}$ *(A4)* $\mathbb{E}(Q^4) < \infty$

where Q is the expected segment length, the expected CPU cost of SPOT for analysing n data points of fixed dimension p is bounded above by $L_p n$ for some constant $L_p < \infty$ dependent on p.

Proof. Proof in Section 4.7. □

Conditions (A1) and (A2) are weak technial conditions. Condition (A3) states that expected penalised likelihood value obtained with the true changepoint and parameter values with a fully multivariate penalty will be greater than the expected penalised cost given by fitting a single segment. Condition (A4) restricts the probability of observing very large segments. As a consequence the expected number of changepoints is an increasing linear function of n.

4.4 Simulations

We now explore the performance of our method through a range of simulations. We begin by defining our performance metrics. Firstly throughout we use $\boldsymbol{\tau} := \{\tau_1, \ldots, \tau_m\}$ and $\hat{\boldsymbol{\tau}} := \{\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{m}}\}$ to denote the set of true changepoints and the set of estimated changepoints respectively. A common approach for evaluating changepoint methods is to examine true and false discovery rates. We say that the changepoint estimate τ_i has been detected if

$$\min_{1 \leq j \leq \hat{m}} |\hat{\tau}_j - \tau_i| \leq h.$$

Throughout this section we set $h = 10$ although it should be noted that in reality the desired accuracy would be application specific and whilst the specific values vary with h the conclusions of the study do not. We denote the set of correctly estimated changes by $\boldsymbol{\tau}_c$. Then we define the true discovery rate (TDR) and false discovery rate (FDR) as follows,

$$TDR := \frac{|\boldsymbol{\tau}_c|}{|\boldsymbol{\tau}|}, \; FDR := \frac{|\hat{\boldsymbol{\tau}}| - |\boldsymbol{\tau}_c|}{|\hat{\boldsymbol{\tau}}|}.$$

The TDR and FDR describes how accurately a method locates multivariate changepoints. However we are also very interested in whether or not the methods return accurate subsets. Let $\boldsymbol{\tau}^k$ and $\hat{\boldsymbol{\tau}}^k$ denote the set of true and estimate changepoints that affect variable k respectively. Then for each k we have the corresponding true and false positive rates TDR^k and FDR^k. Then we define the Variable Average True Detection Rate (VATDR) and the Variable Average False Detection Rate (VAFDR) as

$$VATDR := \frac{1}{p}\sum_{k=1}^{p} TDR^k, \; VAFDR := \frac{1}{p}\sum_{k=1}^{p} FDR^k.$$

Intuitively if a method correctly estimates subsets then we would expect the VATDR to be close to one and the VAFDR to be close to zero.

An important concern is whether or not the segmentation allows us to accurately us to estimate the model parameters. Therefore we also report the Mean Square Error (MSE),

$$MSE := \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i\|_2^2.$$

Unless otherwise indicated, In order to satisfy the assumptions for a computational cost of $\mathcal{O}(n)$, the number of changes is proportional to the length of the data. For a given change, the probability of a variable being affected by a change is either $\{.2,\ .5,\ .8\}$, with probabilities $\{1/2, 3/8, 1/8\}$ respectively.

4.4.1 Optimality Gap

We begin by examining the optimality gap incurred from solving (4.2.2) via our approximate method SPOT, that is, the difference in penalised cost between the approximate solution and the exact solution. We compute exact solutions using the SMOP dynamic program. Due to the large computational cost of SMOP, we are limited in the size and range of datasets we can consider. We simulated 100 datasets of size $n = 100$ and $p = 3$. The data is normally distributed with two changes in mean. The magnitudes of the changes are uniformly distributed on $[0.8, 1.3]$. We use a cost function based on the log likelihood for normal data with known variance.

Results from this simulation are shown in Figure 4.4.1. For over sixty of the examples, we observe no optimality gap. The largest gap observed is less than five percent. For 89 examples, the gap is less than a single percent. The advantage of this optimality gap is a much lower computational cost. The average computation time for SPOT is .0995 seconds compared with almost four hours for SMOP.

In order to measure the changepoint accuracy we compared the performance of the algorithms in two scenarios. In both scenarios, we observe two changes in the mean of normally distributed data. The first scenario has changepoints at 33 and 66, while the second has changepoints at 20 and 85. For all the datasets the affected subsets are $S_1 = (0, 0, 1)$ and $S_2 = (0, 1, 1)$. The first set of datasets represent an ideal setting with a large minimum segment length, meaning the approximation error from our method should be small. The second set of datasets represent a more challenging case as the first and final segment lengths are short. Because of the short segment lengths we would expect the approximation from our method to be poorer.

The results of this simultation are presented in Figure 4.4.1 (b) and (c). In both examples, the exact approach correctly locates more changes. Furthermore, the vari-

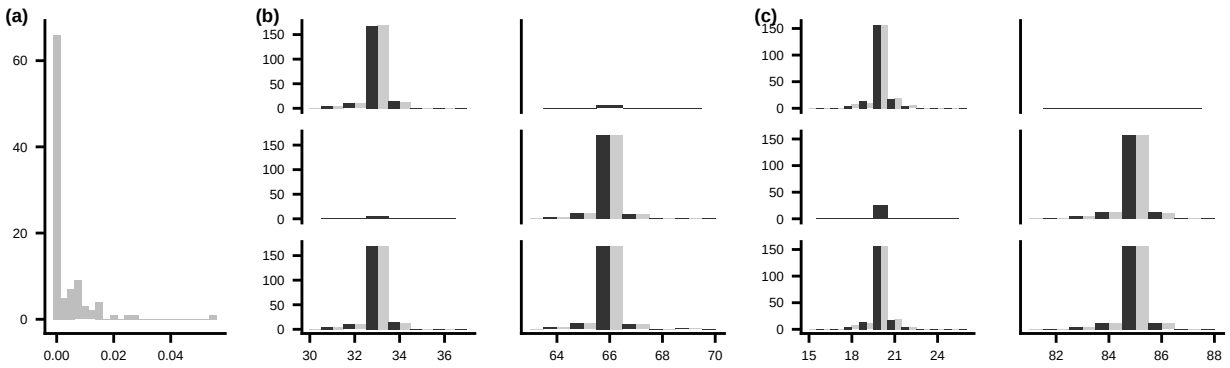

Figure 4.4.1: (a) Histogram of observed percentage optimality gap for SPOT; Histogram of estimated changepoints using SMOP (black) and SPOT (grey) over 100 repetitions. Example (b) has large segment lengths while (c) has short minimum segment lengths.

ance of the exact changepoint location estimates is lower than the variance of the approximate estimates. This is evidence that the approximation does in fact reduce statistical power. However the difference in accuracy between the methods is small.

4.4.2 Comparison with Fully Multivariate Model

Although simulation studies on small datasets have value, it is necessary to evaluate the performance of SPOT on larger datasets. Since an exact optimisation is computationally infeasible for larger datasets, we compare our approximate optimisation of (4.3.2) with an exact optimisation of (4.2.1) (i.e. assuming a change in all variables at each changepoint).

For this simulation, we consider datasets with length ranging from $n = 1000$ to $n = 100000$ and and dimension $p = 250, 1000, 2500$. For each n, p pair we simulated 100 datasets with changes in mean. The size of the changes are uniformly distributed between .5 and .8. Finally the minimum segment length is 10.

Some results from this simulation are shown in Figure 4.4.2. Boxplots were generated using the ggplot2 package (Wickham, 2016) using default settings. For each method and n, p pair, we plot a box consisting of a bold black line in the center, lower

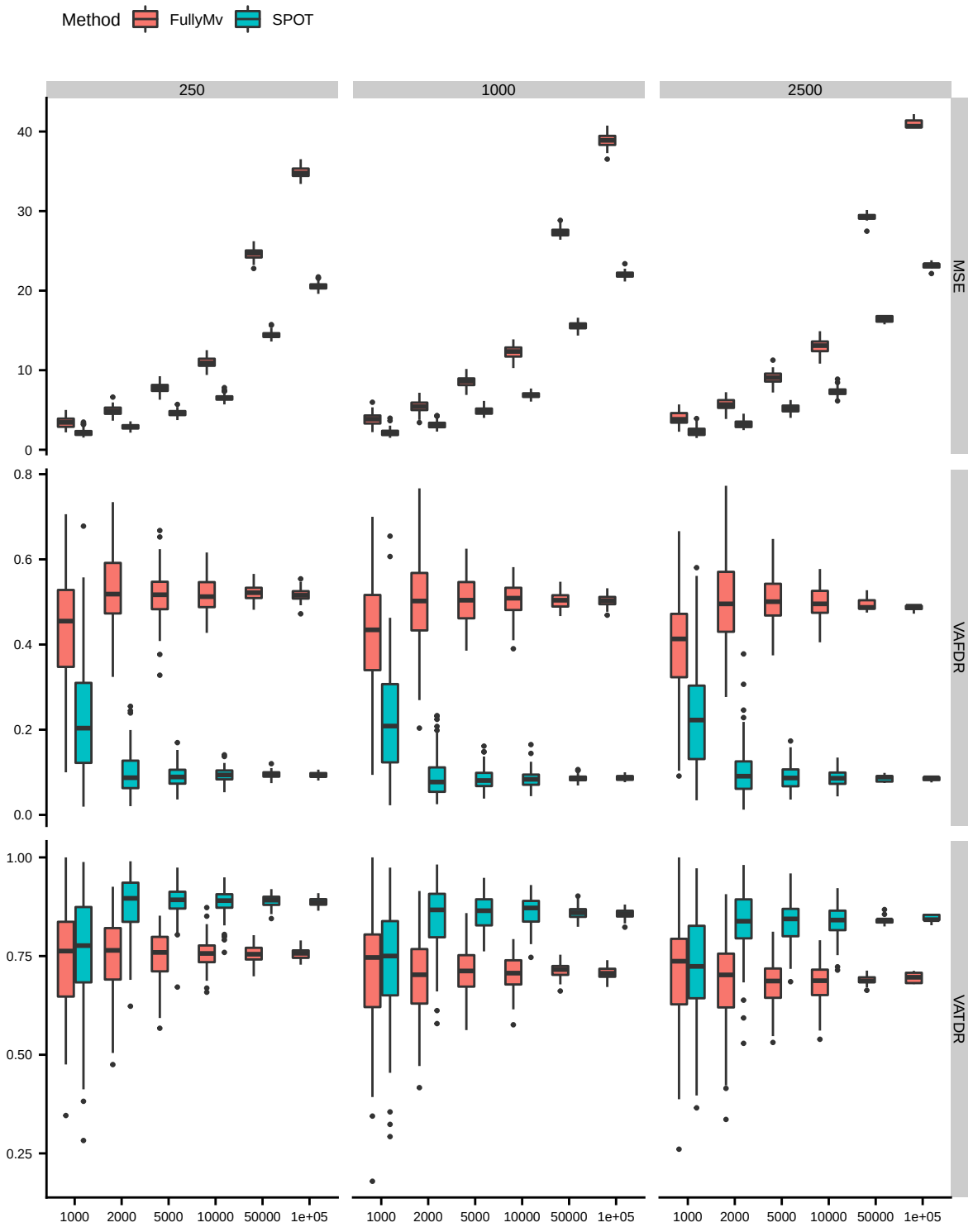

Figure 4.4.2: Boxplot comparison of SPOT and fully multivariate segmentations for different metrics over $n = \{1000, 2000, 5000, 10000, 50000, 100000\}$ and $p = \{250, 1000, 2500\}$.

and upper hinges (bottom and top lines of the box) and whiskers which extend out from the hinges. The bold black line gives the median. while the lower and upper hinges correspond to the 25th and 75th percentiles respectively. The lower/upper whisker extends from the 25/75th percentile to the smallest/largest value no further than 1.5 times the length of the box from the lower/upper hinge. Points beyond the whiskers are outliers. There are clear improvements in MSE, VATDR and VAFDR. SPOT achieves a lower MSE for every n, p pair. Furthermore the magnitude of this improvement increases with p and n. SPOT takes variables not changing into account, which means more data can be used to estimate the parameters producing more accurate estimates and a lower MSE.

We also observe improvements in the variable average detection rates. As the length of the data increases SPOT achieves a substantially lower VAFDR. We expect this metric to favour our approach, as the fully multivariate approach will falsely detect changes in variables that do not change. However there is also an improvement in the VATDR, implying that our approach accurately locates more changepoints for individual series. However, SPOT does miss some changes. The VATDR plateaus at around .8, implying that one fifth of the changepoints for each time series are missed on average.

4.4.3 Comparison with Other Methods

We now compare the performance of SPOT against two other state of the art algorithms, E-Divisive (T. Wang and Samworth, 2018) and Inspect (T. Wang and Samworth, 2018). We use the default values in the InspectChangepoint (T. Wang and Samworth, 2016) and ECP (James and Matteson, 2015) packages. When measuring computation time for Inspect, we do not count the time taken to identify the optimal penalty. This increases the computational time by a further order of magnitude. Both of these methods are fully multivariate and thus it is not relevant to report VATDR and VAFDR. We consider datasets of size $n = \{200, 400, 600, 800\}$ and $p = \{5, 10, 20\}$. For each parameter set, we generate 1000 normally distributed datasets with changes in mean. The size of the changes are uniformly distributed between 0.5 and 0.8, which

satisfies the minimum step size assumption utilised in (T. Wang and Samworth, 2018).

The results are shown in Figure 4.4.3, where again we use boxplots to compare the different methods. The E-Divisive method performs the worst across all metrics. In particular the TDR for the nonparametric method is significantly lower. This reflects the fact that a nonparametric approach has lower power. Furthermore the computational cost of E-Divisive is much larger. Our algorithm achieves significantly lower MSE than both Inspect and E.divisive. On the other hand, both methods achieve similar performance across the other metrics.

The Inspect method is designed primarily for high dimensional datasets. It is natural therefore to compare SPOT with Inspect in this setting. We ran both methods on datasets ranging in size $n = \{200, 500, 1000, 2000, 10000\}$ and $p = \{50, 250, 750\}$. The size of the changes are uniformly distributed between .5 and .8. Due to the computational cost we were not able to run E-Divisive. The results of this simulation are shown in Figure 4.5.1. We can see that Inspect performs much better in this setting. Inspect reports a higher TDR at the cost of a higher FDR. SPOT achieves a lower MSE, particularly for larger values of n, however this difference decreases as p grows. Finally, we note that SPOT has a much shorter runtime.

4.5 Applications

4.5.1 Genetics Data

We begin by considering the comparative genomic hybridisation (CGH) dataset from Bleakley and Vert, 2011 available in the ecp R package (James and Matteson, 2015). This dataset has been previously analysed in the literature (T. Wang and Samworth, 2018) and thus makes a useful comparison. CGH is a technique that detects abnormalities in chromosomal copy number by comparing the fluorescence intensity levels of DNA fragments from a test sample against a reference sample. This dataset examines the log intensity ratio measurements of 43 individuals at 2215 loci on their genome. Each individual has a bladder tumour. Copy number variations that are shared across multiple individuals are more likely to be related to the disease, thus it

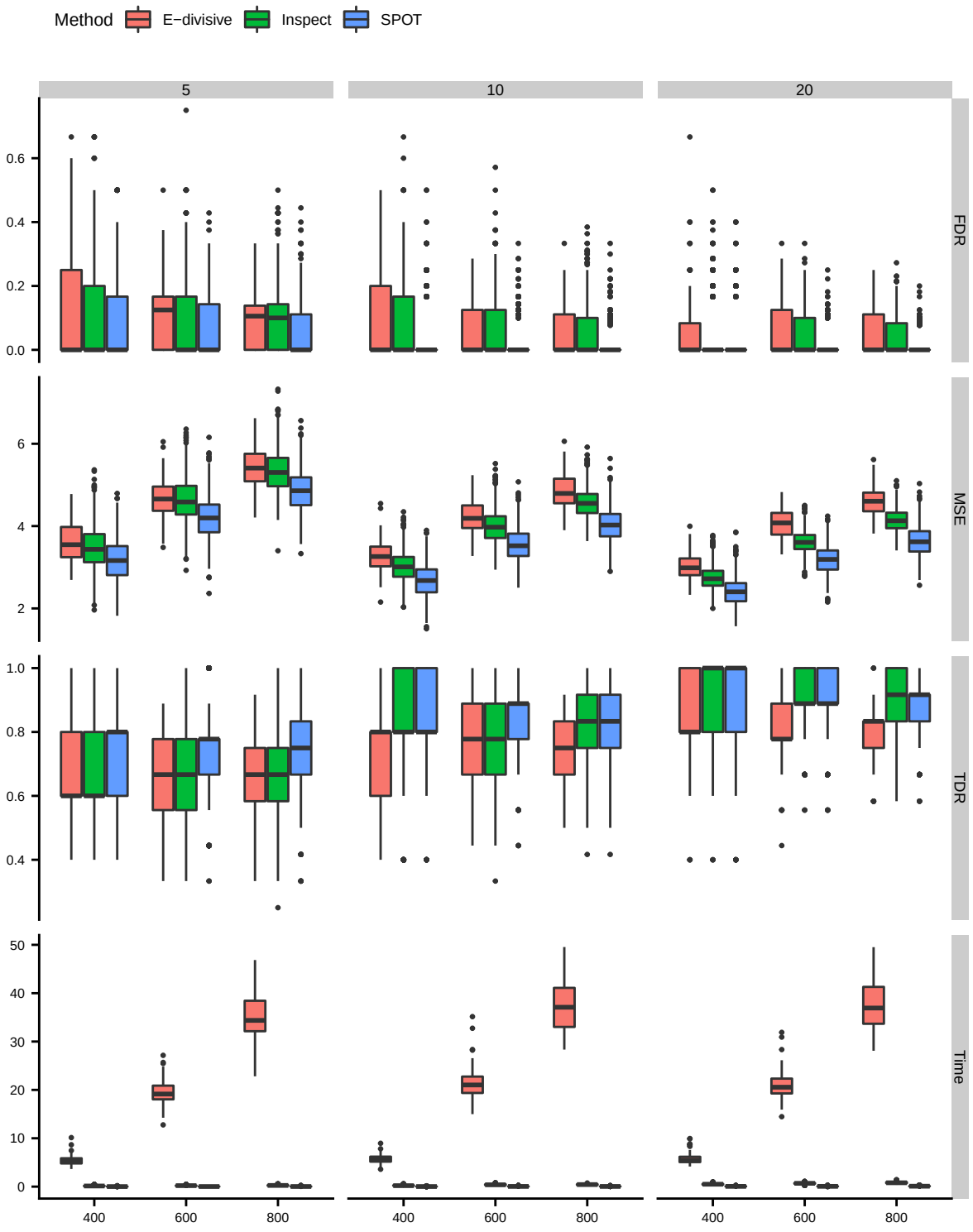

Figure 4.4.3: Boxplot comparison of SPOT, E-Divisive and Inspect for different metrics over $n = \{400, 600, 800\}$ and $p = \{5, 10, 20\}$

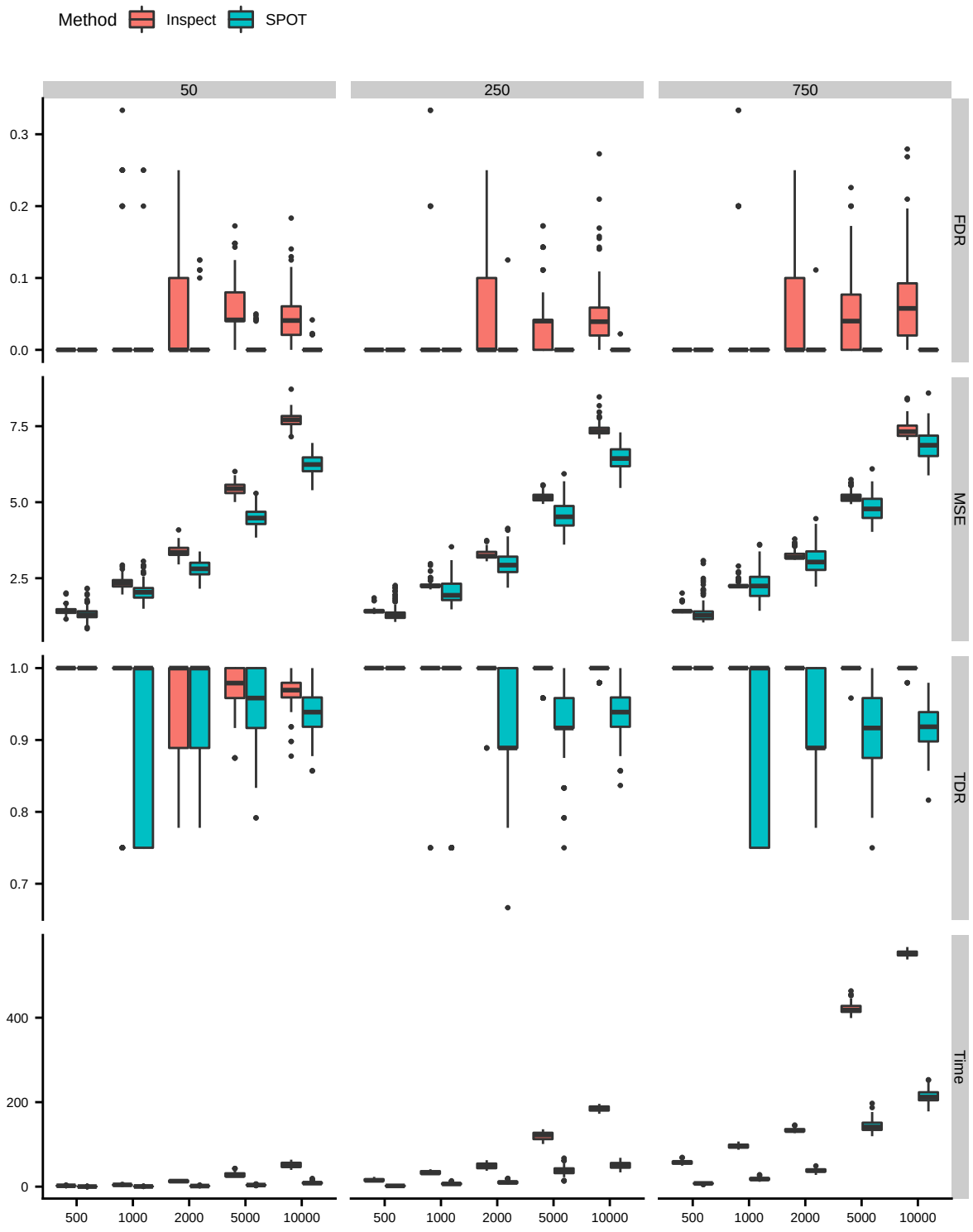

Figure 4.5.1: Boxplot comparison of SPOT and Inspect for different metrics over $N = \{500, 1000, 2000, 10000\}$ and $p = \{50, 250, 750\}$

is interesting to detect changepoints in this dataset, as well as identify which series the changepoints affect.

The observations for the first and third patients are shown in Figure 4.5.2. As in previous analysis, we assume the data is normally distributed and changes in copy number variation correspond to changes in mean. We note that the variance of the underlying dataset does not equal one, and thus standard penalties are inappropriate for the penalised cost framework. Therefore, we scale our penalties by the mean standard deviation for the series. This is equivalent to the penalties $\alpha = 2(.143)\log p$ and $\beta = 2(.143)\log n$, where $n = 2215$ and $p = 43$. Finally we removed some outliers from the dataset. Both the transformed and raw data are available on request.

We applied SPOT and Inspect to the full standardized data. Under default settings, as implemented in the ECP package, E.divisive locates 54 changepoints. In their paper T. Wang and Samworth, 2018 report the test statistic for each changepoint, only accepting changepoints whose test statistic is above a certain threshold. The default threshold for Inspect produces far too many changepoints as it is too low. Therefore the authors only include the thirty most significant changepoints which we repeat here.

The resulting segmentations for three individuals are shown below in Figure 4.5.2 Comparing the segmentations, we can immediately see the advantage of the subset approach. SPOT produces a segmentation with 67 changepoints, but still produces parsimonius segmentations for individual series. In particular we can see that the second series is only affected by two of these changes. On the other hand even restricting to the thirty changepoints with the largest test statistics, Inspect clearly still overfits on a series by series basis.

4.5.2 Syrian Civil War

The Violations Documentation Center in Syria (VDC) is a humanitarian organisation that records violence due to the Syrian Civil War (Violations Documentation Center in Syria, 2019). As part of this work, they have created an open source dataset of confirmed deaths. This dataset includes the name of the victim, the date and region

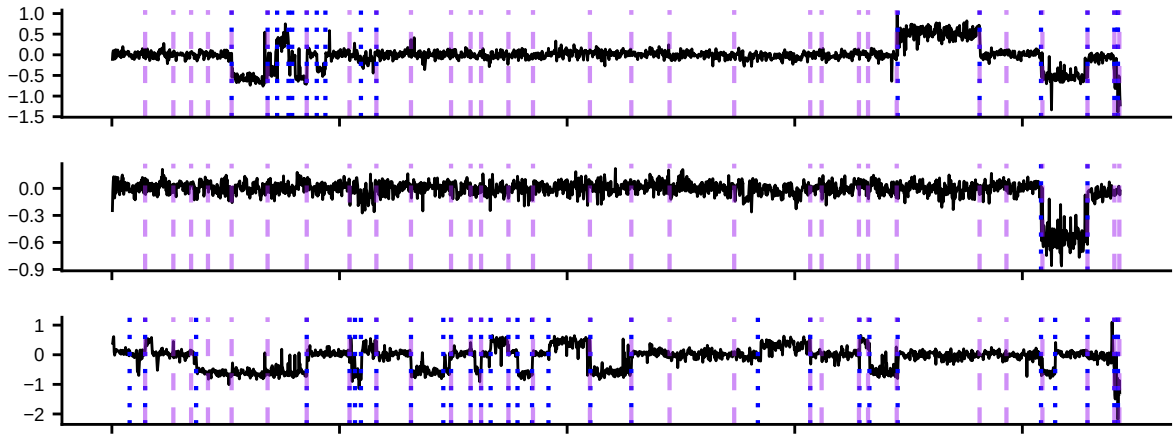

Figure 4.5.2: Segmentations for three individuals obtained from applying SPOT (dotted) and Inspect (dashed) to the normalised CGH dataset.

where they died as well as other information such as the organisation responsible. Using this data we can construct a time series of deaths per day for each of the 16 regions defined by the VDC. Note that as before we remove some outliers and both the original data and the standardized data are available on request.

The data for eight regions is shown in Figure 4.5.3. Note that these regions account for over ninety six percent of the deaths. Guha-Sapir et al., 2018 use this dataset to measure the number of deaths due to different weapons, as well as the number of deaths in different regions. This analysis is primarily focused on high level statistics. While this is useful, there is also a benefit in analysing the data at a more granular level. In particular we can see that the average number of deaths per day changes drastically and frequently over time. Identifying these changes is useful as it provides a simple, data driven way to understand the evolution of the war over time.

There are a number of challenges with modelling this data. The data is discrete and non negative. Therefore it is inappropriate to model it as Normal. Secondly there a large number of zeros. Therefore we use a cost function based on the likelihood for the Zero Inflated Poisson model i.e. if $X_{i,j}$ is the number of deaths in jth location on

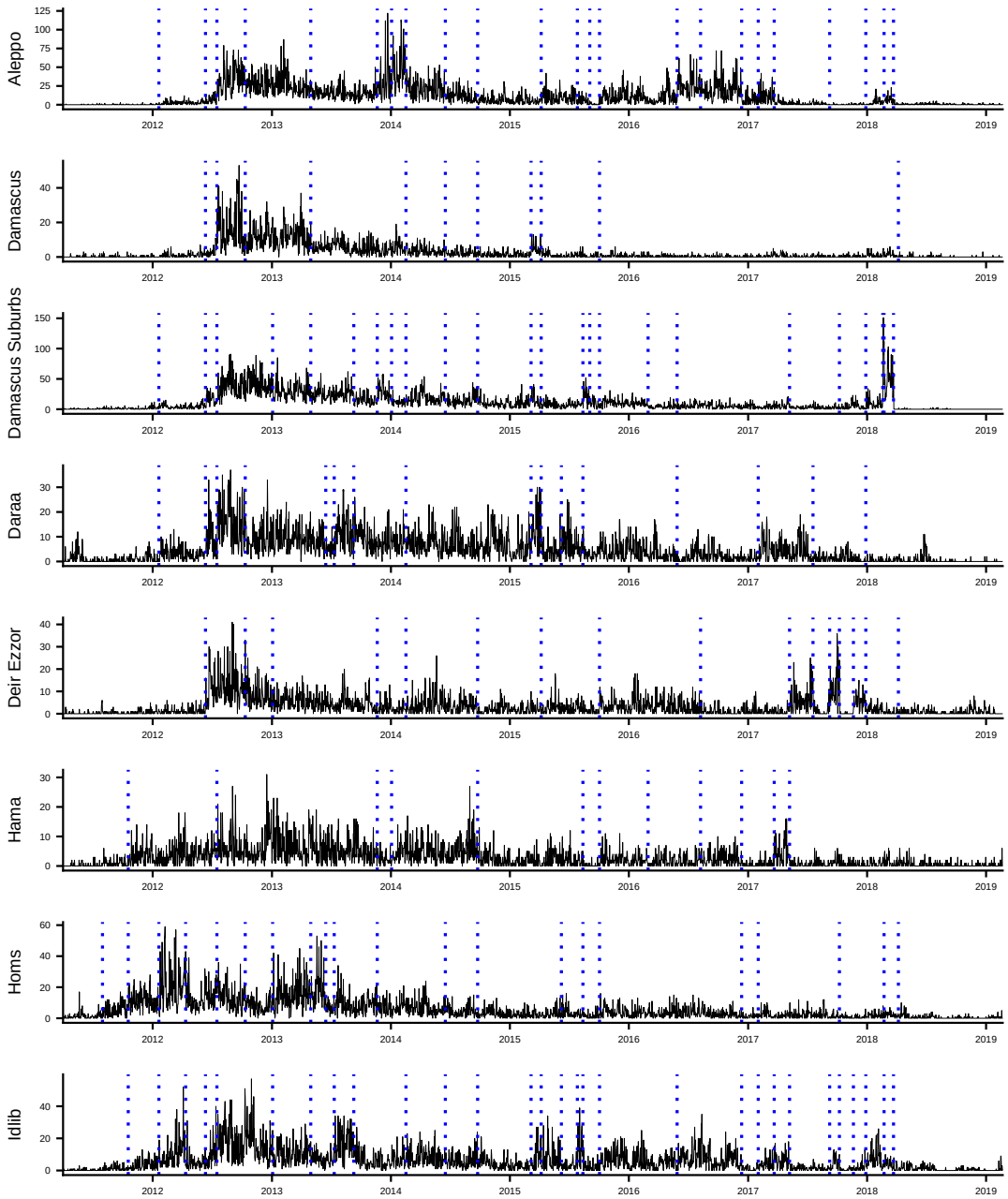

Figure 4.5.3: Deaths per day due to Syrian Civil War in eight regions as defined by the VDC. Changepoints as located by SPOT are indicated by the dotted lines.

the ith day and belongs to the kth segment we have that

$$Pr(X_{i,j} = 0) = \pi_k + (1 - \pi_k) \exp -\lambda_k \tag{4.5.1}$$

$$Pr(X_{i,j} = x_i) = (1 - \pi_k) \frac{\lambda_k^{x_i} \exp -\lambda_k}{x_i!}. \tag{4.5.2}$$

Note each segment features two parameters, the rate parameter λ_k and the inflation parameter π_k. We use the EM algorithm to fit these parameters, when evaluating the cost function. Finally the data is also overdispersed. Therefore, in order to apply SPOT, we scale our penalties by the square root of the average dispersion value. The resulting segmentation is shown in Figure 4.5.3

Using these penalties we locate forty changepoints. However, as with the previous example, the method still returns a parsimonius segmentation for each of the individual series. In order to validate the results, we showed them to an expert on the Syrian conflict. We identified two aspects of the segmentation which match expert understanding of how the war developed over time. Firstly, there is a period of over two years where the level of violence in Damascus is at a constant low level. Damascus, as the capital of Syria, is the center of power for the goverment and as goverment forces started winning the war, violence in the region substantially reduced. We note that a fully multivariate method would not be able to capture this pattern, as other locations do feature changes during this time. Secondly during the same period we see a number of dramatic changes in Deir Ezzor. The expert recognised this as a strategic pattern of violence in the region, where periods of intense fighting are punctuated with strategic calm. Thus, we argue that the segmentation given by SPOT does a good job of capturing the evolution of the war over time.

4.6 Conclusion

In this paper we have presented the SPOT method, a dynamic program for detecting changes in multivariate data where only a subset of the variables may change at any point. This approach has a number of positive qualities. The algorithm can be applied to a range of different types of datasets and distributions. It is computationally

efficient, with cost that under certain conditions is linear in the number of data points. Finally, despite being an approximation it is accurate, always recovering a better segmentation than the equivalent approach which assumes every variable changes. In simulations, SPOT outperforms other state of the art methods across a range of metrics.

4.7 Proof of Main Results

Proof of Theorem 4.3.1. Given m changepoints $(\tau_1, \ldots, \tau_m)$, we define the sum of squares,

$$S_n(\tau_1, \ldots, \tau_m) = \sum_{j=1}^{m_0+1} \sum_{i=\tau_{j-1}+1}^{\tau_j} \left\{X_i - \bar{X}(\tau_{j-1}, \tau_j)\right\}^2.$$

Similarly let $(\hat{\tau}_{1,m}, \ldots, \hat{\tau}_{m,m})$ denote the set of m changepoints that minimise $S_n(\tau_1, \ldots, \tau_m)$. Then the maximum likelihood estimate for σ^2 given m changepoints is given by

$$\hat{\sigma}_m^2 := \frac{S_n(\hat{\tau}_1, \ldots, \hat{\tau}_m)}{n}.$$

Yao, 1988 demonstrates that a consistent estimate of m_0 is given by the m that minimises

$$SC(m) = 2^{-1} n \log \hat{\sigma}_m^2 + m \log n$$

subject to $m \leq m_U$. We can define windowed equivalents of these estimators as follows,

$$\hat{\sigma}_{w,m}^2 := \frac{\hat{S}_n(\tau_1, \ldots, \tau_m)}{n} \text{ and } \hat{S}_n(\tau_1, \ldots, \tau_R) := \sum_{r=1}^{R+1} \sum_{i=\tau_{r-1}+1}^{\tau_r} \left\{X_i - \bar{X}(\tau_{r-1}, \tau_{r-1} + w)\right\}^2.$$

We are interested in how these estimators behave. Lemma B.1.4 in the appendix shows that the windowed variance estimator converges to the true variance, i.e. that $\hat{\sigma}_{w,m_0}^2 \to \sigma^2$ in probability. Then by Lemma B.1.5 we have that $\Pr(\hat{m} \geq m_o) \to 1$. In other words the estimator does not underestimate the number of changes asymptotically.

Now we only need to show that asymptotically the windowed estimator does not overestimate the number of changes. Let $Y_i = X_i - \theta_i$ for all $1 \leq i \leq n$. Then for any

$\epsilon > 0$ with probability approaching one,

$$\sum_{i=1}^{n} Y_i^2 > n(\sigma^2 - \epsilon).$$

Using Yao's bound B.1.3, for $m_0 < m < m_U$, for any $\epsilon > 0$, with probability approaching one,

$$\begin{aligned}
2\{\hat{SC}(m) - \hat{SC}(m_0)\} &\geq 2\{SC(m) - \hat{SC}(m_0)\} \\
&= n \log \hat{\sigma}_m^2 - n \log \hat{\sigma}_{w,m_0}^2 + 2(m - m_0) \log n \\
&\geq n \log \hat{\sigma}_m^2 - n \log \left(\sum_{i=1}^{n} Y_i^2 + \log n \right) + 2(m - m_0) \log n \\
&= n \log \left\{ 1 - \left(\frac{\sum_{i=1}^{n} Y_i^2 - n\hat{\sigma}_m^2 + \log n}{\sum_{i=1}^{n} Y_i^2 + \log n} \right) \right\} + 2(m - m_0) \log n \\
&\geq n \log \left\{ 1 - \left(\frac{\{\epsilon + (m - m_0 - 1)2(1 + \epsilon)\}\sigma^2 \log n + \log n}{n(\sigma^2 - \epsilon) + \log n} \right) \right\} + 2(m - m_0) \log n.
\end{aligned}$$

Using the fact that $\log 1 - x > (1 + \epsilon)(-x)$ for small $x > 0$ we have that the right hand side is greater than

$$-(1 + \epsilon) \frac{\{\epsilon + (m - m_0 - 1)2(1 + \epsilon)\}\sigma^2 \log n + \log n}{\sigma^2 - \epsilon + \frac{\log n}{n}} + 2(m - m_0) \log n$$

for large n. Since this is positive for small ϵ, we have that $\Pr(\hat{SC}(m) - \hat{SC}(m_0) > 0) \to 1$, completing the proof. □

Proof of Theorem 4.3.2. Firstly since $\hat{F}(n)$ is a penalised log likelihood with sub optimal parameters,

$$G(\hat{\tau}, \hat{\mathcal{S}}) \leq \hat{F}(n).$$

Then we only need to show that

$$\hat{F}(n) \leq G(\tau^{FMV}, \mathcal{S}^{FMV}) = F^{FMV}(n).$$

We proceed via strong induction. For data of length one there is only a single possible segmentation, hence the statement holds for $n = 1$. Assume that for all $k < n$ that $\hat{F}(k) \leq F^{FMV}(k)$.

Now by definition we have that

$$\hat{\mathcal{C}}(t, s) \leq \mathcal{C}(t, s).$$

Then

$$\hat{F}(n) = \min_{0 \leq k < n} \hat{F}(k) + \hat{\mathcal{C}}(k, n) + \beta \leq \min_{0 \leq k < n} F(k) + \mathcal{C}(k, n) + \beta = F^{FMV}(n)$$

This completes the proof. □

Proof of Theorem 4.3.3. Suppose that (4.3.7) holds for some $T > s$. Then we have that,

$$\hat{F}(t) + \hat{C}(t, s) - p\alpha \geq \hat{F}(s).$$

Adding $\hat{C}(s, T) + \beta$ to both sides of this equation gives,

$$\hat{F}(t) + \hat{C}(t, s) + \hat{C}(s, T) - p\alpha + \beta \geq \hat{F}(s) + \hat{C}(s, T) + \beta$$

However by Lemma B.1.6 in the appendix we have that

$$\hat{F}(t) + \hat{C}(t, T) + \beta \geq \hat{F}(t) + \hat{C}(t, s) + \hat{C}(s, T) - p\alpha + \beta \geq \hat{F}(s) + \hat{C}(s, T) + \beta$$

Hence t cannot be the most recent changepoint prior to T. □

Proof of Theorem 4.3.4. Let $G(s, t)$ denote the minimum value of the approximate cost function defined in the original text for data $\boldsymbol{X}_{s:t}$. By definition $G(s, t)$ is independent of data occuring before s and after t since it starts at s. Furthermore

$$\hat{F}(t) \leq \hat{F}(s) + G(s, t) + p\alpha,$$

since the right hand term is equivalent to having a fully multivariate change at time s and $\hat{F}(t) = G(0, t)$.

The pruning condition described in Theorem 4.3.4 is difficult to work with as it is dependent on the time t. Therefore we use a more stringent pruning condition that is independent of t. In particular we say that time point $t - k$ is pruned if

$$C(t - k, t) - 2p\alpha \geq G(t - k, t), \text{ where } C(s, t) = \sum_{j=1}^{p} \sum_{i=s+1}^{t} \mathcal{D}^j(s, t).$$

To see why this condition is more stringent note that if this condition holds we have that

$$\hat{F}(k) + \hat{C}(t - k) - p\alpha \geq \hat{F}(k) + C(t - k) - p\alpha \geq \hat{F}(k) + G(t - k, t) + p\alpha \geq \hat{F}(t).$$

For a positive integer $k \leq t$, let $I_{t,k}$ be an indicator of whether or not the observation k has not been pruned at time t. Then the overall computational cost of processing an observation at time $t+1$ is $\mathcal{O}(1 + \sum_{j=1}^{t} I_{t,k})$. Furthermore since the data-generating process is time invariant, and our pruning condition only depends on the data $\boldsymbol{X}_{(t-k+1):t}$ we have that $\mathbb{E}(I_{t,k}) = E_k$ independent of t. Hence the expected computational cost is bounded by nL_n where

$$L_n = 1 + \sum_{j=1}^{n-1} E_j.$$

By Lemma B.1.7 we have that

$$L := \lim_{n\to\infty} L_n < \infty$$

Then since L_n is an increasing sequence we have that the computational cost of analysing n points is bounded above by Ln. □

Chapter 5

Changepoint Analysis of Promotions

5.1 Introduction

The data science team at Tesco is modelling and forecasting data at ever greater granularity. As a result, business decisions can be made with greater accuracy and confidence. In this chapter, we consider the problem of modelling sales of individual products. A significant difficulty with modelling individual products is that the behaviour can change over time. For example, the sales of an individual product will be significantly impacted by whether or not there is a promotion going on. Therefore, in order to be able to accurately model and forecast at the individual product level, it is important to be able to take such changes into account.

In the previous two chapters, we studied a changepoint model that allows for some series to not be affected by the change. This feature is particularly valuable when analysing sales data of products that may be affected by promotions. For such data, we would expect that related or similar products may change at the same time. For example, changes in price due to promotions occur on a single date across multiple products. However, we would not expect sales of every product to be affected by a change. For example, a change in the price of an ice cream product should not have a large effect on sales of bread. This may also be different across time, sales of cranberry

sauce may be linked to a discount in the price of Turkey joints around Christmas time, but not at other times of the year. In all these examples, it is important to be able to detect changes across multiple series without fitting changes.

In this chapter, we explore and highlight some of the challenges associated with analysing real data via a changepoint analysis. In particular we use dual penalty changepoint model discussed in Chapters 3 and 4 to analyse the effect of promotions on sales data. We use the SPOT algorithm discussed in the previous chapter for this problem as it detects changes in multivariate time series, where changes occur in just a subset of the variables under observation. Crucially these subsets are not required to be specified in advance and may vary across changepoints. Although the SPOT algorithm is general we still need to preprocess the data to account for missing values, seasonality and trend which we assume do not change over time. After preprocessing, we use the SPOT algorithm to detect multivariate changes in mean, which are associated with increases (or decreases) of sales.

We are particularly interested in analysing how sales promotions for a single product can affect sales of similar products which are not affected by the promotion. Therefore we investigate whether there are interesting patterns in the resulting segmentation, such as whether some series frequently change together and whether or not these changes can be explained by promotions. Patterns that are not explained by promotions may indicate that there is an interesting relationship between the products, such as substitution effects. Finally we note that this application nicely illustrates some of the challenges involved with applying the SPOT method to a real world problem.

The structure of this chapter is as follows. In Section 5.2, we describe the data, highlighting any issues that make modelling the data more difficult. In particular, we highlight features of the data that we need to take account of via preprocessing. The preprocessing steps are then described in Section 5.3. In Section 5.4 we discuss the resulting multivariate segmentation, and explore whether or not the segmentation identifies any interesting patters. Finally we give concluding remarks in Section 5.5.

5.2 Data

We examine product level data over a 4 year period. In particular, we examine the daily sales data for products, aggregated across all stores of the same type in the estate. During this period, there are over 4 million different unique product codes. Rather than working with the entire dataset, we focus on products within a single product area which contains the majority of food products within Tesco; over 2600 products.

Within this group, there are products that are not sold throughtout the entire period. Therefore, we restrict our dataset to products with at least 1500 days of sales. From discussions with the Data Science Team, we learned that the Christmas and Easter periods are currently modelled separately to the rest of the year. Since these periods are very short, lasting no more than two weeks, it is very difficult to detect changes during these periods but they may drastically affect the detection of changes near to these periods. Furthermore, it would be very difficult to attribute changes to anything other than the holiday effect. Therefore, we do not include the Christmas and Easter periods in our analysis.

For each product and store type, we have the daily quantity of product sold at the store type and, the daily sales value of the product sold at the store type. Dividing these we can calculate the price of each product, as well as locate any changes in the price that would indicate a promotion. This data takes into account any multibuy offers as the price per unit would show a decrease. The unit price for a single product over this time period is shown in Figure 5.2.1. This shows clear separation of higher average price (which increases over time) and a lower sales prices. There may also be some smaller offers that results in more local changepoint effects.

The time series of total quantity, sold across all EXTRA stores for two further products from the cereal category, is shown in Figure 5.2.2. The data is clearly non stationary and the effects differ across products. Common to both series are trend, seasonality and weekday effects. Furthermore, there are periods where the sales jump, before returning to their previous levels shortly after. These jumps correspond with

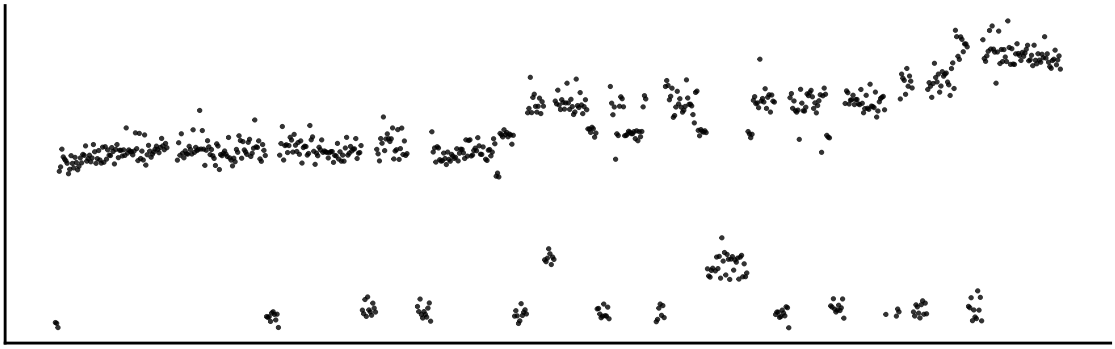

Figure 5.2.1: Daily price for a single product. Note that it has multiple price regimes that it switches between. The variance in the price is likely due to the effect of multibuys.

changes in the price of the product, indicating that a promotion was occuring which increased sales.

The trend, seasonality and weekday effects of the data make a changepoint analysis more difficult. In particular, we would expect that the trend, seasonality and weekday effects to be constant over time. This mixture of local effects (changepoints) and global effects (seasonality) complicates modelling and there are currently no changepoint methods that address this problem sufficiently. Therefore, rather than run a changepoint analysis on the raw data, which would necessitate a change in these global features, we first model the global features of the data and then run a changepoint analysis on the fitted residuals. This is further detailed in the following section.

5.3 Analysis

In this section we describe how we analyse the raw data. Our analysis has two components. Firstly, we estimate a set of promotion dates from the price data using a univariate changepoint analysis. These estimates become a ground truth for which we compare our later multivariate analysis. Note this comparison is only valid for series which are affected by the promotions. We then use a log linear model to remove the effects of seasonality, trend and day of the week. To capture the weekday

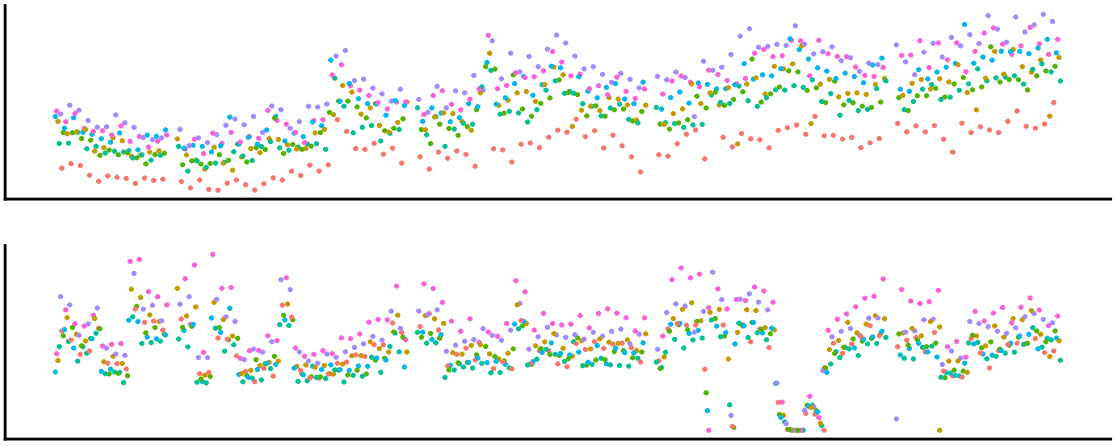

Figure 5.2.2: Daily quantity sold for two products across all EXTRA stores. The colour indicates the day of the week. Note that in both series there is a persistent day of the week effect, while in the first we can also observe trend and seasonality. Furthermore, the second series has prominent jumps that correspond with changes in the price.

effect, we include a categorical weekday variable in the design matrix. The assumed linear trend is captured via the difference between the start date and the date of an observation. For seasonality we model a yearly frequency using sine and cosine waves as the weekday effect is captured by the weekday dummy variables. We chose this approach due to the fact that we only observe 4 years worth of data, and thus a more complex model may overfit the data and obscure changes. However we accept that other approaches for modelling seasonality may be appropriate and even preferable.

One difficulty in fitting this model is that changes in level due to promotions will worsen the model fit and produce less accurate parameter estimates. This will increase the variance of the residuals making it more difficult to locate changes. Therefore, we also include the univariate promotion segmentation from the first step as a feature in our model, allowing us to model the impact of discounts separately.

Examining the data in Figure 5.2.2, we can see that the variance of the series changes over time. In particular, the variance appears proportional to the average sales. This is unsurprising, since we are working with count data. Therefore we

apply a log transform to the data to standardise the data. Again we note that other transforms, such as the Box-Cox or Anscombe transforms, may be appropriate. The full model can be expressed as

$$\log(\text{Quantity}) = \text{Trend+Seasonality+Weekday Effect+Discount Effect+Full Residuals}$$

or more mathematically as

$$\begin{aligned}\log(y_{t,j}) = \beta_{0,j} + \beta_{1,j}t + \beta_{2,j}\sin\frac{t}{365} + \beta_{3,j}\cos\frac{t}{365} \\ + \sum_{k=1}^{6}\beta_{3+k,j}I(\text{day}(t) == k) + \beta_{10}\delta_{t,j} + \epsilon_{t,j} \\ \text{where } \epsilon_{t,j} \sim \mathcal{N}(0, \sigma_j^2)\end{aligned}$$

where $Y_{t,j}$ is the quantity sold for product j at time t, I is an indicator function, $\delta_{t,j}$ is an indicator variable equal to one if there is a discount in variable j at time t.

We are primarily interested in detecting changes in products that cannot be explained by promotions. However these changes are likely to be caused by promotions in similar products and occur at the same time. Removing the effect of these promotions makes it more difficult to locate changes in products for which the price does not change but sales are affected by the promotion. Thus applying a changepoint analysis directly to the full residuals of this dataset is inefficient. Therefore we fit the changepoint analysis to the sum of the fitted residuals and the discount effect i.e.

$$\text{Partial Residuals} = \text{Discount Effect} + \text{Full Residuals}$$

or more mathematically as

$$p_{t,j} = \beta_{10}\delta_{t,j} + \epsilon_{t,j}.$$

We handle the issue of missing data by sampling from the above model. Whether or not a discount is occuring is inferred from the previous time point, as discounts last longer than a single day. Then a residual component is sampled from a normal distribution with mean zero and variance equal to the variance of the residuals for the series. We do not take dependence between the residuals into account for this example, however it may be possible to improve the analysis by doing so.

This approach gives us a full set of partial residuals, which we model in the following section using the multivariate changepoint techniques discussed in the previous chapters. In particular, we will apply the dual penalty approach, discussed in Chapter 3, to detect subset multivariate changepoints. Due to the size of the data, we will apply the approximate method, SPOT, to solve the resulting optimisation problem.

5.4 Results

The results presented here apply the methods in the previous sections to the data described in Section 5.2. We begin by demonstrating that we can estimate the dates of promotions from the price data for each series, using a changepoint analysis. For each time series of prices we fit a univariate changepoint analysis. Recall that we do not use a multivariate approach as the signal to noise ratio for these series is very high, as seen in Figure 5.2.1. For each series we look to identify changes in mean using the PELT algorithm. We use a minimum segment length of 15 days, as we do not expect the effect of promotions to be shorter than this. Finally due to the small variance we do not use standard penalties. Instead the β penalty is set to 0.2. Note this procedure is not optimal and it would be preferable to scale the data so that it has variance one. This is difficult in this situation, due to the fact that the within segment variance of the price for some series is zero.

The result of this analysis for a single product is shown in Figure 5.4.1. We can see that the changepoint analysis picks up the majority of large shifts in the price. However, we note that more subtle shifts in price may be missed. We argue that this is acceptable, since small shifts in price are less likely to correspond with promotions.

Before studying the multivariate changepoint analysis, it is important to consider whether the model described in the previous section fits the data well. The model fit for four series from the Cereal category is shown in Figure 5.4.2. We can see that the trend and seasonality effects have been estimated well. However the model shows some bias. We can see that sales for Sundays (orange) tend to be overestimated. Furthermore, the model for the second series, overestimates the quantity sold during the

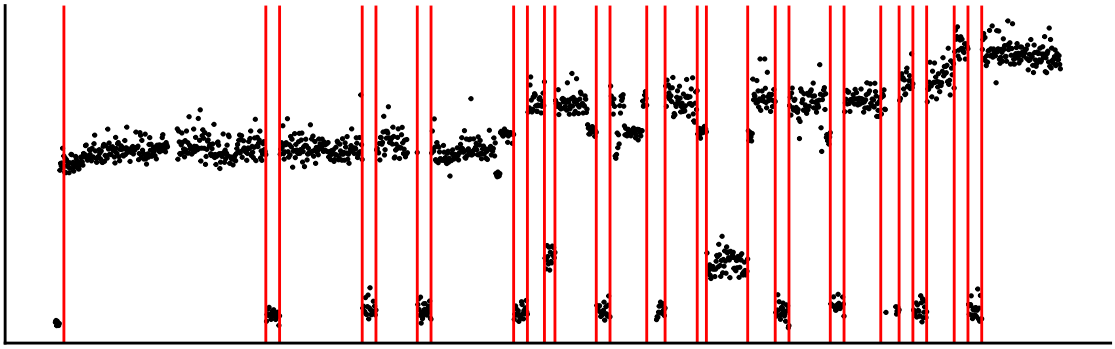

Figure 5.4.1: Daily price for a single product with fitted changepoints. We can see that the model picks up large changes in price that likely correspond with promotions. Furthermore, the model does not seem to overfit changes i.e. there are not many changes that do not seem to correspond with a promotion. Note however that more subtle changes are missed by the algorithm (such as at the start of the data due to the minimum segment length).

first year. These modelling issues mean that we do not have an ideal setting for applying a multivariate changepoint analysis. However, this is a significant improvement over the raw data and, the residuals should still give a good segmentation.

The full residuals for these series are shown in Figure 5.4.3. Examining these plots we can see that the full residuals are not always stationary. Furthermore, we can see that the orange points are more likely to be lower than the points for other days, reflecting the bias seen in Figure 5.4.2.

Finally, we also observe some large outliers. For example in series 2 we observe a very small value near the middle of the series which significantly distorts the view of the series. These very large outliers may introduce false changes in the analysis. Therefore, before applying the changepoint analysis, we first remove these outliers.

We cannot assume the full residuals are stationary, since there may be changepoints in the data that are not due to discount effects. Therefore, we utilise windowed median and maximum absolute deviation (mad) estimators, to get the mean and variance at a point in time. We say a point is an outlier if it is 3.5 (mad) standard deviations or more away from the median. The residuals for the same series after

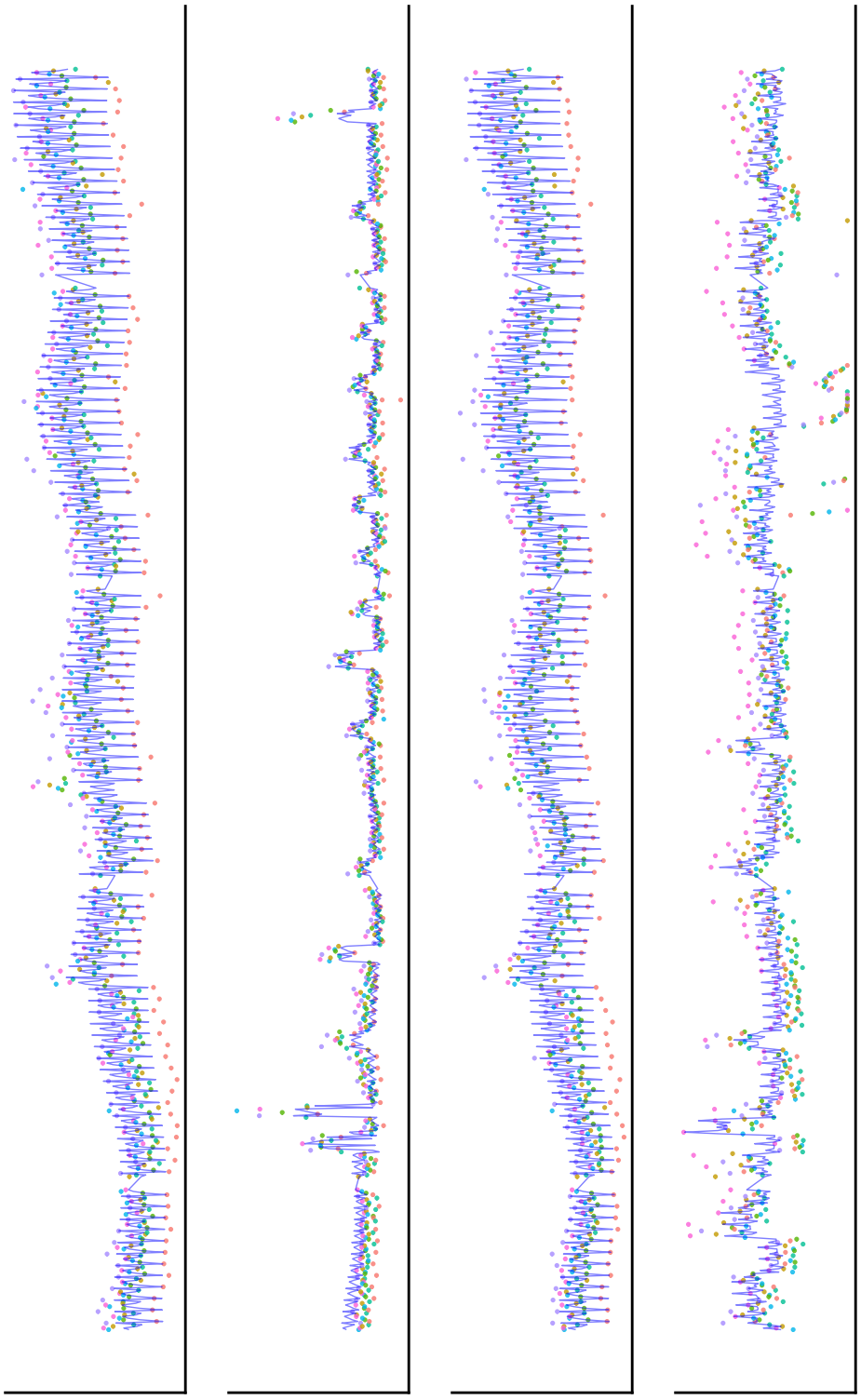

Figure 5.4.2: Quantity sold per day for four products. Colour indicates day. Blue line the denotes the full fitted model including the discount effect.

removing outliers are shown in Figure 5.4.4.

Finally, we note that there are some days where sales of products are unexpectedly zero. In order to apply our multivariate changepoint analysis we require all the series to have the same length. We treat this as a missing data problem. Therefore we sample from our full residual model, in order to fill in values for missing dates.

We apply our changepoint analysis to the partial residuals (i.e. the sum of the residuals and the discount effect). In this example, we just consider products in the junior area "Cereal", across four different store types. This gives $p = 360$ series of length $n = 1478$. We apply the SPOT algorithm using a cost function that detects changes in mean and variance of normally distributed data. We set $\beta = 4 \log n$ and $\alpha = 4 \log p$, the default penalties for this approach.

By way of example, consider the four series in Figure 5.4.4, we depict the full multivariate changepoint analysis of these in Figure 5.4.6. The changepoints detected by the SPOT algorithm are denoted by the dotted lines while the dashed lines indicate changes in price as detected in the univariate analysis, which indicate promotions. The second and fourth products feature a number of discount periods which are all located by SPOT. The method also detects changes which are not explained by discounts. The first and third products feature very few price changes. However, the SPOT method still locates a large number of changes. Some of these changes can be attributed to structure in the data not captured by the linear model. For example, the first and third series have clear downward trends which the model reports as changes mean. We note that there is a positive correlation between these two products. The multivariate changepoint analysis we employ assumes independence between series. As a result, we are more likely to report false changes due to the dependence.

The changes in price provide a ground truth for when promotions occur. We are interested in understanding how the changes reported by SPOT and the ground truth differ. We begin by considering how the model performs on series with a large number of changes in price. In particular, we focus on series with between 30 and 40 changes in price over the period. The changepoint analysis results for five series is shown in Figure 5.4.7. In order to compare the results from the price analysis, to those from

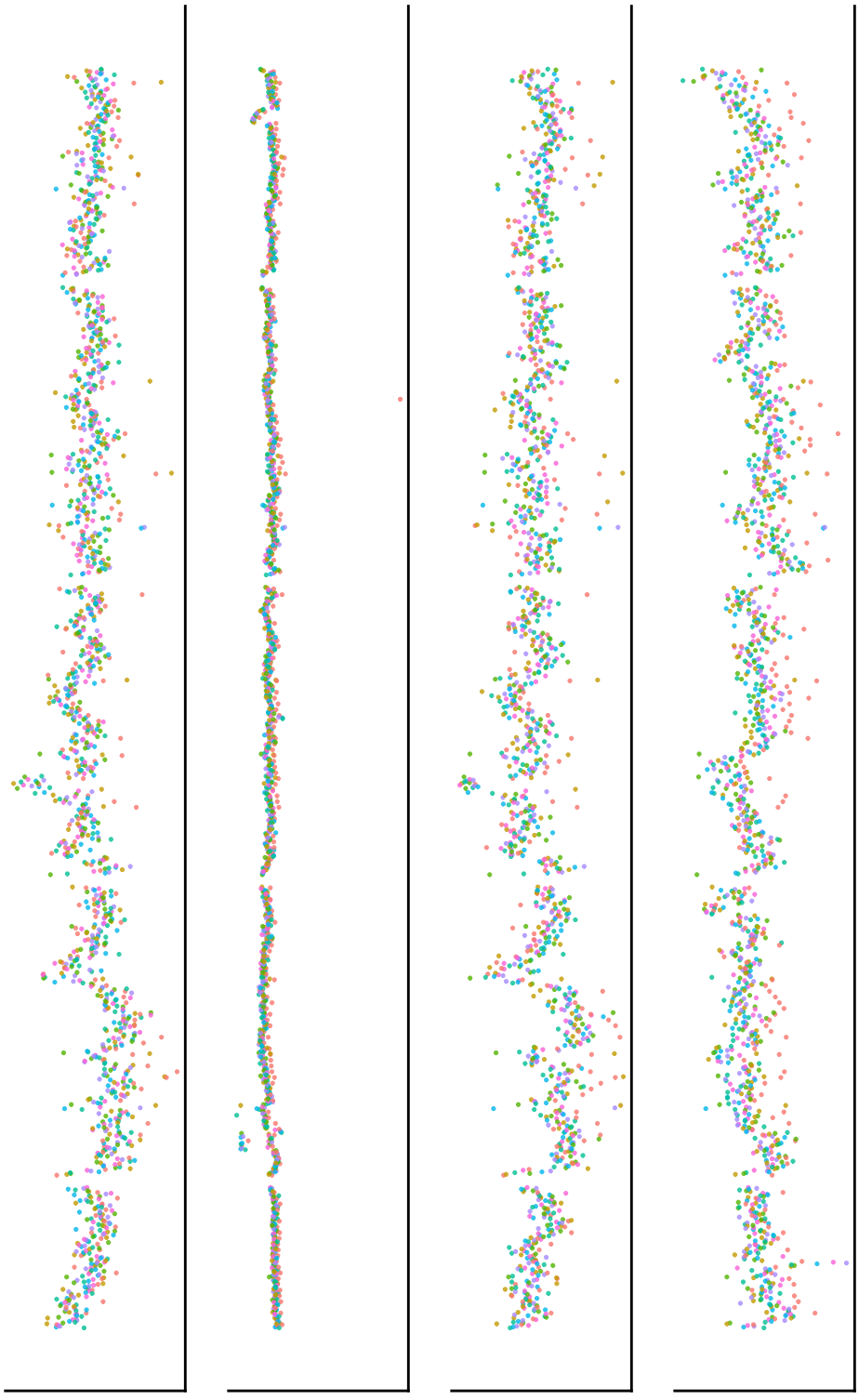

Figure 5.4.3: Fitted residuals from Figure 5.4.2. Note the weekday effect has been removed. However there are large outliers. Furthermore, the series appear to be non stationary at certain times (for example the start of the first and third series appears to have either a trend or a change in mean).

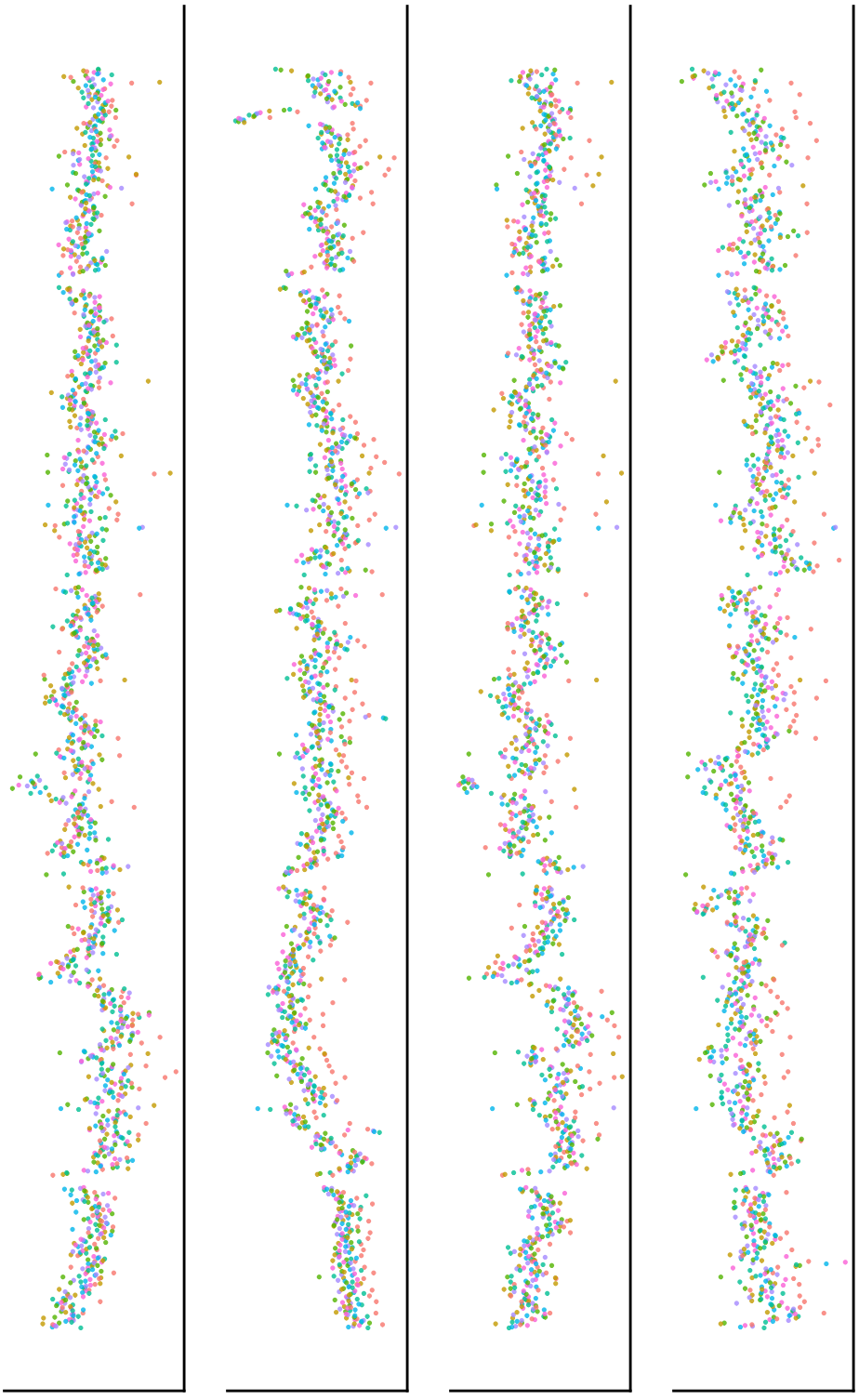

Figure 5.4.4: Same plot as Figure 5.4.3 with outliers removed. We can see that the second series is also non stationary at times.

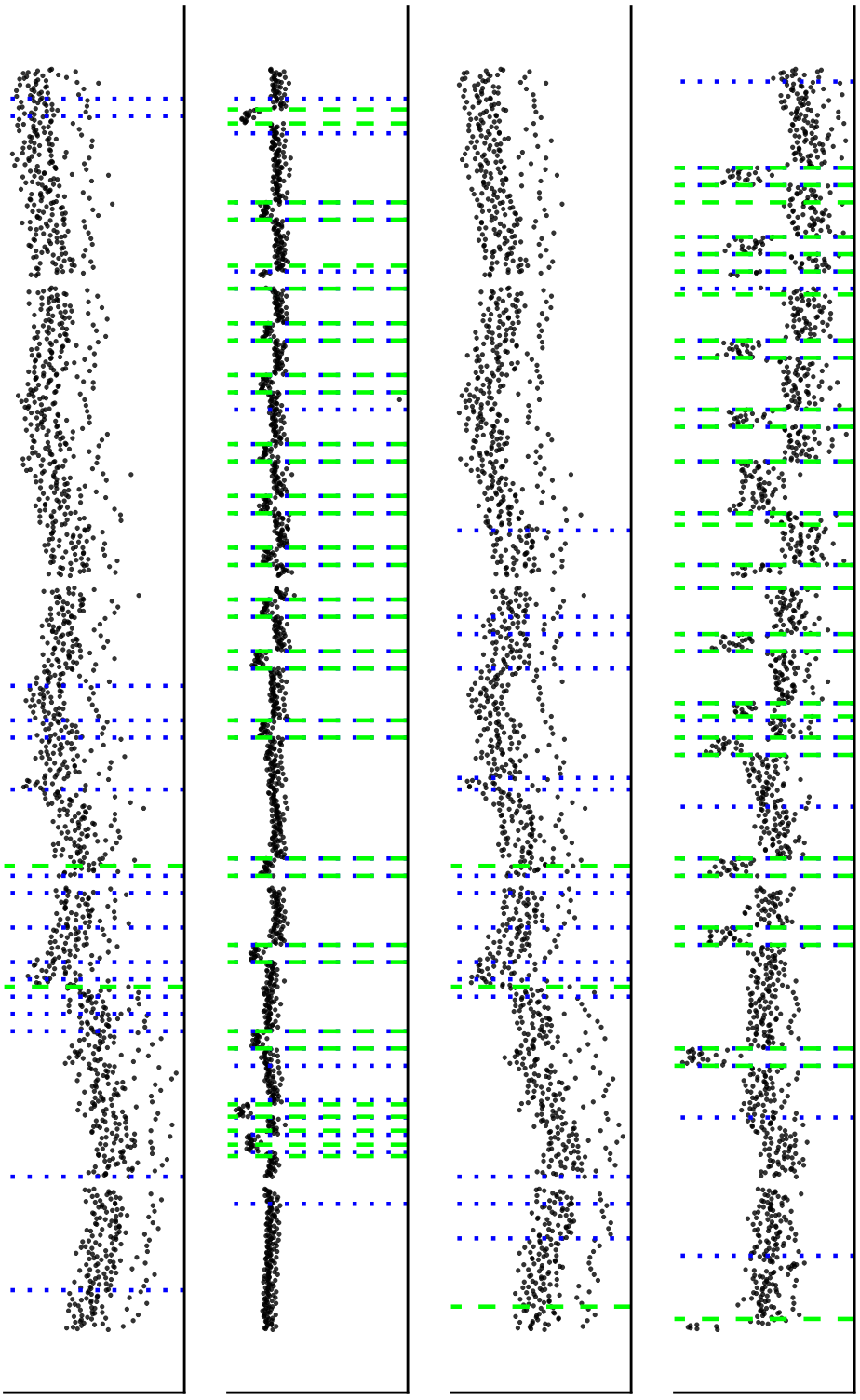

Figure 5.4.5: Changes reported by SPOT (dotted lines) and by price changes (dashed lines) for the raw data.

SPOT, we partition the changes. We say that a change in price is predicted by SPOT, if SPOT reports a change within three days of the discount. Changes in price which have been detected by SPOT are shown in green (dashed lines), while changes in price which have been missed are in red (solid lines). Finally, we also report changes detected by SPOT which do not correspond to any change in price. These are denoted by the purple (dotted lines).

Examining the results, we can see that SPOT detects most of the changes in price. However, there are instances where the method struggles. In particular, if changes in price occur close to each other in the same series, the method is less effective especially with the minimum segment length of 15 days. This is a weakness of the SPOT method discussed in the previous chapter. This can be seen in the first few red lines in series 1,2 and 5. SPOT often returns many more changes than there are changes in price. There are multiple possible reasons for this. Firstly, the algorithm may be responding to signals not fully captured by the model. For example, the first two series have a slight trend which the model reports as a change in mean. Secondly there may be inaccuracies due to the approximation used by SPOT. For example, SPOT locates many changes in the third series which do not correspond to discounts. We would expect some of these to be due to overfitting by SPOT. Finally, some of these changes are accurate representations of the data. For example, there are changes in the fourth series, which do not correspond to changes in price, but do correspond with clear changes in the quantity sold. Given the abruptness and scale of these changes, it seems likely that this product is responding to changes in the price of similar products or some external marketing by the individual manufacturers.

We also consider series that have relatively few changes in price. SPOT performs worse for these series. We report a large number of missed changes and few detected changes. The method appears to estimate the changepoint location with less accuracy. For example in the fifth series, we can see that SPOT misspecifies the location of a large change. This is due to the fact that SPOT tries to group changes together and a large number of series have a change near that time. We could attempt to control this by varying the penalty values, moving away from the defaults. While this is a

clear issue, we note that changes in price for these series are less reliable indicators of large changes in quantity sold. For example, the second and third series do not seem to be affected by changes in price. These may be series whose price changes do not affect sales.

In Figure 5.4.8, we identified a series with multiple abrupt, large changes which were unexplained by changes in price, which we shall refer to as the target product. We would like to investigate whether or not there is any correlation between when these changes occur and when explained changes occur for a different product. If this correlation were strong, then we would say there is a relationship between the products, such as competition between the products. We now explore whether this approach works in practice. We separate the changes detected by SPOT, into changes that are predicted by price and those that are not. Then for each series, we count the number of predicted changes which co-occur with unpredicted changes in the series. If this number is high, then the products should be related and vice versa. We then sort the products by this number.

The target product is an affordable health cereal and thus we would expect changes in other health cereals to explain the changes. In fact, we do observe a strong correlation between the target product and a number of other affordable health cereals. Upon showing these results to members of the Data Science team, we were informed that they would expect there to be relationships between the target product and the identified products. We argue that this is evidence the method is finding actual patterns between the products. However we note that other variants of the target product are less predictive under this measure which is surprising, and this is potentially evidence that the method is not accurately identifying relationships which one would expect. In order to further examine the relationship between the target product and identified products, we also plot the top four products and the target product in Figure 5.4.9. We can see that these products are so predictive because they have discount periods so frequently, thus the comparison is not necessarily appropriate. Further work could consider whether a different measure could more accurately identify related series from the changepoint analysis.

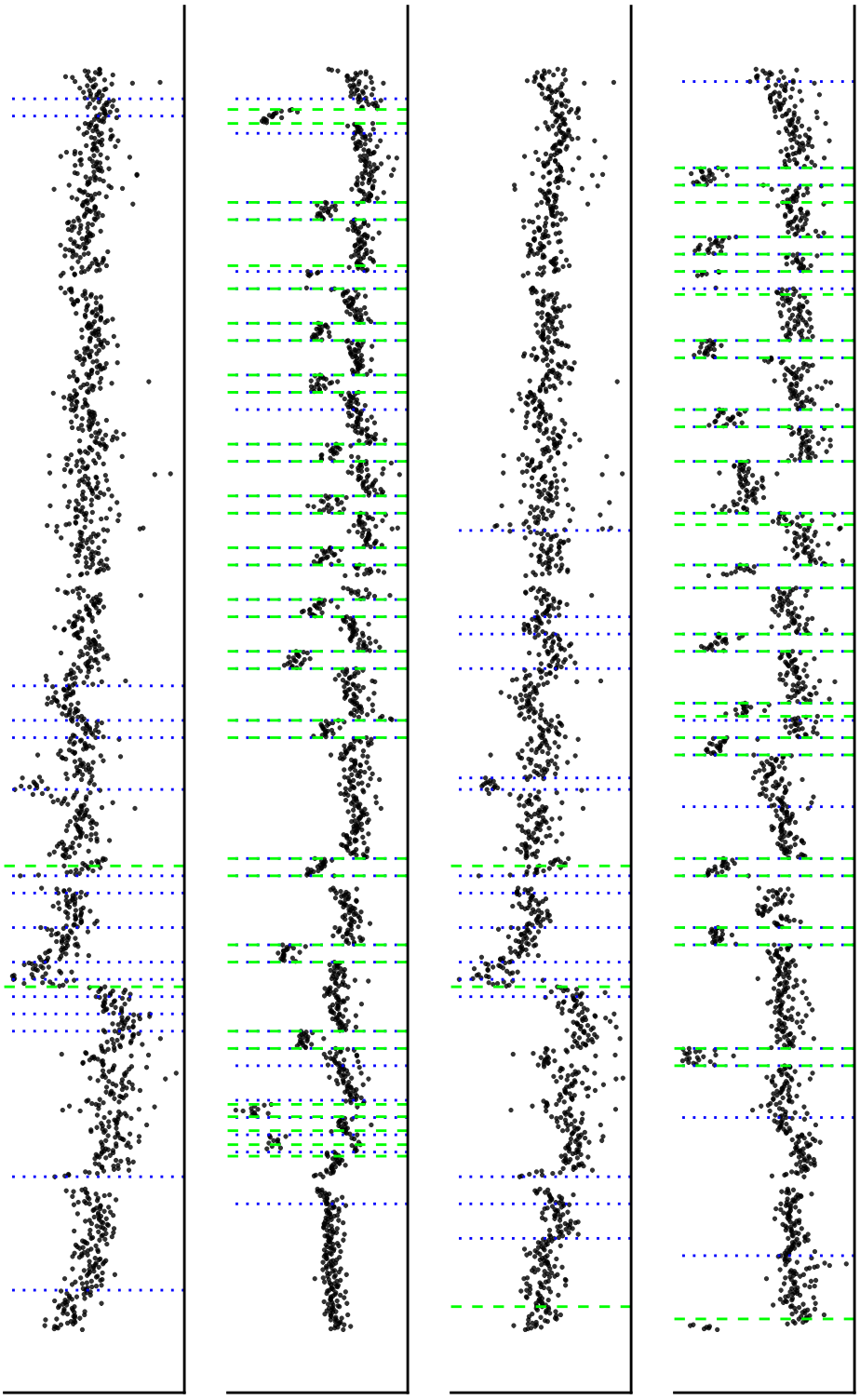

Figure 5.4.6: Changes reported by SPOT (dotted lines) and by price changes (dashed lines).

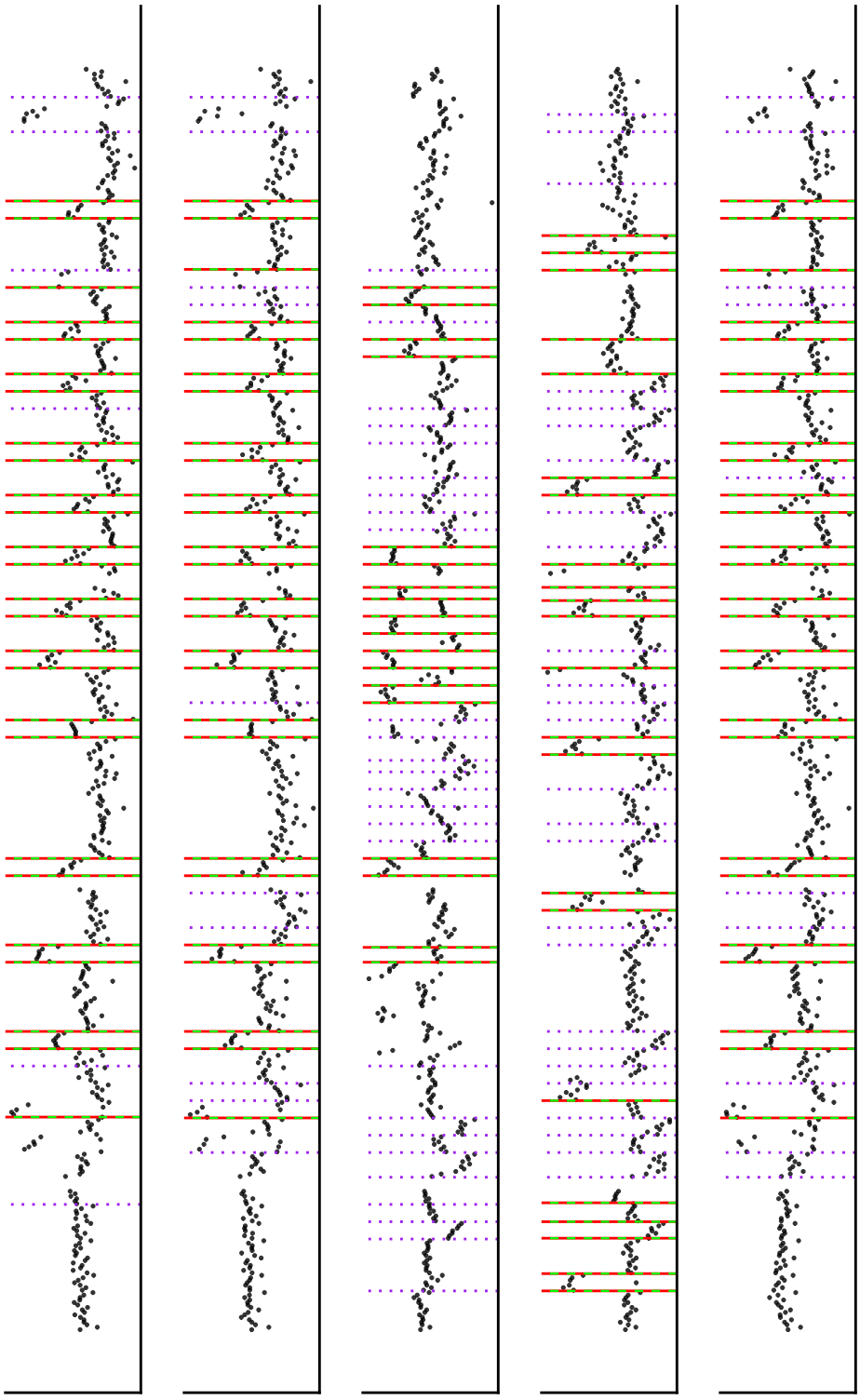

Figure 5.4.7: Changes reported by SPOT and by price changes for series with a large amount of discount periods. Dashed lines indicate change in price has been detected by SPOT, solid lines indicate change in price has been missed by SPOT while dotted lines indicate that SPOT has detected a change not related to a change in price.

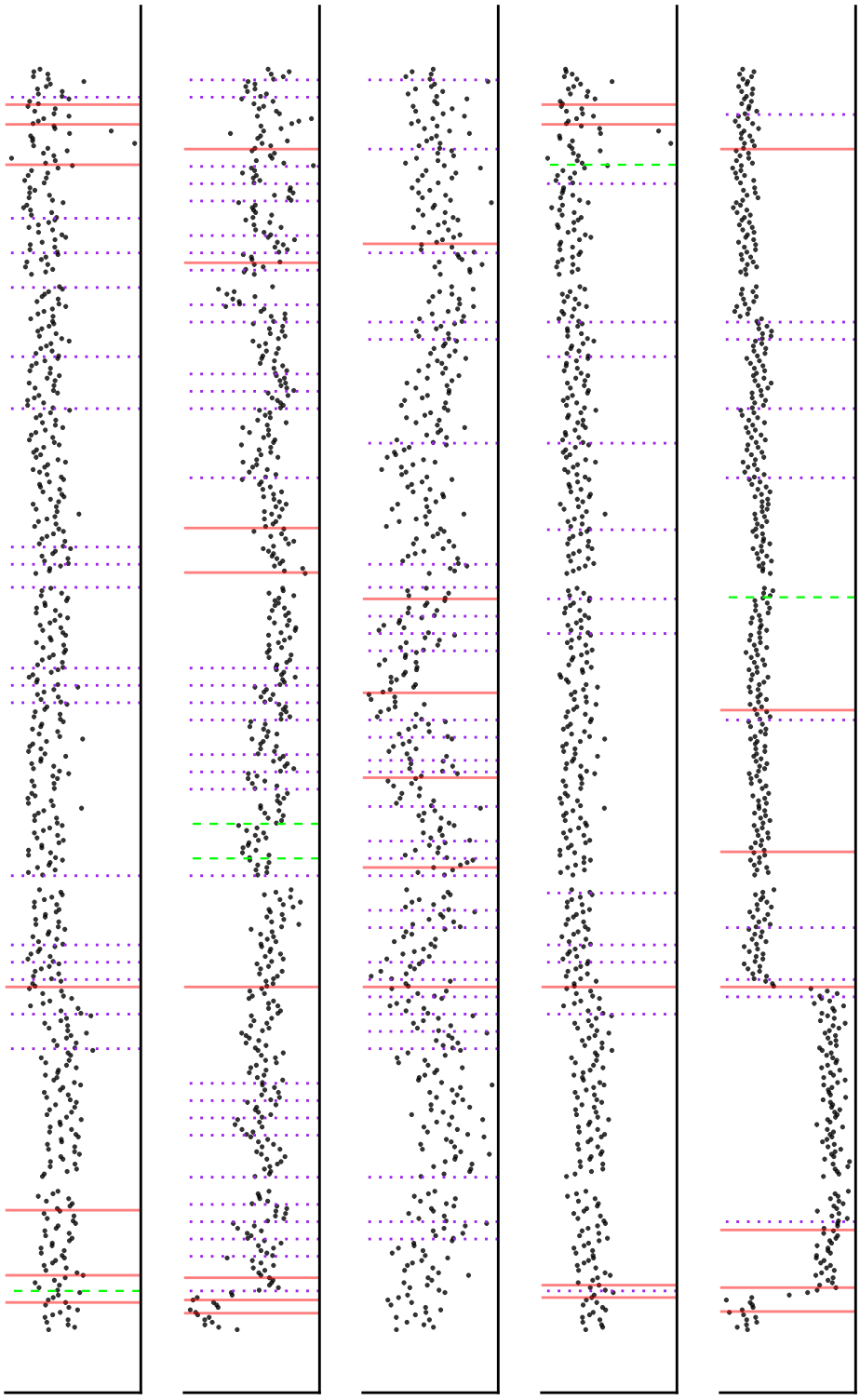

Figure 5.4.8: Changes reported by SPOT and by price changes for series with a small number of discount periods. Dashed lines indicate change in price has been detected by SPOT, solid lines indicate change in price has been missed by SPOT while dotted lines indicate that SPOT has detected a change not related to a change in price.

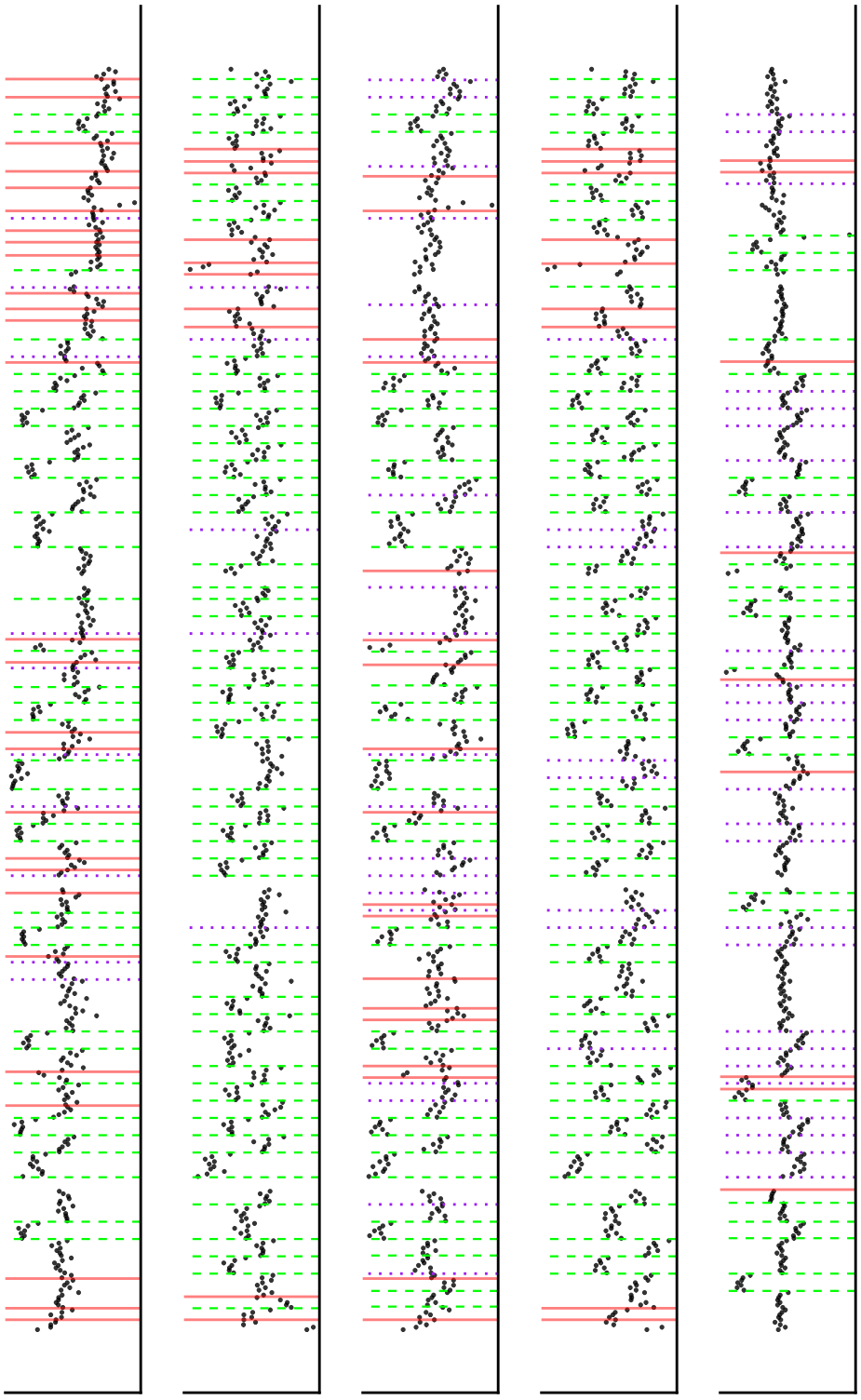

Figure 5.4.9: 4 products that are most predictive of unexplained changes in target product sales. Bottom: Changepoint plot for target product sales.

5.5 Conclusion

In this chapter, we consider whether a multivariate changepoint analysis can be used to study the impact of promotions on products unaffected by the promotion. Using changes in price as a ground truth, we examined whether a multivariate analysis can identify promotions that do occur. We also investigated whether there was any interesting patterns in the changes that don't correspond with promotions.

Our results show that a multivariate changepoint analysis reliably detects changes in product sales due to promotions. Particularly for series where they occur frequently, the method identified promotions with high accuracy. Furthermore the segmentation locates changes in products that are not affected by promotions. These changes are particularly interesting as they may indicate relationships between products. We investigated a simple approach for identifying such relationships given a segmentation. Using domain knowledge provided by the Data Science team, we validated that this approach may in fact identify real relationships between products. However we note that the evidence was mixed and the true relationships we detected may have been identified by chance.

We note that there was a significant challenge with this approach. The method reported a large number of changes where no promotion occured, especially for series with few promotions. In some cases, these changes may correspond with interesting changes (i.e. changes due to a promotion in a related product), however they may also be caused by overfitting changes. There are two possible causes for overfitting of changes. Firstly, extra changes may be fitted due to approximation error in the fitting method. Secondly, there may be extra structure in the data, not captured by the model, which induces changes. It is challenging to separate overfitted changes from interesting changes without examining each product individually. As a result, it is difficult to determine whether or not, interesting patterns occur in the changes that do not correspond to promotions.

Future work may want to consider, whether a more complex model for the data can reduce the issue of overfitting changes, without reducing the accuracy of the

changepoint estimates. Consideration of a joint estimation of the global and local parameters may have benefits here. Furthermore a significant difficultly with a multivariate changepoint analysis, is the difficulty in understanding the output given that only a small subset of the analysed series can be visually examined. Therefore, it may be worthwhile investigating new ways of plotting and analysing multivariate changepoint output.

Chapter 6

Changes in Covariance

6.1 Introduction

Data of increasing size and complexity are being collected in an ever growing list of fields. A common issue in handling such data is that the underlying distributional properties can change over time. Statistical models must take account of this heterogenity for accurate inference. One approach is to assume that the changes occur at a small number of discrete time points known as changepoints. Changepoint methods are relevant in a wide range of applications including genetics (Hocking et al., 2013), network traffic analysis (Rubin-Delanchy et al., 2016) and oceanography (Carr et al., 2017). We consider the specific case where the covariance structure of the data changes at each changepoint. This problem is relevant in a number of applications. For example, Stoehr et al. (2020) examine changes in the covariance structure of functional Magnetic Resonance Imaging (fMRI) data, where a failure to satisfy stationarity assumptions can significantly contaminate any analysis. Furthermore, Wied et al. (2013) and Berens et al. (2015) examine how detecting changes in the covariance structure of financial time series can be used to improve stock portfolio optimisation.

The changepoint problem has a long history in the statistical literature, dating back at least as far as Page (1954). The literature contains two distinct but closely related problems, online changepoint detection and offline changepoint detection. Online changepoint detection considers the case where data is observed sequentially over

time and the aim is to detect any changes as quickly as possible. In the offline setting, the data is observed as a single batch and we aim to locate potentially multiple changepoints. We focus on the latter problem however, readers interested in the former should see Tartakovsky et al. (2014) for a thorough review.

The literature on detecting changes in multivariate time series has grown substantially in the last few years. In particular, many authors consider changes in the high dimensional setting, that is, where the number of the parameters of the model, is significantly larger than the number of data points. Much of this work considers changes in expectation where the series are uncorrelated (Grundy et al., 2020; Horváth and Hušková, 2012). Furthermore a number of authors have examined changes in expectation where only a subset of variables under observation change (Enikeeva and Harchaoui, 2019; Jirak et al., 2015; T. Wang and Samworth, 2018). Separately a number of authors have considered changes in second order structure of high dimensional time series models including auto-covariance and cross-covariance (Cho and Fryzlewicz, 2015), changes in graphical models (Gibberd and Nelson, 2014, 2017) and changes in network structure (D. Wang, Yu, and Rinaldo, 2018).

The problem of detecting changes in the covariance structure has been examined in both the low dimensional and high dimensional setting. In the low dimensional setting J. Chen and Gupta (2004) and Lavielle and Teyssiere (2006) utilise a likelihood based test statistic and the Schwarz Information Criterion (SIC) to detect changes in covariance of normally distributed data. Aue, Hörmann, et al. (2009) consider a nonparameteric test statistic for changes in the covariance of linear and non-linear multivariate time series. Matteson and James (2014) study changes in the distribution of (possibly) multivariate time series using a clustering inspired nonparametric test statistic that claims to handle covariances. In the high dimensional setting, Avanesov and Buzun (2018) and D. Wang, Yu, and Rinaldo (2017) study test statistics based on the distance between sample covariances, utilising the operator norm and ℓ_∞ norm respectively.

In this work, we propose a novel method for detecting changes in the covariance structure of high dimensional time series. We study a test statistic inspired by a

distance metric intuitively defined on the space of positive definite matrices. The primary advantage of this metric is that under the null hypothesis of no change, it is independent of the underlying covariance structure which is not the case for other methods in the literature. Using results from Random Matrix Theory (RMT), we study the asymptotic properties of this test statistic, when the dimension of the data is of comparable size to (but still smaller than) the sample size. The structure of this discussion is as follows. In Section 6.2, we discuss an important limitation of current state of the art methods, and introduce a two sample test statistic that does not suffer this limitation. In Section 6.3, we derive an asymptotic distribution for the test statistic using Random Matrix Theory (RMT). In Section 6.4, we discuss how this test statistic can be used to detect changes in the covariance structure of time series. In Section 6.5, we study the finite sample performance of our approach on simulated datasets and compare it with other state of the art methods. Finally in Section 6.6, we use our method and other state of the art methods to examine how changes in the covariance structure of pixel intensities can be used to detect changes in the amount of water on the surface of a plot of soil.

6.2 Two Sample Tests for the Covariance

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^p$ be independent p dimensional vectors with

$$\mathrm{Cov}\left(\boldsymbol{X}_i\right) = \boldsymbol{\Sigma}_{i,p}, \text{ for } 1 \leq i \leq n. \tag{6.2.1}$$

where each $\boldsymbol{\Sigma}_{i,p} \in \mathbb{R}^{p \times p}$ is full rank. Furthermore, let $\boldsymbol{X}_{n,p}$ denote an $n \times p$ matrix defined by $\boldsymbol{X}_{n,p} := (\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_n^T)$. Our primary interest in this paper is to develop a testing procedure that can identify a change in the covariance structure of the data over time. For now, let us consider the case of a single changepoint. We compare a null hypothesis of the data sharing the same covariance versus an alternative setting that allows a single change at time τ. Formally we have

$$H_0 : \boldsymbol{\Sigma}_0^* = \boldsymbol{\Sigma}_{1,p} = \cdots = \boldsymbol{\Sigma}_{n,p} \tag{6.2.2}$$

$$H_1 : \boldsymbol{\Sigma}_1^* = \boldsymbol{\Sigma}_{1,p} = \cdots = \boldsymbol{\Sigma}_{\tau,p} \neq \boldsymbol{\Sigma}_{\tau+1,p} = \cdots = \boldsymbol{\Sigma}_{n,p} = \boldsymbol{\Sigma}_2^*, \tag{6.2.3}$$

where τ is unknown. We would like to be able to distinguish between the null and alternative hypothesis, and locate the changepoint τ under the alternative. We are interested in the setting where the dimension of the data p, is of comparable size to the length of the data, n. In particular, we require that for all pairs n, p, the set

$$\mathcal{T}_{n,p}(\alpha) := \{t \in \mathbb{Z}^+ \text{ such that } p + \alpha < t < n - p - \alpha\} \tag{6.2.4}$$

is non empty, where $\alpha \in \mathbb{Z}^+$ is a problem dependent positive constant. Note $\mathcal{T}_{n,p}(\alpha)$ defines the set of possible candidate changepoints, while $p + \alpha$ is the minimum distance between changepoints or minimum segment length. Then for each candidate changepoint $t \in \mathcal{T}_{n,p}(\alpha)$, a two sample test statistic $T(t)$ can be used to determine if the data to the left and right of the changepoint have different distributions. If the two sample test statistic for a candidate exceeds some threshold, then we say a change has occured and an estimator for τ is given by the value $t \in \mathcal{T}_{n,p}(\alpha)$ that maximises $T(t)$.

Let $\bar{\Sigma}(\cdot, \cdot)$ be a sample covariance estimator defined as follows,

$$\bar{\Sigma}(p, q) := \frac{1}{q - p} \sum_{i=p+1}^{q} \boldsymbol{X}_i \boldsymbol{X}_i^T.$$

For a given changepoint candidate τ, we can detect whether a change has occured by measuring the distance between the sample covariance estimates, $\bar{\Sigma}(0, \tau)$ and $\bar{\Sigma}(\tau, n)$. A natural choice for the distance measure is given by the magnitude of the matrix $\bar{\Sigma}(0, \tau) - \bar{\Sigma}(\tau, n)$. Indeed, we can express three of the most important test statistics in the literature, Aue, Hörmann, et al. (2009), Avanesov and Buzun (2018), and D. Wang, Yu, and Rinaldo (2017) as,

$$\max_{\ell < \tau \leq n - \ell} \|\alpha_{\tau,1} \bar{\Sigma}(0, \tau) - \alpha_{\tau,2} \bar{\Sigma}(\tau, n)\|, \tag{6.2.5}$$

where $\{\gamma_{\tau,1}\}_{\tau=\alpha+1}^{n-\alpha}$, $\{\gamma_{\tau,2}\}_{\tau=\alpha+1}^{n-\alpha}$ are sequences of normalizing constants, α is the minimum segment length and $\|\cdot\|$ is some norm which measures the size of the difference matrix (such as the operator norm or infinity norm), .

The difference matrix above may seem like an intuitive approach to detect changepoints, however it can be difficult to use in practice. Under the null hypothesis, we

can express (6.2.5) as

$$\max_{\alpha<\tau\leq n-\alpha} \|\Sigma_0^{\frac{1}{2}}(W_1 - W_2)\|$$

where $W_1 \sim W_p(\tau, \boldsymbol{I})$ and $W_2 \sim W_p(n-\tau, \boldsymbol{I})$ and $W_p(t, \boldsymbol{V})$ is the p dimensional Wishart distribution with t degrees of freedom and scale matrix $\boldsymbol{V}$. As a result, the scale of the difference matrix is a function of the underlying covariance, Σ_0, and a test statistic based on the difference matrix must be corrected to account for this. For example, Aue, Hörmann, et al., 2009 normalize their test statistic using the sample covariance for the whole data, Avanesov and Buzun, 2018 use a bootstrap procedure which assumes knowledge of Σ_0 and D. Wang, Yu, and Rinaldo, 2017 use a threshold which is a function of Σ_0. All these approaches require estimating Σ_0 in practice. This is impractical under the alternative setting, since estimating the segment covariances requires knowledge of the changepoint.

Therefore, it is natural to ask whether there are alternative ways of measuring the distance between covariance matrices. In the univariate setting, a common approach is to evaluate the logarithm of the ratio of the segment variances (J. Chen and Gupta, 1997; Inclan and Tiao, 1994; Killick, Eckley, et al., 2010). This is in contrast with the change in expectation problem where it is more common to measure the difference between sample means. In the variance setting, a ratio is more appropriate for two reasons. Firstly, since variances are strictly positive, if the underlying variance is quite small then the absolute difference between the mean values will also be small whereas the ratio is not affected. Secondly, under the null hypothesis of no change, the variances will cancel and the test statistic will be independent of the variance. Thus, there is no need to estimate the variance when calculating the threshold.

We propose to extend this ratio idea from the univariate setting to the multivariate setting by studying the multivariate ratio matrix,

$$R(A, B) := (B^T B)^{-1} A^T A, \tag{6.2.6}$$

where $A \in \mathbb{R}^{n\times p}$ and $B \in \mathbb{R}^{m\times p}$. Ratio matrices are widely used in multivariate analysis to compare covariance matrices (Finn, 1974). In particular, we are often interested in functions of the eigenvalues of these matrices (Lawley, 1938; Potthoff

and Roy, 1964; Wilks, 1932). Here we are interested in the following test statistic,

$$T(A,B) = \sum_{j=1}^{p} \left(1 - \lambda_j(R(A,B))\right)^2 + \left(1 - \lambda_j^{-1}(R(A,B))\right)^2, \tag{6.2.7}$$

where $\lambda_j(R(A,B))$ is the jth largest eigenvalue of the matrix $R(A,B)$.

The proposed test statistic has two valuable properties that may not be immediately obvious. Firstly it is symmetric i.e $T(\boldsymbol{X}_1, \boldsymbol{X}_2) = T(\boldsymbol{X}_2, \boldsymbol{X}_1)$. This is important for a changepoint analysis as the segmentation should be the same regardless of whether the data is read forewards or backwards. Secondly, the distribution of the test statistic, $T(\boldsymbol{X}_1, \boldsymbol{X}_2)$ is independent of the covariances of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, if the covariances of these matrices are equal. This is extremely valuable in the changepoint setting as under the null hypothesis of no change, the two samples will have the same covariance and thus the test statistic for each candidate will not depend on the underlying covariance Σ_0. Therefore this test statistic has the same useful properties that the ratio approach has in the variance setting. The following result demonstrates that the test statistic in (6.2.7) does indeed have these properties.

Proposition 6.2.1. *Let $\boldsymbol{X}_1 \in \mathbb{R}^{n_1 \times p}$ and $\boldsymbol{X}_2 \in \mathbb{R}^{n_2 \times p}$ be random matrices drawn from some distribution L_1 with covariance $\boldsymbol{\Sigma}$. Then we have that*

1. *$T(\boldsymbol{X}_1, \boldsymbol{X}_2) = T(\boldsymbol{X}_2, \boldsymbol{X}_1)$ (Symmetry)*
2. *The distribution of $T(\boldsymbol{X}_1, \boldsymbol{X}_2)$ does not depend on the covariance matrix $\boldsymbol{\Sigma}$*

Proof. Proof in Appendix C.2. □

Note test statistics that can be expressed via equation (6.2.5) do not have this property. However these properties are clearly not unique to our chosen test statistic T, and there are many other possible choices (such as $\log^2 x$).

It is both possible and interesting to study the properties of this test statistic in the finite dimensional setting (i.e. where p is fixed). However in this work, our focus is on high dimensional problems where the dimension of the data is of comparable size to the length of the data. In the next section, we consider the properties of this test statistic as a two sample test under the null hypothesis of no change and compute the

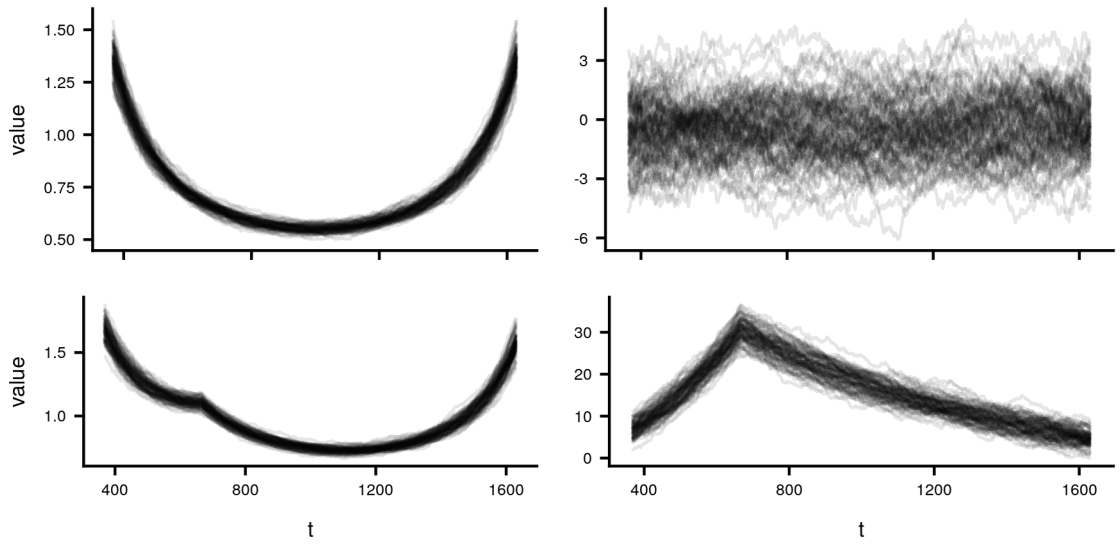

Figure 6.3.1: Test statistic T defined in (6.2.7) applied to a 100 different data sets before (left) and after standardisation (right) using (6.3.4) under the null setting (top) and alternative setting (bottom) with $n = 2000$, $p = 100$ and $\tau = 666$.

asymptotic moments of the distribution using results from Random Matrix Theory. We have chosen T as it is possible to compute these moments analytically, which is not true for other functions.

6.3 Random Matrix Theory

We now describe some foundational concepts in Random Matrix Theory (RMT), before discussing how these ideas are utilised to identify the asymptotic distribution of our test statistic under the null hypothesis. RMT concerns the study of matrices where each entry is a random variable. In particular, RMT is often concerned with the behaviour of the eigenvalues and eigenvectors of such matrices. Interested readers should see Tao (2012) for an introduction and Anderson et al. (2010) for a more thorough review.

A key object of study in the field is the Empirical Spectral Distribution (ESD),

defined for a $p \times p$ matrix $\boldsymbol{A}$ as

$$F^{\boldsymbol{A}}(x) := \frac{1}{p}\sum_{j=1}^{p} I(\lambda_{p-j}(\boldsymbol{A}) \leq x) \tag{6.3.1}$$

where I is an indicator function. In other words, the ESD of $\boldsymbol{A}$ is a discrete uniform distribution placed on the eigenvalues of $\boldsymbol{A}$. Several authors have established results on the limiting behaviour of the ESD as the dimension tends to infinity, the so called Limiting Spectral Distribution (LSD). For example, Wigner (1967) demonstrate that if the upper triangular entries of a Hermitian matrix $\boldsymbol{A}$ have mean zero and unit variance, then $F^{1/\sqrt{p}\boldsymbol{A}}(x)$ converges to the Wigner semicircular distribution.

The LSD of the ratio matrix, defined in (6.2.6), was shown to exist in Yin et al., 1983 and computed analytically in Silverstein, 1985. The following two assumptions are sufficient for the LSD of an F matrix to exist.

Assumption 6.3.1. *Let $\boldsymbol{X}_{n_1,p} \in \mathbb{R}^{p\times n_1}$ and $\boldsymbol{X}_{n_2,p} \in \mathbb{R}^{p\times n_2}$ be random matrices with independent not necessarily identically distributed entries $\{X_{n_1,i,j}, 1 \leq i \leq n_1, 1 \leq j \leq p\}$ and $\{X_{n_2,k,j}, 1 \leq k \leq n_2, 1 \leq j \leq p\}$ with mean 0, variance 1 and fourth moment $1+\kappa$. Furthermore, for any fixed $\eta > 0$,*

$$\frac{1}{n_1 p}\sum_{j=1}^{p}\sum_{i=1}^{n_1} \mathbb{E}|X_{n_1,i,j}|^4 \boldsymbol{I}(|X_{n_1,j,k}| \geq \eta\sqrt{n_1}) \to 0 \tag{6.3.2}$$

$$\frac{1}{n_2 p}\sum_{j=1}^{p}\sum_{i=1}^{n_2} \mathbb{E}|X_{n_2,i,j}|^4 \boldsymbol{I}(|X_{n_2,j,k}| \geq \eta\sqrt{n_2}) \to 0 \tag{6.3.3}$$

as n_1, n_2, p tend to infinity subject to Assumption 6.3.2.

Assumption 6.3.2. *The sample sizes n_1, n_2, and the dimension p grow to infinity such that*

$$\gamma_{n_1} := \frac{p}{n_1} \to \gamma_1 \in (0,1),\ \gamma_{n_2} := \frac{p}{n_2} \to \gamma_2 \in (0,1) \text{ and } \boldsymbol{\gamma} := (\gamma_1, \gamma_2).$$

For simplicity, we will refer to the limiting scheme described in Assumption 6.3.2 as $\boldsymbol{n} \to \infty$.

Let $\boldsymbol{X}_{n_1,p}, \boldsymbol{X}_{n_2,p}$ be matrices satisfying Assumptions 6.3.1 and 6.3.2. Furthermore, let $F_{\boldsymbol{n}}$ denote the ESD of $R(\boldsymbol{X}_{n_1,p}, \boldsymbol{X}_{n_2,p})$. Then Silverstein, 1985 demonstrate that

$F_{\boldsymbol{n}}$ converges almost surely to the non random distribution function

$$F_{\boldsymbol{\gamma}}(dx) = \frac{1-\gamma_2}{2\pi x(\gamma_1+\gamma_2 x)}\sqrt{(b-x)(x-a)}I_{[a,b]}(x)dx \text{ as } \boldsymbol{n}\to\infty$$

where

$$a = \frac{(1-h)^2}{(1-\gamma_2)^2}, \quad b = \frac{(1+h)^2}{(1-\gamma_2)^2}, \quad h = \sqrt{\gamma_1+\gamma_2-\gamma_1\gamma_2}.$$

The LSD, $F_{\boldsymbol{\gamma}}$ provides an asymptotic centering term for functions of the eigenvalues of random ratio matrices. In particular, for any function f, we have that,

$$\mathbb{E}_{F_n}(f) = \frac{1}{p}\sum_{i=1}^{p} f(\lambda_i(R(\boldsymbol{X}_{n_1,p}, \boldsymbol{X}_{n_2,p}))) \to \int f(x)dF_{\boldsymbol{\gamma}}(x) = \mathbb{E}_{F_{\boldsymbol{\gamma}}}(f) \text{ as } n_1, n_2, p \to \infty$$

by the definition of weak convergence. This allows us to account for bias in the statistic as seen in Figure 6.3.1.

The rate of convergence of $|\mathbb{E}_{F_n}(f) - \mathbb{E}_{F_{\boldsymbol{\gamma}}}(f)|$ to zero was studied in Zheng, 2012 and found to be $1/p$. In particular, the authors establish a central limit theorem for the quantity,

$$G_{\boldsymbol{n}}(x) := p\left[F_{\boldsymbol{n}}(x) - F_{\boldsymbol{\gamma}}(x)\right].$$

We can apply this result to our problem in order to demonstrate that our two sample test statistic converges to a normal distribution with known mean and variance terms.

Theorem 6.3.1. *Let $X_{n_1} \in \mathbb{R}^{n_1\times p}$ and $X_{n_2} \in \mathbb{R}^{n_2\times p}$ be random matrices satisfying Assumptions 6.3.1 and 6.3.2 and $T(\dot{;})$ be defined as in* (6.2.7)*. Then we have that as $\boldsymbol{n}\to\infty$,*

$$T(X_{n_1}, X_{n_2}) - p\int f^*(x)dF_{\boldsymbol{\gamma}}(x) \to N(\mu(\boldsymbol{\gamma}), \sigma^2(\boldsymbol{\gamma}))$$

where

$$\begin{aligned}
f^*(x) &= (1-x)^2 + (1-1/x)^2 \\
\mu(\boldsymbol{\gamma}) &= 2K_{3,1}\left(1-\frac{y_2^2}{h^2}\right) + \frac{2K_{2,1}y_2}{h} + 2K_{3,2}\left(1-\frac{y_1^2}{h^2}\right) + \frac{2K_{2,2}y_1}{h} \\
\frac{1}{2}\sigma^2(\boldsymbol{\gamma}) &= K_{2,1}^2 + 2K_{3,1}^2 + K_{2,2}^2 + 2K_{3,2}^2 + \\
&\quad \frac{J_1K_{2,1}}{h} + \frac{J_1K_{2,1}}{h(h^2-1)} - \frac{J_1K_{3,1}(h^2+1)}{h^2} - \frac{J_1K_{3,1}}{h^2(h^2-1)} + \\
&\quad \frac{J_2K_{2,1}2h}{(h^2-1)^3} + \frac{J_2K_{3,1}}{h^2} + \frac{J_2K_{3,1}(1-3h^2)}{h^2(h^2-1)^3}
\end{aligned}$$

and

$$K_{3,1} = \frac{h^2}{(1-y_2)^4}, \;\; K_{2,1} = \frac{2h(1+h^2)}{(1-y_2)^4} - \frac{2h}{(1-y_2)^2},$$
$$K_{3,2} = \frac{h^2}{(1-y_1)^4}, \;\; K_{2,2} = \frac{2h(1+h^2)}{(1-y_1)^4} - \frac{2h}{(1-y_1)^2},$$
$$J_1 = -2(1-y_2)^2 \; and \; J_2 = (1-y_2)^4$$
$$h = \sqrt{y_1 + y_2 - y_1 y_2}, \;\; y_1 = \frac{p}{n_1} \; and \; y_2 = \frac{p}{n_2}.$$

Proof. Proof in Appendix C.2. □

Using Theorem 6.3.1, we can properly normalise T such that it can be applied to a changepoint analysis. In particular, we have that under the null hypothesis

$$T(\bar{\Sigma}(0,t), \bar{\Sigma}(t,n)) - p \int f^*(x) dF_{\boldsymbol{\gamma}_{t/n}} \to N(\mu(\boldsymbol{\gamma}_{t/n}), \sigma^2(\boldsymbol{\gamma}_{t/n})) \tag{6.3.4}$$

as n, p tend to infinity, where $\gamma_{t/n} := (p/t, p/(n-t))$ and f is as defined in Theorem 6.3.1. Thus we utilise the normalised test statistic, $\tilde{T}$,

$$\tilde{T}(t) := \sigma^{-1/2}(\boldsymbol{\gamma}_{t/n}) \left(T(\bar{\Sigma}(0,t), \bar{\Sigma}(t,n)) - p \int f^*(x) dF_{\boldsymbol{\gamma}_{t/n}} - \mu(\boldsymbol{\gamma}_{\tau/n}) \right),$$

which under the null hypothesis converges pointwise to a standard normal random variable.

The asymptotic moments of the test statistic, T, depend on the parameter $\boldsymbol{\gamma}_{t/n}$, and as t approaches p (or equivalently $n-p$) the mean and variance of the test statistic dramatically increase. In the context of changepoint analysis, this implies that the mean and variance increase at the edges of the data. We note that this is a common result for changepoint test statistics. We can significantly reduce the impact of this by the above standardisation. This can be seen empirically in Figure 6.3.1. After standardisation, the test statistics for the series with no change, do not appear to have any structure. Similarly, the test statistics for the series with a change show a clear peak at the changepoint. Importantly we can now easily distinguish the test statistic under the null and alternative hypotheses, and this normalization does not require knowledge of the underlying covariance structure.

6.4 Practical Considerations

Before we can apply our method to real and simulated data, we need to address three practical concerns. In particular, we need to choose a threshold for rejecting the null hypothesis of no change, determine an appropriate minimum segment length and address the issue of multiple changepoints.

6.4.1 Threshold for Detecting a Change

Firstly, we need to select an appropriate threshold for determining whether or not to reject the null hypothesis. We choose to utilise the asymptotic distribution of the test statistic, and assume independence between the test statistic value for different candidates. In particular, we say that

$$\max_{\alpha<t<n-\alpha} \tilde{T}(t) \approx \max_{\alpha<t<n-\alpha} Z_t$$

where Z_t are standard normal variables. Thus, we need to choose a threshold, β_n, dependent on the length of the data, such that

$$P\left(\max_{\alpha<t<n-\alpha} Z_t > \beta_n\right) \to 0 \text{ as } n \to \infty.$$

Note however, that if we choose β to be too large then our method will be overly conservative. Motivated by results from univariate changepoint analysis (Csorgo and Horváth, 1997), we choose $\beta = \log(n)$ to balance these competing priorities. As we shall see in the simulation study, this choice gives high probability of detection with a low risk of false positives.

6.4.2 Minimum Segment Length

Secondly, we must also consider an appropriate choice for the minimum segment length parameter, α. In many applications, domain specific knowledge may be used to increase this parameter. However, it is also important to consider the smallest value that will give reliable results in the general case. The minimum segment length must grow sufficiently fast to ensure that $\tilde{T}(t)$ converges to a normal distribution.

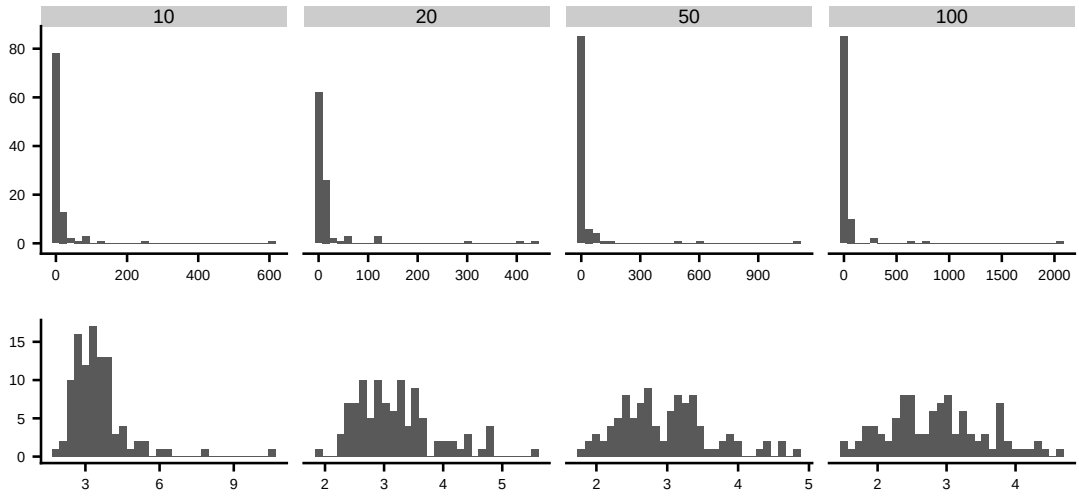

Figure 6.4.1: Histogram of values of $\max_{\alpha<t<n-\alpha} \tilde{T}(t)$ applied to 100 datasets of length $n = 2000$ with no change for $p = \{10, 20, 50, 100\}$ with $\alpha = p$ (top) and $\alpha = 4p$ (bottom).

Outside the asymptotic regime, it is possible for the ratio matrix to have very large eigenvalues. Thus for candidate changepoints t close to p (or by symmetry $n-p$), the probability of observing spuriously large values of $\tilde{T}(t)$ becomes much larger. This can be seen in Figure 6.4.1. When $\alpha = p$ (the smallest possible value), we observe extremely large values of the test statistic that would make identifying a true change almost impossible. On the other hand, when $\alpha = 4p$, the test statistic behaves more reliably.

We need $p/(p + \alpha_{n,p})$ to converge to $\gamma_\alpha \in (0, 1)$ or equivalently $\alpha_{n,p} = \mathcal{O}(p)$ for the asymptotic results to hold. However it is important that $\alpha_{n,p}$ not be too small in the finite sample setting as this may significantly limit the types of datasets that we can apply the method to. Therefore, we require a sequence $\alpha_{n,p}$ that appropriate manages this trade off. In Section 6.5, we analyse the effect of different sequences in the finite sample setting via a simulation study. D. Wang, Yu, and Rinaldo, 2017 choose . We demonstrate that the sequence proposed by D. Wang, Yu, and Rinaldo, 2017, $\alpha_{n,p} = p \log n$, is also appropriate for our setting producing a low false positive rate. However, we find that it is quite conservative for larger values of p.

6.4.3 Multiple Changepoints

Finally, we also consider the extension to multiple changes. In this setting, we have a set of m unknown ordered changepoints, $\boldsymbol{\tau} := \{0 = \tau_0, \tau_1, \ldots \tau_m, \tau_{m+1} = n\}$ such that,

$$\Sigma_i = \Sigma_k^*,\ \tau_k < i \leq \tau_{k+1},\ 1 \leq k \leq m+1,$$

where Σ_t is the covariance matrix of the ith vector. We are interested in estimating the number of changes m, and the set of changepoints $\boldsymbol{\tau}$. The classic approach to this problem is to extend a method defined for the single changepoint setting to the multiple changepoint setting, via an appropriate search method such as dynamic programming (Killick, Fearnhead, et al., 2012) or binary segmentation (Scott and Knott, 1974). For this work, we do not consider the dynamic programming approach. The dynamic programming approach minimises the within segment variability through a cost function for each segment. This is not compatible with our approach which maximises the distance between segments. Therefore, for our simulations with multiple changepoints, we utilise the classic binary segmentation procedure.

The binary segmentation method extends a single changepoint test as follows. Firstly, the test is run on the whole data. If no change is found then the algorithm terminates. If a changepoint is found, it is added to the list of estimated changepoints, and the binary segmentation procedure is then run on the data to the left and right of the candidate change. This process continues until no more changes are found. Note the threshold β, and the minimum segment length α, remain the same.

Finally, we note that a number of extensions of the traditional binary segmentation procedure have been proposed in recent years (Fryzlewicz, 2014, 2020; Olshen et al., 2004). We do not use these search methods in our simulations, as they incorporate additional hyperparameters that affect performance. However it is not difficult to incorporate our proposed test statistic into these methods.

6.5 Simulations

In this section, we compare our method with other state of the art methods in the literature, namely the methods of Aue, Hörmann, et al., 2009; Avanesov and Buzun, 2018; D. Wang, Yu, and Rinaldo, 2017. Software implementing these methods is not currently available and as a result, we have implemented each of these methods according to the descriptions in their respective papers. More complete descriptions of these methods are provided in Section 2.3.2. The methods have been implemented in the R programming language.

Simulation studies in the current literature for changes in covariance structure are very limited. D. Wang, Yu, and Rinaldo, 2017 do not include any simulations. Aue, Hörmann, et al., 2009; Avanesov and Buzun, 2018 only consider the single changepoint setting, and do not consider random parameters for the changes. Furthermore to our knowledge, no papers compare the performance of different methods. While theoretical results are clearly important, it is also necessary to consider the finite sample performance of any estimator, and we now study the finite sample properties of our approach on simulated datasets. For all our simulations, we sample an initial covariance matrix, Σ_0 from a Wishart distribution with diagonal covariance. For the ith changepoint, we sample a positive definite transition matrix $\boldsymbol{\Delta}_i$ as follows,

$$\begin{aligned} \boldsymbol{W}_i &\sim \text{Wishart}(\boldsymbol{I}_p, p) & \boldsymbol{W}_i &= \boldsymbol{Q}_i \boldsymbol{R}_i \\ \lambda_j(\Delta_i) &\sim \text{Gamma}(5, 0.2) & \boldsymbol{\Delta}_i &= \boldsymbol{Q}_i^T \boldsymbol{\Lambda}_i \boldsymbol{Q}_i \end{aligned}$$

where $\boldsymbol{\Lambda}_i$ is a diagonal matrix with $\boldsymbol{\Lambda}_{jj} = \lambda_j$. Note taking the QR decomposition from a random Wishart is equivalent to uniformly sampling from the set of real valued orthonormal matrices. The Gamma distribution was chosen to ensure that the eigenvalues are positive, and that the determinant of the matrix does not get too large or small. The covariance matrix for the new segment is given by,

$$\Sigma_i = \Delta_i^{1/2} \Sigma_{i-1} \Delta_i^{1/2}.$$

Throughout this section, the significance thresholds for each method are set as follows unless otherwise stated. The threshold for our method is set to the default

setting of $\log n$ as discussed in the Section 6.5. In Remark 2.1, Aue, Hörmann, et al., 2009 state that the asymptotic distribution of their test statistic after standardisation can be approximated by a standard normal distribution. Therefore we set the threshold for detecting a change to be the 95% quantile or 1.96. Note this could be increased, reducing the probability of overfitting changes but also reducing the power of the method. This approach also requires a plug in estimator for the long run covariance of the vectorized second moment of the data. Since there is no temporal structure in the simulated datasets we consider, this long run covariance is exactly the covariance of the vectorized second moment and we use the empirical estimate as our plug in estimator. This should improve the performance of the method compared with a generic plug in estimator for the long run covariance. This matrix has dimension $p(p+1)/2$ where p is the dimension of the data, and must be inverted which significantly limits the size of datasets we can consider with this method. As a result, we do not include this method in simulations with large datasets.

D. Wang, Yu, and Rinaldo, 2017 do not provide a practical default threshold for their method, instead providing an interval of consistent thresholds which is defined by theoretical quantities such as the minimum size of a change, the minimum distance between changes and a bound on the tails of the data, B. A lower bound on the minimum threshold is given by $B^2\sqrt{p \log n}$. The value B bounds the square root of the largest eigenvalue of the covariance of the underlying data, which implies the largest eigenvalue is a lower bound for B. Note this value is not available in practice so we approximate this quantity with the largest eigenvalue of the data. Thus a lower bound for the threshold is given by $\lambda_{\max}(X)\sqrt{p \log n}$. Again if this value was increased, the method would lose power but be less likely to overfit changes.

The rest of the section is structured as follows. We begin by assessing the chosen default values for the minimum segment length parameter and threhold for determining a change proposed in Section 6.4. We then study the properties of our method and previously discussed methods in the single changepoint setting, considering both random and fixed changepoint locations. Finally we examine the performance of the different methods in the multiple changepoint setting.

6.5.1 Assesment of minimum segment length and threshold

In order to control the false positive rate of the method, we need appropriate choices of the minimum segment length, α, and the threshold, β. In the previous section, we proposed a default value of $\beta = \log n$. For this threshold to be appropriate, it should produce a low false positive rate, that goes to 0 as n grows. We generated 100 datasets of length $n = \{200, 500, 1000, 2000, 5000, 10000\}$ and $p = \{3, 5, 10, 20, 50, 100\}$ and applied the proposed method to each. We set $\alpha_{n,p} = p \log n$ for all scenarios.

The results of this analysis can be seen in Figure 6.5.1. We can see that for all values of p, the FPR decreases as n grows. However, we do note that the FPR is higher for smaller values of p. This is likely due to the fact that, for small values of p, the test statistic has not reached the limiting regime and thus, the threhold is misspecified.

From the above experiments we can see that, setting $\alpha = p \log n$ suitably controls the FPR. However, this may be overly conservative. For large values of p, this minimum segment length becomes very large. Therefore, it is worthwhile investigating whether lower values can be used. With this in mind, we repeated the experiment with $\alpha_{n,p} = 1.5p$. The results of this analysis can be seen in Figure 6.5.1. For small and moderate values of p, this produces a much larger FPR. However for large values ($p >= 20$), the method performs equally well. This indicates that for large values of p, smaller values of α can be considered. In settings where smaller values of α are required, it is necessary to ensure that the results are robust to small changes in the value of α.

6.5.2 Single Changepoint

We now compare our approach with some state of the art methods Aue, Hörmann, et al., 2009; Avanesov and Buzun, 2018; D. Wang, Yu, and Rinaldo, 2017, which are labelled in graphs as Aue, Av.Buzun and Wang. Our approach is labelled Ratio i.e. Ratio matrix. For all our simulated examples, we let the minimum segment length or distance between changes be $2p \log n$ as this is required by the method of D. Wang,

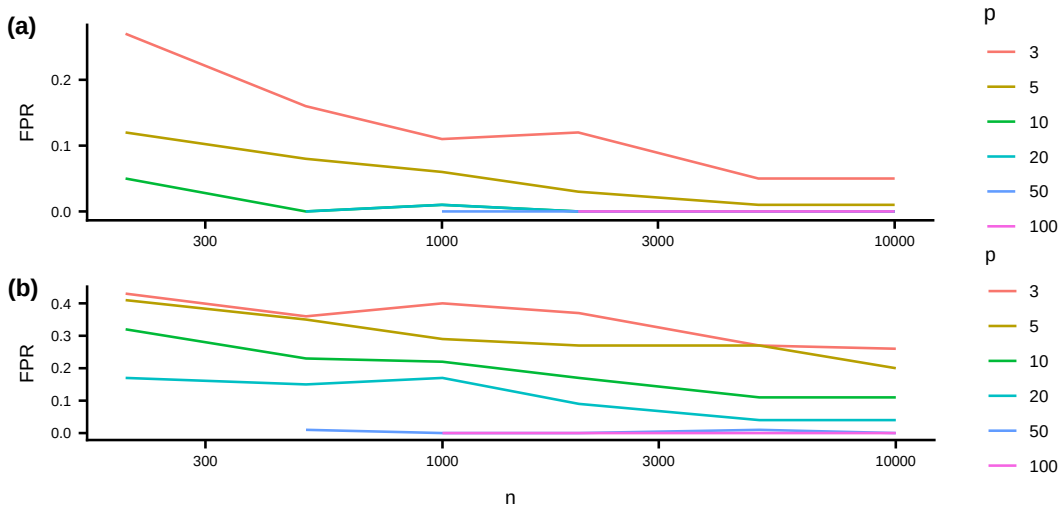

Figure 6.5.1: (a) False postive rate for our approach for 100 datasets with no change with $\alpha_{p,n} = p \log n$; (b) Same with $\alpha_{p,n} = 1.5p$.

Yu, and Rinaldo, 2017.

We compare the four approaches on a set of 100 datasets with a change at $\tau = \lfloor n/3 \rfloor$. We consider two settings with the first case having a moderate value for p ($p = 15, n = 500$), and the second case having a larger value for p ($p = 100, n = 2000$). Importantly in the second setting, we should be closer to the asymptotic regime for our method as n and p are larger. For each dataset we computed the test statistic as well as the difference between the truth and the changepoint estimates for each method. Note that the method of Aue, Hörmann, et al., 2009 is not computable for the $p = 100$ case, and as a result is not included for this case. The results of this simulation can be seen in Figure 6.5.2.

In the small p case, our approach and the method of Aue, Hörmann, et al., 2009 clearly outperform the other methods. Looking at Figure 6.5.2 (a), the methods labelled Wang and Av.Buzun are very poorly peaked indicating they have not detected a change. On the other hand, the methods labelled Aue and Ratio have large peaks indicating clear localization of the change. Neither the Wang nor the Av.Buzun method accurately locates the changepoints as can be seen in Figure 6.5.2 (b). However the Av.Buzun method performs particularly poorly. It is not entirely clear why this is the case, however we offer two possible explanations. Firstly the method was developed

for the truly high dimensional setting and thus may lack power in this application. Secondly the method incorporates hyperparameters which we set to default settings. The method may be more effective if these were fine tuned to the problem. Finally we note that our approach gives the most accurate changepoint estimates in terms of concentration around the true value. This is likely due to the fact that the Aue test statistic decays slowly after the change, which leads to changepoint estimates which are biased to the right.

In the large p setting (Figure 6.5.2b), the proposed Ratio approach clearly outperforms the Wang method. As in the low dimensional case, the Wang test statistic is nearly completely flat and fails to estimate the changepoint location, while the Av Buzun method completely fails to detect any signal. On the other hand, our approach is clearly peaked and gives very accurate changepoint estimates. However we note that there is a slight bias to the right in the changepoint location.

In order to further compare these approaches, we also apply the methods to examples with a single changepoint at a random location. We generated 1000 datasets of length $n = \{200, 500, 1000, 2000, 5000\}$ and $p = \{3, 5, 10, 20, 50, 100\}$, where the change is sampled uniformly over $\{\lfloor p \log n \rfloor + 2, \ldots, n - \lfloor p \log n \rfloor\}$. For each dataset and method we computed a changepoint candidate (ignoring whether the change was significant or not). We then calculated the error in estimating the changepoint location as the absolute difference between the true change and the estimate. The results for the different approaches are shown in Figure 6.5.3. We can see that for larger values of n and p, our approach gives estimates with the lowest error. The error for our approach reduces substantially as p increases. This is very clear for $p \geq 10$, however there is also a substantial improvement going from $p = 3$ to $p = 10$. This improvement is not seen in the other methods.

6.5.3 Multiple Change Points

We now explore the performance of our method on simulated data sets with multiple changepoints. We begin by defining our performance metrics. Firstly throughout we use we use $\boldsymbol{\tau} := \{\tau_1, \ldots, \tau_m\}$ and $\hat{\boldsymbol{\tau}} := \{\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{m}}\}$ to denote the set of true

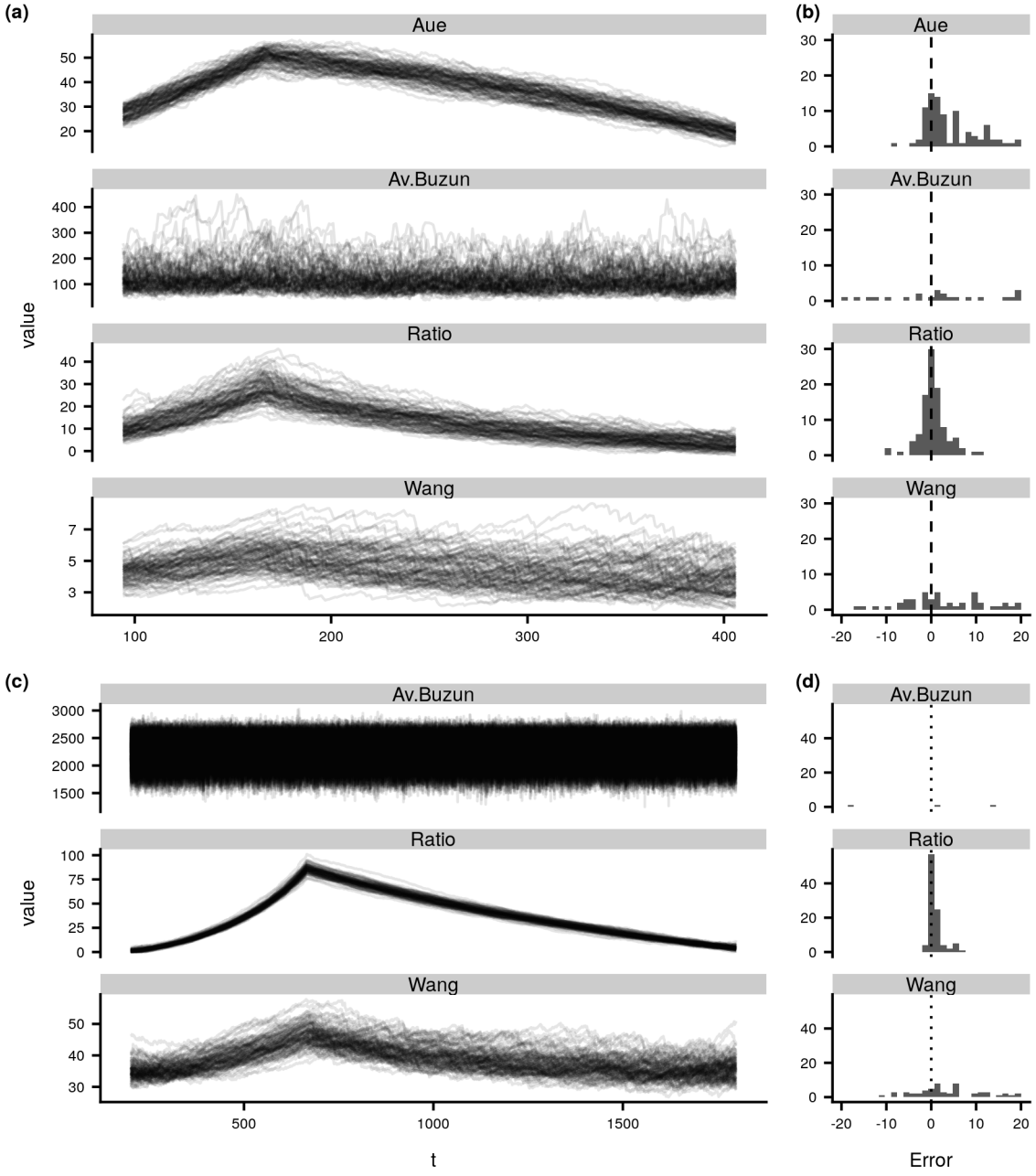

Figure 6.5.2: (a) Test statistic at each time point from a 100 different data sets under the alternative setting with $p = 15$, $n = 500$ and a changepoint at $n/3$. (b) Histogram of the difference between the estimated changepoint location and the true changepoint. (c) Same as (a) for $p = 100$ and $n = 2000$. (d) Same as (b) for $p = 100$ and $n = 2000$.

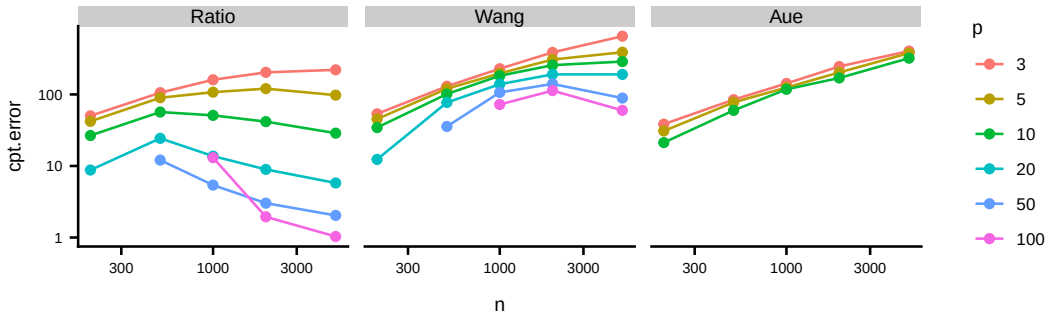

Figure 6.5.3: Mean absolute difference between the estimated and true changepoint locations from 100 different data sets (ignoring whether the changepoint is significant) for three different methods over increasing values of p and n.

changepoints and the set of estimated changepoints respectively. A common approach for evaluating changepoint methods is to examine true and false discovery rates. We say that the changepoint τ_i has been detected correctly if

$$\min_{1\leq j\leq \hat{m}} |\hat{\tau}_j - \tau_i| \leq h.$$

Note that this is an adaptation of the changepoint location error used in the previous section for the multiple changepoint setting. Throughout this section, we set $h = 20$ although it should be noted that in reality the desired accuracy would be application specific and while the specific values vary with h, the conclusions of the study do not. We denote the set of correctly estimated changes by $\boldsymbol{\tau}_c$. Then we define the true discovery rate (TDR) and false discovery rate (FDR) as follows,

$$TDR := \frac{|\boldsymbol{\tau}_c|}{|\boldsymbol{\tau}|},\ FDR := \frac{|\hat{\boldsymbol{\tau}}| - |\boldsymbol{\tau}_c|}{|\hat{\boldsymbol{\tau}}|}.$$

We also consider whether or not the resulting segmentation allows us to estimate the true underlying covariance matrices. Therefore for each method, we also compute the mean absolute error (MAE) and spectral mean absolute error (SMAE) as follows,

$$MAE := \frac{1}{n}\sum_{i=1}^{n} \|\hat{\Sigma}_i - \Sigma_i\|_1 \text{ and } SMAE := \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p} |\lambda_j(\hat{\Sigma}_i) - \lambda_j(\Sigma_i)|$$

We consider datasets with 5 changepoints uniformly sampled with minimum segment length $p \log n$, where $p = \{5, 10, 20, 50, 100\}$ and $n = \{200, 500, 1000, 2000, 5000\}$.

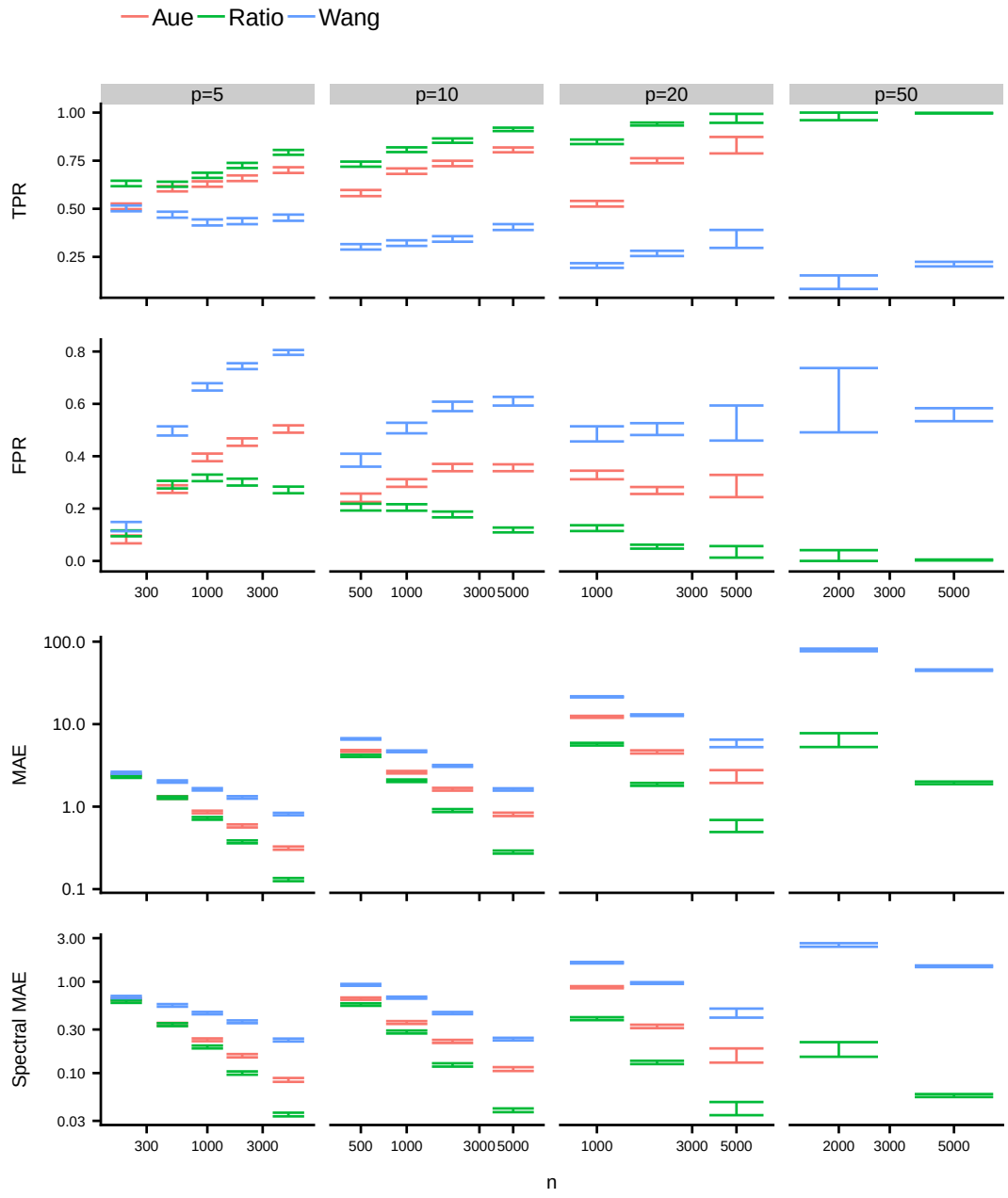

Figure 6.5.4: Each error bar gives a bootstraped 95% confidence intervals for the average error for that method across 1000 replications each with 4 changepoints. Note our approach out performs the others across almost all parameter combinations.

The true covariance matrices are generated as in the previous section. For each (n, p) pair, we generated 1000 datasets and applied our method, the Aue method and the Wang method to each dataset. As it performed so poorly in the single changepoint setting, we do not include the Av.Buzun method in this simulation. Furthermore for $p = 50$, the minimum segment length for the Aue method is 1275, which means the minimum data length is 6375. This is longer than the longest data set we consider. Therefore we do not run the Aue method for $p > 50$. Using the resulting segmentations, we then calculated the error metrics for each method. In order to compare the different approaches in a statistically sound manner, we calculated confidence intervals for the mean error across the replications for each method and (n, p) pair via bootstrap resampling. If there is no overlap in the confidence intervals then there is a statistically significant difference in the average error for the methods.

The results of this analysis are shown in Figure 6.5.4. The worst performer across all metrics is the Wang method. Notably the true positive rate for the method decreases as p grows. This is in striking contrast with the other methods which become more accurate for larger values of p as one may expect. This may be due to the fact that, the Wang method only considers the first principal component of the difference matrix, ignoring the remainder of the spectrum. For larger values of p, this quantity may account for less of the overall change. Furthermore, the method also has the highest false positive rate, indicating that adapting the threshold would not lead to more accurate changepoint estimates. The Aue method outperforms the Wang method and is competitive with our approach for small values of n and p. However for larger values of n and p, our method outperforms it with higher TPR and much lower FPR. Importantly, the FPR for the Aue method increases with n. This is due to the fact that the threshold is based on the asymptotic distribution and does not take the length of the data into account.

6.6 Application: Detecting changes in moisture levels in soil

In this section, we investigate whether changes in the covariance structure of soil data correspond with shifts in the amount of moisture in the soil. There is significant interest in developing new techniques to better understand how water is absorbed and travels through soil. This is an important question and is relevant to a variety of industrial applications such as farming and construction (Hillel, 2003). An important challenge in this area is measuring the level of moisture in the soil. A widely used approach is to place probes at different depths and locations in the soil which measure the level of moisture. However this approach has a number of limitations. Firstly we only measure soil near the probe, which means a lot of information is lost. Secondly the probes do not give any information about what happens at the surface. This issue becomes particularly important if the soil is very dry as moisture can struggle to move through dry soil and the water can stay on the surface. Similarly when the soil reaches saturation water can also stay on the surface. How fast water drains from the surface is indicative of the moisture level of the soil. To measure this across a site more easily, scientists are investigating the use of cameras to capture the soil surface.

We analyse images from an experiment studying moisture on the surface of the soil. A camera was placed over a large plot of soil and took a set of 589 pictures over a day. Examples of these photos can be seen in Figure 6.6.1. At different times, different amounts of rainfall are simulated and the amount of water on the soil surface changes. This is particular obvious in the small trench that runs through the center of the plot. At different times during the observation period, streams of water of different volumes appear in the trench. We wish to segment the data based on the flow of the stream, partitioning the data into wet and dry periods.

The intensity of a set of pixels over time is shown in Figure 6.6.2 . We can see that the mean level is clearly nonstationary. This nonstationary behaviour may be attributed to two causes, changes in the background light intensity (due to a cloud passing by) and changes in the wetness of the soil which changes how much light

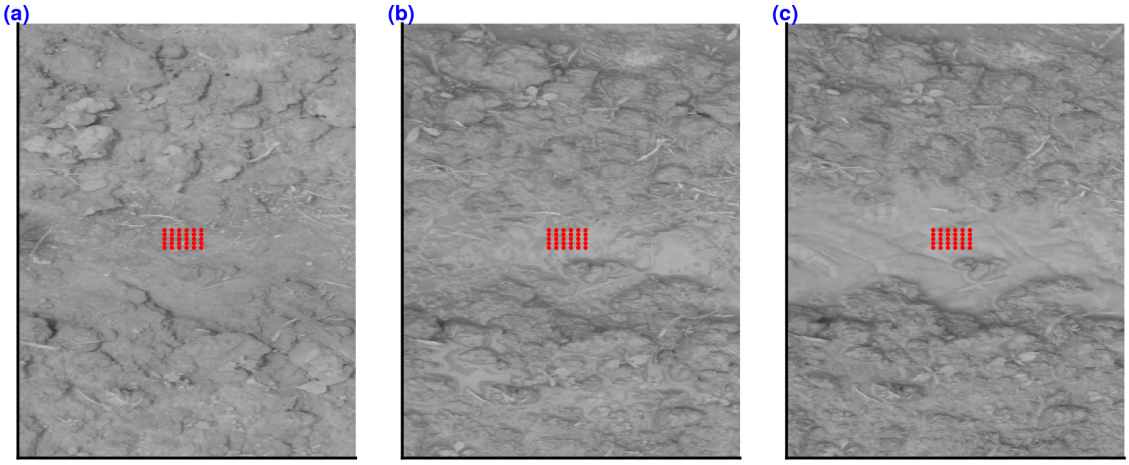

Figure 6.6.1: Soil at different times with different levels of moisture. The soil starts off dry and then at different times varying amounts of moisture are added. The red dots indicate the 30 pixels we analyse for changes in covariance.

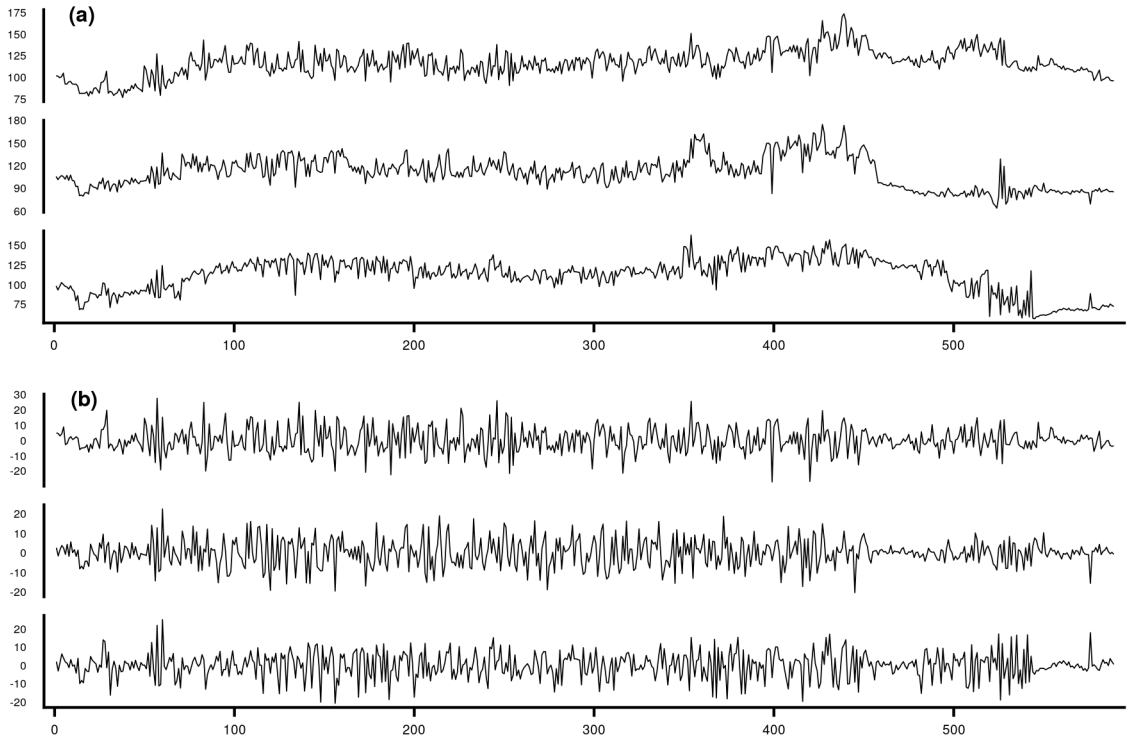

Figure 6.6.2: (a) Raw grayscale intensities for three pixels. (b) Standardised intensities for the same three pixels.

is reflected. Since changes in the mean intensity are not necessarily associated with changes in the wetness, we instead focus on changes in the covariance structure. When pixels become wet, we expect that the correlation between the pixels should increase as they become more alike as the surface becomes uniformly water instead of the variable soil surface. Thus changes in the covariance structure of the pixels may correspond with changes in the wetness.

The data consists of 589 images with resolution 1480×690. The original pictures are in colour but were transferred to grayscale for computational purposes. Each pixel provides information about a very small part of the pit. In order to increase the amount of information in each image, we compressed the images further by averaging over each 3×3 block, which gives an image with resolution 690×230. Note this approach to compression is naive and more advanced approaches may lead to better results. We analysed two subsets of pixels ($p = 10, 30$) in the center of the images which can be seen in Figure 6.6.2. These pixels are arranged in a grid with space between the pixels to to reduce the correlation between pixels. We run a multiple changepoint analysis on the smaller subset using our approach as well as the Aue and Wang methods. We also ran a multiple changepoint analysis on the larger subset. However due to the dimension of the larger dataset, the Aue method can only identify a single changepoint for this data. Since we know that there are more changes we do not use the Aue method when analysing the larger dataset.

In order to analyse the covariance structure of the data, we first need to transform the data to have stationary mean. There are two obvious approaches to this task, calculating a time varying estimate of the mean and differencing until stationary. The latter approach induces autocorrelation into the data which is problematic, therefore we choose the former approach. Estimating the mean of this series is challenging as there is stochastic volatility and the smoothness of the function appears to change over time. As a result, standard smoothing methods such as LOESS and windowed mean estimators may be inappropriate. We use a Bayesian Trend Filter with Dynamic Shrinkage (Kowal et al., 2019) which is robust to these issues. We use this method as implemented in the DSP package (Kowal, 2020). We then transform the data to

stationary, by taking the difference between the raw data and the estimated mean.

The transformed data for a subset of the pixels can be seen in Figure 6.6.2. We can see that the transformed data has a stationary mean, however the variance is clearly nonstationary. We combined the three methods with the standard binary segmentation procedure in order to detect multiple changes in covariance. The minimum segment length was set to 25. The thresholds for significance for each method were again set to the defaults as discussed in the previous section. The results of this analysis are shown in Table 6.6.1. We begin by discussing the results for the smaller subset and then move on to the larger subset afterwards.

In order to validate our results we worked with scientists currently studying this data and and identified three clear time points where there is a substantial change in the amount of water on the surface at the relevant pixels. The first change is somewhat gradual going from very dry at time $t = 64$ to very wet from time $t = 76$. The second and third changes are more abrupt, with a substantial increase in the amount of water at time $t = 350$ and a corresponding sharp decrease at time $t = 450$. The Aue method reports 7 changepoints, the Wang method reports 5 changepoints and our method locates 8 changepoints. All methods detect the first and last changes. However the Wang method does not detect any change near the second anticipated changepoint. All of the methods appear to overfit changepoints, in the sense that they report changes that do not correspond with clear changes in the amount of water on the surface. For our method and the Aue method, the majority of these overfitted changes occur when the soil is dry (before t=64 and after t=450). During these periods the amount of light exposure varies much more from image to image which may explain these nuisance changes.

For the larger dataset, the minimum segment length was set to 60 (twice the number of variables) and the thresholds were set to their defaults. The results were broadly similar for our method and quite different for the Wang method. Our approach reports 6 changes again detecting the three obvious changes in the video. We note that the reduced number of changepoints is primarily due to the increased minimum segment length. The Wang method only reports a single changepoint. This

Method	Small subset ($p = 10$)	Larger subset($p = 30$)
Aue	66, 101, 243, 354, 451, 514, 589	NA
Wang	52,79, 184, 237, 445	445
Ratio	49, 77, 244, 347, 452, 493, 532,562	64, 125, 184, 255, 340, 450, 527

Table 6.6.1: Detected changepoints for each of the three methods when applied to the soil image data. Note the dimension of the larger subset means the Aue method can detect at most one changepoint.

drop in reported changes is caused by the largest eigenvalue of the sample covariance being much larger. As a result, the threshold for detecting a change is 3.5 times larger to account for this and consequently, it appears that the method loses power.

6.7 Conclusion

In this work, we have presented a novel test statistic for detecting changes in the covariance structure of high dimensional data. This geometrically inspired test statistic has a number of desirable properties that are not features of competitor methods. Most notably our approach does not require knowledge of the underlying covariance structure. We utilise results from Random Matrix Theory to derive a limiting distribution for our test statistic. The proposed method outperforms other methods on simulated datasets, in terms of both accuracy in detecting changes and estimation of the underlying covariance model. We then use our method to analyse changes in the amount of surface water on a plot of soil. We find that our approach is able to detect changes in this dataset that are visible to the eye and locates a number of other changes. It is not clear whether these changes correspond to true changes in the surface water and we are investigating this further.

While our method has a number of advantages, it is important to recognise some limitations. Firstly, our method requires calculating the inverse of a matrix at each time point, which is a computationally and memory intensive operation. As a result, our approach is infeasible for larger datasets that can be considered by other methods,

which only require the first principle component. However, as we demonstrate through simulations, there are a wide range of settings where our method produces better results for a marginal increase in computational time. Finally we note that a limitation of our method is that the minimum segment length is bounded below by the dimension of the data. This means that the method cannot be applied to tall datasets ($p > n$) or datasets with short segments.

Chapter 7

Conclusion

In this thesis, we have presented new methodology for detecting changepoints in multivariate data. We have considered two distinct settings; firstly we consider data with changepoints where not every variable under observation is affected by the change, and secondly high dimensional data with changes in covariance structure. Our goal in this work has been to introduce new computationally efficient algorithms that can detect changes in these settings with high accuracy for potentially very large datasets.

The vast majority of methodology available for detecting changes in multivariate data assumes that every variable is affected by each changepoint and ignores the question of estimating affected variables. In Chapters 3 and 4 we consider the subset multivariate changepoint model, which uses a doubly penalised cost function approach in order to simultaneously estimate both the locations of changepoints and the set of variables affected by each change. While this approach offers a number of advantages from a statistical perspective, it requires solving a challenging discrete optimisation problem via a computationally intensive dynamic program, SMOP, that is infeasible for even small datasets. We make two key contributions to the literature on this problem. Firstly in Chapter 3 we introduce a new algorithm, PSMOP, for computing an exact solution to the discrete optimisation problem. This method incorporates a preprocessing step which utilises novel search space reduction techniques to remove bad candidate changepoints. In simulations, we demonstrate that the preprocessing step significantly reduces the computational cost of computing an exact solution and,

PSMOP has a much lower computational cost than the original dynamic program SMOP. Furthermore, we demonstrate the subset multivariate changepoint approach can outperform both state of the art univariate and multivariate methods.

Although the PSMOP procedure is considerably more efficient than the original dynamic program, it is still infeasible for datasets of moderate size. Therefore in Chapter 4, we introduce an approximate dynamic program, SPOT, which can efficiently compute near optimal solutions to the discrete optimisation problem for even very large datasets. This approximation uses windowed cost functions, which evaluate model parameters on a subset of the data. Although our approach is not exact, we demonstrate that a classic consistency result can be extended to windowed cost functions and that SPOT is guaranteed to outperform equivalent methods which assume every variable is affected by the change. Furthermore, under mild conditions on the number of changepoints, we demonstrate that the computational cost of the algorithm is linear in both the number of datapoints and the dimension of the data. In simulations, we observe that SPOT can identify changepoints and affected variables in very large datasets and outperforms other multivariate methods.

In Chapter 6 we consider the problem of detecting changes in the covariance structure of data, where the dimension of the data is large compared to the length of the data. Our key contribution in this chapter is a new test statistic for detecting changes, for which the distribution under the null hypothesis of no change is independent of the structure of the underlying covariance of the data. As a result the threshold for determining a change does not depend on the data. To our knowledge, there are no other methods for detecting changes in covariance that have this property. Using results from Random Matrix Theory, we derived a limiting distribution for our test statistic. We then developed a rigourous simulation study, to analyse the finite sample properties of our estimator and compare our approach with other state of the art methods. To our knowledge, this simulation study is the first to rigourously compare methods for detecting changes in covariance. Finally we used the new method to detect changes in the amount of moisture on the surface of soil.

7.1 Further Directions

We now discuss three possible directions in which the work presented in this thesis could be extended and developed further in the future:

- Develop data driven strategies for selecting the penalty parameters α and β
- Detecting subset multivariate changepoints in data with dependence between variables
- Further develop the theoretical results presented in Chapter 6

7.1.1 Data Driven Penalty Selection

In the univariate setting, selecting an approriate value for the penalty parameter β can be very challenging. One solution to this problem is to use data driven strategies to set the parameter value. For example, Haynes et al., 2017 propose a procedure that efficiently solves the univariate penalised optimisation problem for a range of penalties. Then the optimal parameter value can be determined via an elbow plot or by comparing the segmentations with domain knowledge. Given that both SMOP and SPOT use two parameters, identifying optimal parameter values may be more challenging and a strategy for correctly setting them even more valuable. There are two key challenges with developing such a strategy for the dual penalty setting.

Firstly we need to be able to efficiently compute segmentations for a range of penalties. In the single penalty setting, given an interval of possible penalty values, there is a discrete (and typically small) number of possible segmentations. The CROPS algorithm proposed by Haynes et al., 2017 efficiently computes a set of penalties $\{\beta_1, \ldots, \beta_f\}$ such that for all $\beta \in (\beta_{i-1}, \beta_i)$ the optimal segmentation is the same. Thus it is possible to evaluate every possible optimal segmentation within the interval. We could propose a similar method for the dual penalty setting, which would look to identify a discrete set of pairs $\{(\alpha_1, \beta_1,), \ldots, (\alpha_f, \beta_f)\}$ such that for all $\alpha \in (\alpha_{i-1}, \alpha_i)$ and $\beta \in (\beta_{i-1}, \beta_i)$ the implied segmentations would be the same. Note a further extension would be to extend such work to the approximate method SPOT.

The second challenge would be to understand the relationship between the parameters α and β. In the univariate setting, the relationship between the number of changepoints and the penalty parameter value is well understood. If $\beta_u > \beta_l$, the optimal segmentation implied by β_u will have at most as many changes as the optimal segmentation implied by β_l. However it is not possible to make a similar statement about the dual penalty framework. In particular, we do not fully understand how changes in the α and β parameters change the segmentations. For example, we might assume that by increasing the α penalty, the number of variables affected by each change will decrease. However as the α parameters increases, changepoint locations may become unprofitable. As a result, the number of variables affected by each changepoint increases. Quantifying the impact of changes in the parameters on the resulting segmentation would thus significantly help in analysing and comparing the resulting segmentations.

7.1.2 Dual Penalty Framework with Dependence

The dual penalty framework assumes that the variables under observation are uncorrelated. However in many applications we are interested in examining multiple variables precisely because they are correlated and, in such an application it is necessary to take account of this dependence. If this dependence structure is ignored, the dual penalty approach will be more likely to overfit changepoints (as spuriously large test statistic values may occur across multiple variables) or miss true changes.

There are a range of different possible methods for taking account of this dependence structure, depending on what kind of assumptions we can make about the dependence. If we knew the underlying covariance matrix, we could apply a whitening transformation to the data and then apply the standard dual penalty framework to the transformed data. Note as we saw in Chapter 6, in many settings we do not know a priori the underlying covariance matrix and estimating this quantity in the subset multivariate setting may be quite challenging. Therefore for this to be a workable strategy, we would need a method for estimating the covariance matrix that is robust to the changes in the distribution.

7.1.3 Finite Sample Results for the Covariance Test Statistic

In Chapter 6 we used a key result from Random Matrix Theory to obtain a convergence result for our test statistic. This result allows us to properly normalise the test statistic and develop an appropriate threshold for distinguishing between the null and alternative hypotheses. However this asymptotic result has a significant limitation, it only demonstrates pointwise convergence for the sequence of test statistic values. As a result, the threshold for significance is based on approximating the distribution of the test statistic with the asymptotic distribution. By obtaining a stronger convergence result (such as uniform convergence), we could get valuable information about how the method performs such as the error in the aforementioned approximation. Furthermore, if the results were finite sample in nature, we could derive confidence intervals which are very important in any analysis.

We suggest two possible directions for developing such a convergence rate. The primary result in Chapter 6 is based on results from Z. Bai and Silverstein, 2004 and Zheng, 2012 which develop central limit theorems for tests of the spectra of covariance matrices. These results are limited to the two sample setting, however by extending the results to the multiple testing setting, we would simultaneously get a uniform convergence result for the covariance test statistic. However these results are not finite sample. There has been significant work done on developing concentration inequalities for the trace of functions of random matrices (Guionnet, Zeitouni, et al., 2000). Thus strong finite sample results for our test statistic could be derived by applying these concentration inequalities. Note these results could be used to develop confidence intervals. One limitation of these results is that they are only suitable for Lipschitz functions. Since the F matrix is not bounded, it is not Lipschitz and therefore we would instead have to work with a bounded variant. However these results would still provide valuable information about the method.

Appendix A

Appendix for Chapter 3

A.1 Useful Results

Lemma A.1.1. *Let $\boldsymbol{c}_f$ be changepoint vector such that $c_f^k = v$ and $c_f^l = v' < v$. Furthermore suppose that $\boldsymbol{c}_p$ is a changepoint vector such that $c_p^k = s$ where $v' < s < v$ and $c_p^l < v'$. Then there exists $\boldsymbol{c}$ such that $c^l = v'$ and*

$$\begin{aligned} &F(\boldsymbol{c}) + \sum_{j=1}^{p} \left[I\left(c^j \neq c_f^j\right) \left(\mathcal{C}^j\left(\boldsymbol{c}, \boldsymbol{c}_f\right) + \alpha \right) \right] + m(\boldsymbol{c}, \boldsymbol{c}_f)\beta \\ \leq &F(\boldsymbol{c}_p) + \sum_{j=1}^{p} \left[I\left(c^j \neq c_f^j\right) \left(\mathcal{C}^j\left(\boldsymbol{c}_p, \boldsymbol{c}_f\right) + \alpha \right) \right] + m(\boldsymbol{c}_p, \boldsymbol{c}_f)\beta \end{aligned}$$

and if there is equality then the segmentation implied by having $\boldsymbol{c}_p$ prior to $\boldsymbol{c}_f$ is equal to the segmentation implied by $\boldsymbol{c}$.

Proof. Firstly let $\boldsymbol{c}$ be a changepoint vector defined as

$$c^j = \begin{cases} c_f^j \text{ if } c_f^j \leq v' \\ c_p^j \text{ otherwise.} \end{cases}$$

Similarly let $\tilde{\boldsymbol{c}}$ be a changepoint vector defined as

$$\tilde{c}^j = \begin{cases} \ell(c_p)^j \text{ if } c_p^j = c^j \\ c_p^j \text{ otherwise.} \end{cases}$$

By construction we have that $\ell(\boldsymbol{c_p}) \prec \tilde{\boldsymbol{c}} \prec \boldsymbol{c} \prec \boldsymbol{c}_f$. Thus we have two implied segmentations given by,

$$\boldsymbol{c}_f, \boldsymbol{c}_p, \ell(\boldsymbol{c}_p) \text{ and } \boldsymbol{c}_f, \boldsymbol{c}, \tilde{\boldsymbol{c}}, \ell(\boldsymbol{c}_p).$$

By construction these two segmentations give the same set of changepoints which proves the second claim. Furthermore it is not guaranteed that $\tilde{\boldsymbol{c}} = \ell(\boldsymbol{c})$, which proves the first claim. □

Lemma A.1.2. *Let $\boldsymbol{c}_p, \boldsymbol{c}_f, \boldsymbol{c}$ be changepoints such that $c_p^k = t < v = c_f^k$, $\boldsymbol{c}_p \prec \boldsymbol{c}_f$ and if $c_f^j = t$ then $c_p^j = t$. Furthermore let $c^k = s$ and $c^j = c_p^j$ for $j \neq k$. Then*

$$m(\boldsymbol{c}_p, \boldsymbol{c}) + m(\boldsymbol{c}, \boldsymbol{c}_f) \leq m(\boldsymbol{c}_p, \boldsymbol{c}_f) + 1.$$

Proof. We break this proof into two cases, the case where $c^j = t$ for some $j \neq k$ and the complement. In the former case, we have that $m(\boldsymbol{c}_p, \boldsymbol{c}) = 0$. Then since $\boldsymbol{c}$ and $\boldsymbol{c}_p$ disagree on at most one changepoint location, we have that

$$m(\boldsymbol{c}_p, \boldsymbol{c}) + m(\boldsymbol{c}, \boldsymbol{c}_f) = m(\boldsymbol{c}, \boldsymbol{c}_f) \leq m(\boldsymbol{c}_p, \boldsymbol{c}_f) + 1.$$

For the second case we have that,

$$c^j \neq t \text{ for } 1 \leq j \leq p \implies c_p^j \neq t \text{ for } j \neq k \text{ and } c_f^j \neq t \text{ for } 1 \leq j \leq p.$$

Then

$$m(\boldsymbol{c}, \boldsymbol{c}_f) = m(\boldsymbol{c}_p, \boldsymbol{c}_f) \text{ and } m(\boldsymbol{c}_p, \boldsymbol{c}) = 1,$$

which gives

$$m(\boldsymbol{c}_p, \boldsymbol{c}) + m(\boldsymbol{c}, \boldsymbol{c}_f) = m(\boldsymbol{c}_p, \boldsymbol{c}) + 1.$$

□

A.2 Proofs for Section 3.2

Proof of Proposition 3.3.1. Firstly note that if $c_f^j = t$ for some $j \neq k$, then by Lemma A.1.1 it must be the case that $c_p^j = t$ or there exists $\boldsymbol{c}'$ such that $c'^j = t$, $\boldsymbol{c}'$ gives

an equivalent segmentation to $\boldsymbol{c}$ and the cost of having $\boldsymbol{c}'$ be the changepoint vector prior to $\boldsymbol{c}_f$ is bounded above by the equivalent cost for $\boldsymbol{c}_p$. If the latter case is true, let $\boldsymbol{c}_p = \boldsymbol{c}'$. Then let $\boldsymbol{c}$ be a changepoint vector such that $c^j = s$ for some $1 \leq j \leq p$ and $c^j = c_p^j$ otherwise. Note by construction $\boldsymbol{c}_p, \boldsymbol{c}$ and $\boldsymbol{c}_f$ satisfy Lemma A.1.2. Then

$$
\begin{aligned}
&F(\boldsymbol{c}_p) + \sum_{j=1}^{p} \left[I\left(c_p^j \neq c_f^j\right) \left(\mathcal{C}^j\left(\boldsymbol{c}_p, \boldsymbol{c}_f\right) + \alpha\right)\right] + m(\boldsymbol{c}_p, \boldsymbol{c}_f)\beta \\
&\quad = F(\boldsymbol{c}_p) + \sum_{j \neq k} \left[I\left(c_p^j \neq c_f^j\right) \left(\mathcal{C}^j\left(\boldsymbol{c}_p, \boldsymbol{c}_f\right) + \alpha\right)\right] + m(\boldsymbol{c}_p, \boldsymbol{c}_f)\beta + \mathcal{C}^k(t, v) + \alpha \\
&\quad = F(\boldsymbol{c}_p) + \sum_{j \neq k} \left[I\left(c^j \neq c_f^j\right) \left(\mathcal{C}^j\left(\boldsymbol{c}, \boldsymbol{c}_f\right) + \alpha\right)\right] + m(\boldsymbol{c}_p, \boldsymbol{c}_f)\beta + \mathcal{C}^k(t, v) + \alpha \\
&\quad > F(\boldsymbol{c}_p) + \mathcal{C}^k(t, s) + \alpha + (m(\boldsymbol{c}_p, \boldsymbol{c}_f) + 1)\beta + \\
&\quad \sum_{j \neq k} \left[I\left(c^j \neq c_f^j\right) \left(\mathcal{C}^j\left(\boldsymbol{c}, \boldsymbol{c}_f\right) + \alpha\right)\right] + \mathcal{C}^k(s, v) + \alpha \\
&\quad \geq F(\boldsymbol{c}_p) + \mathcal{C}^k(t, s) + \alpha + m(\boldsymbol{c}_p, \boldsymbol{c})\beta + \\
&\quad \sum_{j \neq k} \left[I\left(c^j \neq c_f^j\right) \left(\mathcal{C}^j\left(\boldsymbol{c}, \boldsymbol{c}_f\right) + \alpha\right)\right] + \mathcal{C}^k(s, v) + \alpha + m(\boldsymbol{c}, \boldsymbol{c}_f)\beta \\
&\quad \geq F(\boldsymbol{c}) + \sum_{j=1}^{p} \left[I\left(c^j \neq c_f^j\right) \left(\mathcal{C}^j\left(\boldsymbol{c}, \boldsymbol{c}_f\right) + \alpha\right)\right] + m(\boldsymbol{c}_c, \boldsymbol{c}_f)\beta \geq F(\boldsymbol{c}_f).
\end{aligned}
$$

□

Proof of Proposition 3.2.3. Firstly by the definition of A_τ we have that

$$
\boldsymbol{c}, \boldsymbol{c}' \in A_\tau \implies \exists 1 \leq k, l \leq p \text{ such that } c^k = \tau \geq c'^k \text{ and } c'^l = \tau \geq c^l,
$$

which implies the result. Secondly if $\boldsymbol{c} \prec \boldsymbol{c}'$ we have that $c^j < \tau$ for $1 \leq j \leq p$ which completes the proof. □

A.3 Proofs for Section 3.3.1

The proof of Proposition 3.3.1 is quite technical but the motivation is simple. We demonstrate that, given any model that includes a segment starting from $t+1$ to v in variable k, we can construct a better model by breaking this segment into two parts, if Proposition 3.3.1 is satisfied.

Proof of Proposition 3.3.2. To begin with assume for a contradiction that $\boldsymbol{c}_p = \ell(\boldsymbol{c}_f)$. By Lemma A.1.1 one of the following two statements must be true, $\boldsymbol{c}_f^j \geq v$ or there exists a sequence of changepoint vectors,

$$\boldsymbol{c}_f = \boldsymbol{c}_{f,0}, \boldsymbol{c}_{f,1}, \ldots, \boldsymbol{c}_{f,g-1}, \boldsymbol{c}_{f,g}$$

such that

$$\boldsymbol{c}_{f,q-1} = \ell(\boldsymbol{c}_{f,q}) \text{ for } 1 \leq q \leq g, c_{f,q}^k = t \text{ for } 0 \leq q < g \text{ and } c_{f,g}^k \geq v.$$

In the latter case, proving that $\boldsymbol{c}_p \neq \ell(\boldsymbol{c}_f)$ is equivalent to showing that $\boldsymbol{c}_{f,g-1} \neq \ell(\boldsymbol{c}_{f,g})$. Note this is equivalent to the former case. Thus going forward we assume that $\boldsymbol{c}_f^k \geq v$.

Now let $\boldsymbol{c}$ be a changepoint vector such that $c^k = s$ and $c^j = c_p^j$ otherwise. By the same argument as in the proof of Proposition 3.3.1, we have that $\boldsymbol{c}_p, \boldsymbol{c}$ and $\boldsymbol{c}_f$ satisfy Lemma A.1.2. Similarly, let $\boldsymbol{c}_v$ be a changepoint vector such that $c_v^k = v$ and $c^j = c^j$ otherwise and note that $\boldsymbol{c}, \boldsymbol{c}_v$ and $\boldsymbol{c}_f$ satisfy Lemma A.1.2. Then by the following

chain of inequalities, we have a contradiction.

$$
\begin{aligned}
F(\boldsymbol{c}_f) &= F(\boldsymbol{c}_p) + \sum_{j=1}^{p} \left[I\left(c_p^j \neq c_f^j\right) \left(\mathcal{C}^j\left(c_p^j, c_f^j\right) + \alpha \right) \right] + m(\boldsymbol{c}_p, \boldsymbol{c}_f)\beta \\
&= F(\boldsymbol{c}_p) + \sum_{j \neq k} \left[I\left(c_p^j \neq c_f^j\right) \left(\mathcal{C}^j\left(c_p^j, c_f^j\right) + \alpha \right) \right] + m(\boldsymbol{c}_p, \boldsymbol{c}_f)\beta + \mathcal{C}^k(t, c_f^k) + \alpha \\
&= F(\boldsymbol{c}_p) + \sum_{j \neq k} \left[I\left(c^j \neq c_f^j\right) \left(\mathcal{C}^j\left(c, c_f\right) + \alpha \right) \right] + m(\boldsymbol{c}_p, \boldsymbol{c}_f)\beta + \mathcal{C}^k(t, c_f^k) + \alpha \\
&\geq F(\boldsymbol{c}_p) + \sum_{j \neq k} \left[I\left(c^j \neq c_f^j\right) \left(\mathcal{C}^j\left(c_p^j, c_f^j\right) + \alpha \right) \right] + m(\boldsymbol{c}_p, \boldsymbol{c}_f)\beta + \mathcal{C}^k(t, v) + \mathcal{C}^k(v, c_f^k) + \alpha \\
&> F(\boldsymbol{c}_p) + \sum_{j \neq k} \left[I\left(c^j \neq c_f^j\right) \left(\mathcal{C}^j\left(c_p^j, c_f^j\right) + \alpha \right) \right] + m(\boldsymbol{c}_p, \boldsymbol{c}_f)\beta + \\
&\mathcal{C}^k(t, s) + \mathcal{C}^k(s, v) + 2\alpha + 2\beta + \mathcal{C}^k(v, c_f^k) + \alpha \\
&> F(\boldsymbol{c}_p) + \mathcal{C}^k(t, s) + \alpha + (m(\boldsymbol{c}_p, \boldsymbol{c}_f) + 1)\beta + \\
&\sum_{j \neq k} \left[I\left(c^j \neq c_f^j\right) \left(\mathcal{C}^j\left(c_p^j, c_f^j\right) + \alpha \right) \right] + \mathcal{C}^k(s, v) + \alpha + \beta + \mathcal{C}^k(v, c_f^k) + \alpha \\
&\geq F(\boldsymbol{c}_p) + \mathcal{C}^k(t, s) + \alpha + (m(\boldsymbol{c}_p, \boldsymbol{c}))\beta + \\
&\sum_{j \neq k} \left[I\left(c^j \neq c_f^j\right) \left(\mathcal{C}^j\left(c_p^j, c_f^j\right) + \alpha \right) \right] + m(\boldsymbol{c}, \boldsymbol{c}_f)\beta + \mathcal{C}^k(s, v) + \alpha + \beta + \mathcal{C}^k(v, c_f^k) + \alpha \\
&\geq F(\boldsymbol{c}) + \mathcal{C}^k(s, v) + \alpha + (m(\boldsymbol{c}, \boldsymbol{c}_f) + 1)\beta + \sum_{j \neq k} \left[I\left(c^j \neq c_f^j\right) \left(\mathcal{C}^j\left(c_p^j, c_f^j\right) + \alpha \right) \right] + \mathcal{C}^k(v, c_f^k) + \alpha \\
&\geq F(\boldsymbol{c}) + \mathcal{C}^k(s, v) + \alpha + m(\boldsymbol{c}, \boldsymbol{c}_v)\beta + \sum_{j=1}^{p} \left[I\left(c_v^j \neq c_f^j\right) \left(\mathcal{C}^j\left(c_v^j, c_f^j\right) + \alpha \right) \right] + m(\boldsymbol{c}_v, \boldsymbol{c}_f)\beta \\
&\geq F(\boldsymbol{c}_v) + \sum_{j \neq k} \left[I\left(c^j \neq c_f^j\right) \left(\mathcal{C}^j\left(c^j, c_f^j\right) + \alpha \right) \right] + m(\boldsymbol{c}_v, \boldsymbol{c}_f)\beta \geq F(\boldsymbol{c}_f).
\end{aligned}
$$

□

Proof of Corollary 3.3.3. Firstly let $\boldsymbol{c}_f$ be a changepoint vector such that $\boldsymbol{c} \prec \boldsymbol{c}_f$. If $c_f^j \geq v$ Proposition 3.3.2 states that $\boldsymbol{c}$ is not the optimal prior changepoint vector. Therefore we can safely assume that $c_f^j < v$. Now since $\boldsymbol{c} \prec \boldsymbol{c}_f$ it must be the case that $c_f^k \geq v$. Therefore by Lemma A.1.1 there exists another changepoint $\boldsymbol{c}_*$ with penalised cost at least as good as $\boldsymbol{c}$. If the inequality is strict then $\boldsymbol{c}$ is not the optimal prior

changepoint vector. If there is equality, then $\boldsymbol{c}_*$ gives the same segmentation as $\boldsymbol{c}$ and so we set $\boldsymbol{c}_* = \ell(\boldsymbol{c}_f)$ completing the proof. □

A.4 Proofs for Section 3.3.2

Proof of Proposition 3.3.4. Firstly let $\boldsymbol{c}_o = \ell(\boldsymbol{c}_p)$ and $\boldsymbol{c}$ be a changepoint vector such that $c^k = c_o^k$ and $c^j = c_p^j$ for $j \neq k$. By the definition of $\ell(\boldsymbol{c}_p)$ either $c_o^k < s$ or $c_o^k = s$. If the former case holds, let $t = c_o^k$. If the latter case holds, since $s > 0$ there exists a sequence of vectors

$$\boldsymbol{c}_o = \boldsymbol{c}_{o,1}, \ldots, \boldsymbol{c}_{o,g}$$

such that

$$\boldsymbol{c}_{o,i} = \ell(\boldsymbol{c}_{o,i-1}) \text{ and } c_{o,g}^k < c_p^k.$$

Then let $t = c_{o,g}^k$. The rest of the proof demonstrates that $\boldsymbol{c}$ gives a better solution to the recursion for $\boldsymbol{c}_f$ then $\boldsymbol{c}_p$ and is the same for both cases.

The sequence of changepoint vectors $\boldsymbol{c}_o, \boldsymbol{c}, \boldsymbol{c}_f$ has less or equal unique changepoints than the sequence $\boldsymbol{c}_o, \boldsymbol{c}, \boldsymbol{c}_f$, so

$$m(\boldsymbol{c}_o, \boldsymbol{c}) + m(\boldsymbol{c}, \boldsymbol{c}_f) \leq m(\boldsymbol{c}_o, \boldsymbol{c}_p) + m(\boldsymbol{c}_p, \boldsymbol{c}_f).$$

Then

$$
\begin{aligned}
&F(\boldsymbol{c}_p) + \sum_{j=1}^{p} \left[I\left(c_p^j \neq c_f^j\right) \left(\mathcal{C}^j\left(c_p^j, c_f^j\right) + \alpha \right) \right] + m(\boldsymbol{c}_p, \boldsymbol{c}_f)\beta \\
&= F(\boldsymbol{c}_o) + \sum_{j=1}^{p} \left[I\left(c_o^j \neq c_p^j\right) \left(\mathcal{C}^j\left(c_o^j, c_p^j\right) + \alpha \right) \right] + m(\boldsymbol{c}_o, \boldsymbol{c}_p)\beta \\
&+ \sum_{j=1}^{p} \left[I\left(c_p^j \neq c_f^j\right) \left(\mathcal{C}^j\left(c_p^j, c_f^j\right) + \alpha \right) \right] + m(\boldsymbol{c}_p, \boldsymbol{c}_f)\beta \\
&= F(\boldsymbol{c}_o) + \mathcal{C}^k(t,s) + \mathcal{C}^k(s,v) + 2\alpha + \sum_{j \neq k} \left[I\left(c_o^j \neq c_p^j\right) \left(\mathcal{C}^j\left(c_o^j, c_p^j\right) + \alpha \right) \right] + m(\boldsymbol{c}_o, \boldsymbol{c}_p)\beta \\
&+ \sum_{j \neq k} \left[I\left(c_p^j \neq c_f^j\right) \left(\mathcal{C}^j\left(c_p^j, c_f^j\right) + \alpha \right) \right] + m(\boldsymbol{c}_p, \boldsymbol{c}_f)\beta \\
&> F(\boldsymbol{c}_o) + \mathcal{C}^k(t,v) + \alpha + \sum_{j \neq k} \left[I\left(c_o^j \neq c_p^j\right) \left(\mathcal{C}^j\left(c_o^j, c_p^j\right) + \alpha \right) \right] + m(\boldsymbol{c}_o, \boldsymbol{c}_p)\beta \\
&+ \sum_{j \neq k} \left[I\left(c_p^j \neq c_f^j\right) \left(\mathcal{C}^j\left(c_p^j, c_f^j\right) + \alpha \right) \right] + m(\boldsymbol{c}_p, \boldsymbol{c}_f)\beta \\
&\geq F(\boldsymbol{c}_o) + \sum_{j=1}^{p} \left[I\left(c_o^j \neq c^j\right) \left(\mathcal{C}^j\left(c_o^j, c^j\right) + \alpha \right) \right] + m(\boldsymbol{c}_o, \boldsymbol{c})\beta \\
&+ \mathcal{C}^k(t,v) + \alpha + \sum_{j \neq k} \left[I\left(c^j \neq c_f^j\right) \left(\mathcal{C}^j\left(c^j, c_f^j\right) + \alpha \right) \right] + m(\boldsymbol{c}, \boldsymbol{c}_f)\beta \\
&\geq F(\boldsymbol{c}) + \sum_{j=1}^{p} \left[I\left(c^j \neq c_f^j\right) \left(\mathcal{C}^j\left(c^j, c_f^j\right) + \alpha \right) \right] + m(\boldsymbol{c}, \boldsymbol{c}_f)\beta \geq F(\boldsymbol{c}_f).
\end{aligned}
$$

□

Proof of Proposition 3.3.5. We again consider two cases for this proof, the case where $c_f^j \neq s$ for all $1 \leq j \leq p$ and the complement. In the complement case, since $s < n$ there exists a sequence of vectors

$$\boldsymbol{c}_f = \boldsymbol{c}_{f,0}, \ldots, \boldsymbol{c}_{f,g}$$

such that

$$\boldsymbol{c}_{f,i-1} = \ell(\boldsymbol{c}_{f,i}) \text{ for } 1 \leq i \leq g \text{ and } \boldsymbol{c}_{f,g} \neq s \text{ for } 1 \leq j \leq p.$$

Now if we can show that $\boldsymbol{c}_{f,g}$ is not the optimal prior changepoint vector for any $\boldsymbol{c}_{f,g+1}$, then $\boldsymbol{c}_p$ is not an element of an optimal segmentation. Thus by letting $\boldsymbol{c}_p = \boldsymbol{c}_{f,g}$ and $\boldsymbol{c}_f = \boldsymbol{c}_{f,g+1}$, this case is equivalent to proving the case where $c_f^j \neq s$ for all $1 \leq j \leq p$.

Now let $\boldsymbol{c}_o = \ell(\boldsymbol{c}_p)$ and let $\boldsymbol{c}$ be a changepoint vector such that

$$c^j = \begin{cases} c_o^j \text{ if } c_p^j = s \\ c_f^j \text{ otherwise.} \end{cases}$$

The sequence of changepoint vectors $\boldsymbol{c}_o, \boldsymbol{c}_p, \boldsymbol{c}_f$ has exactly one more change then the sequence $\boldsymbol{c}_o, \boldsymbol{c}, \boldsymbol{c}_f$, so

$$m(\boldsymbol{c}_o, \boldsymbol{c}) + m(\boldsymbol{c}, \boldsymbol{c}_f) + 1 = m(\boldsymbol{c}_o, \boldsymbol{c}_p) + m(\boldsymbol{c}_p, \boldsymbol{c}_f).$$

Finally by the definition of Π_s, we have that for any subset of variables $\mathcal{J}$,

$$\sum_{j \in \mathcal{J}} [\mathcal{C}^j(c_o^j, s) + \mathcal{C}(s, c_f^j) + \alpha] + \beta > \sum_{j \in \mathcal{J}} \mathcal{C}^j(c_o^j, c_f^j).$$

Then

$$\begin{aligned}
& F(\boldsymbol{c}_p) + \sum_{j=1}^{p} \left[I\left(c_p^j \neq c_f^j\right) \left(\mathcal{C}^j\left(c_p^j, c_f^j\right) + \alpha \right) \right] + m(\boldsymbol{c}_p, \boldsymbol{c}_f)\beta \\
= F(\boldsymbol{c}_o) + & \sum_{j | c_p^j \neq s}^{p} I\left(c_o^j \neq c_p^j\right) \left[\mathcal{C}^j(c_o^j, c_p^j) + \alpha \right] + \sum_{j | c_p^j \neq s}^{p} I\left(c_p^j \neq c_f^j\right) \left[\mathcal{C}^j(c_p^j, c_f^j) + \alpha \right] + \\
& \sum_{j | c_p^j = s}^{p} \left[\mathcal{C}^j(c_p^j, s) + \alpha + \mathcal{C}^j(s, c_f^j) + \alpha \right] + m(\boldsymbol{c}_o, \boldsymbol{c}_p)\beta + m(\boldsymbol{c}_p, \boldsymbol{c}_f)\beta \\
> F(\boldsymbol{c}_o) + & \sum_{j | c_p^j \neq s}^{p} I\left(c_o^j \neq c_p^j\right) \left[\mathcal{C}^j(c_o^j, c^j) + \alpha \right] + \sum_{j | c_p^j \neq s}^{p} I\left(c_p^j \neq c_f^j\right) \left[\mathcal{C}^j(c^j, c_f^j) + \alpha \right] + \\
& \sum_{j | c_p^j = s}^{p} \left[\mathcal{C}^j(c^j, c_f^j) + \alpha \right] + m(\boldsymbol{c}_o, \boldsymbol{c})\beta + m(\boldsymbol{c}, \boldsymbol{c}_f)\beta \\
= F(\boldsymbol{c}_o) + & \sum_{j=1}^{p} I\left(c_o^j \neq c^j\right) \left[\mathcal{C}^j(c_o^j, c^j) + \alpha \right] + m(\boldsymbol{c}_o, \boldsymbol{c})\beta + \\
& \sum_{j=1}^{p} I\left(c^j \neq c_f^j\right) \left[\mathcal{C}^j(c^j, c_f^j) + \alpha \right] + \sum_{j | c_p^j = s}^{p} \left[\mathcal{C}^j(c^j, c_f^j) + \alpha \right] + m(\boldsymbol{c}, \boldsymbol{c}_f)\beta \\
\geq F(\boldsymbol{c}) + & \sum_{j=1}^{p} I\left(c^j \neq c_f^j\right) \left[\mathcal{C}^j(c^j, c_f^j) + \alpha \right] + m(\boldsymbol{c}, \boldsymbol{c}_f)\beta \geq F(\boldsymbol{c}_f).
\end{aligned}$$

□

Appendix B

Appendix for Chapter 4

B.1 Appendix

Lemma B.1.1. *As n tends to infinity with probability approaching one,*

$$\sum_{i=1}^{n} Y_i^2 + \log(n) \geq \hat{S}_n(\tau_1^0, \ldots, \tau_{m_0}^0).$$

Proof. To begin with note that

$$\hat{S}_n(\tau_1^0, \ldots, \tau_{m_0}^0) - S_n(\tau_1^0, \ldots, \tau_{m_0}^0) = \tag{B.1.1}$$

$$\sum_{j=1}^{m_0+1} \sum_{i=\tau_{j-1}+1}^{\tau_j} \left\{X_i - \bar{X}(\tau_{j-1}, \tau_j)\right\}^2 - \left\{X_i - \bar{X}(\tau_{m-1}, \tau_{m-1} + w)\right\}^2$$

$$= \sum_{j=1}^{m_0+1} \sum_{i=\tau_{j-1}+1}^{\tau_j} X_i^2 - 2X_i\bar{X}(\tau_{j-1}, \tau_{j-1} + w) + \bar{X}^2(\tau_{j-1}, \tau_{j-1} + w) -$$

$$X_i^2 + 2X_i\bar{X}(\tau_{j-1}, \tau_j) - \bar{X}^2(\tau_{j-1}, \tau_j)$$

$$= \sum_{j=1}^{m_0+1} l_i\big(\bar{X}(\tau_{j-1}, \tau_j) - \bar{X}(\tau_{j-1}, \tau_{j-1} + w)\big)^2 \tag{B.1.2}$$

where $l_j = \tau_j - \tau_{j-1}$. Now,

$$\bar{X}(\tau_{j-1}, \tau_j) - \bar{X}(\tau_{r-1}, \tau_{r-1} + w) \sim \mathcal{N}\left(0, \frac{1}{w} + \frac{1}{l_i - w}\right) \tag{B.1.3}$$

which along with (B.1.2) implies

$$\hat{S}_n(\tau_1^0, \ldots, \tau_{m_0}^0) - S_n(\tau_1^0, \ldots, \tau_{m_0}^0) \sim \sum_{r=1}^{m_0+1} l_i \left(\frac{\sigma^2}{w} + \frac{\sigma^2}{l_i - w}\right) \chi_1^2. \tag{B.1.4}$$

In other words the error at the true set of change points is distributed as a weighted chi-squared distribution. The weights are constant with respect to n,

$$l_i\left(\frac{\sigma^2}{w}+\frac{\sigma^2}{l_i-w}\right)=\frac{l_i}{n}\left(\frac{\sigma^2}{\frac{w}{n}}+\frac{\sigma^2}{\frac{l_i-w}{n}}\right)=q_i\left(\frac{\sigma^2}{t}+\frac{\sigma^2}{q_i-t}\right).$$

Then trivially we have that with probability approaching one,

$$\hat{S}_n(\tau_1^0,\ldots,\tau_m^0)-S_n(\tau_1^0,\ldots,\tau_m^0)\leq\log n.$$

Hence with probability approaching one

$$\sum_{i=1}^{n}Y_i^2+\log(n)-\hat{S}_n(\tau_1^0,\ldots,\tau_{m_0}^0)=$$

$$\sum_{i=1}^{n}Y_i^2-S_n(\tau_1^0,\ldots,\tau_{m_0}^0)+S_n(\tau_1^0,\ldots,\tau_{m_0}^0)-\hat{S}_n(\tau_1^0,\ldots,\tau_{m_0}^0)+\log n$$

$$\geq S_n(\tau_1^0,\ldots,\tau_{m_0}^0)-\hat{S}_n(\tau_1^0,\ldots,\tau_{m_0}^0)+\log n\geq 0.$$

This completes the proof. □

Note that since the error term is independent of n the $\log(n)$ rate is more for convenience then anything else and it could be replaced by any unbounded function of n.

Yao provides two results that we use to prove the consistency of our estimator which we present here for clarity.

Theorem B.1.2 (Yao's Lemma)**.** *Suppose that $Z_1,\ldots,Z_n$ are iid normal with common mean 0 and variance σ^2. Then for any $\epsilon>0$, as $n\to\infty$,*

$$\Pr\left\{\max_{0\leq i<j\leq n}(Z_{i+1}+\ldots Z_j)^2/(j-i)>2(1+\epsilon)\sigma^2\log n\right\}\to 0.$$

Theorem B.1.3 (Yao's bound)**.** *For every $m(m_0<m<m_U)$ and for any ϵ with probability approaching one,*

$$0\leq\sum_{i=1}^{n}Y_i^2-n\hat{\sigma}_m^2\leq\left\{\epsilon+(m-m_0-1)2(1+\epsilon)\right\}\sigma^2\log n$$

We also need two other results which we prove below. For both proofs we restrict our attention to the subset, A_n, which denotes the datasets of size n such that

$$\sum_{i=1}^{n} Y_i^2 + \log n \geq \hat{S}_n(\tau_1^0, \ldots, \tau_{m_0}^0).$$

Proving convergence in probability follows from the fact that the measure of this set tends to one by B.1.1.

Lemma B.1.4. *As $n \to \infty$,*

$$0 \leq \sum_{i=1}^{n} Y_i^2 - n\hat{\sigma}_{w,m_0}^2 = \mathcal{O}_p(\log n).$$

Proof. Let $\lambda = \lfloor n/w \rfloor$ and choose $w_1, \ldots, w_\lambda$ such that

$$\max_{0 \leq i \leq \lambda} w_{i+1} - w_i < \omega.$$

Then on the set A_n we have that

$$\sum_{i=1}^{n} Y_i^2 + \log n \geq \hat{S}_n(\hat{\tau}_1, \ldots, \hat{\tau}_{m_0}), S_n(\tau_1^0, \ldots, \tau_{m_0}^0) \geq S_n(\hat{\tau}_1, \ldots, \hat{\tau}_{m_0}, \tau_1^0, \ldots, \tau_{m_0}^0, w_1, \ldots, w_\lambda). \tag{B.1.5}$$

However,

$$S_n(\hat{\tau}_1, \ldots, \hat{\tau}_{m_0}, \tau_1^0, \ldots, \tau_{m_0}^0, w_1, \ldots, w_\lambda) = \hat{S}_n(\hat{\tau}_1, \ldots, \hat{\tau}_{m_0}, \tau_1^0, \ldots, \tau_{m_0}^0, w_1, \ldots, w_\lambda),$$

so we have that

$$\sum_{i=1}^{n} Y_i^2 + \log n \geq \hat{S}_n(\hat{\tau}_1, \ldots, \hat{\tau}_{m_0}, \tau_1^0, \ldots, \tau_{m_0}^0, w_1, \ldots, w_\lambda) \tag{B.1.6}$$

Now let $v(1,s) < \cdots < v(V(s),s)$ be the elements of $\{\hat{\tau}_1, \ldots, \hat{\tau}_{m_0}\} \cup \{w_1, \ldots, w_\lambda\}$ which are greater then τ_{s-1}^0 but less than τ_s^0. Then,

$$\begin{aligned}
&\hat{S}_n(\hat{\tau}_1, \ldots, \hat{\tau}_{m_0}, \tau_1^0, \ldots, \tau_{m_0}^0, w_1, \ldots, w_\lambda) \\
&= \sum_{s=1}^{m_0} \sum_{k=1}^{V(s)+1} \sum_{i=v(k-1,s)}^{v(k,s)} \left\{ X_i - \bar{X}\left(v(k-1,s), v(k,s)\right) \right\}^2 \\
&= \sum_{i=1}^{n} Y_i^2 - \sum_{s=1}^{m_0} \sum_{k=1}^{V(s)+1} \left(v(k,s) - v(k-1,s)\right) \left(\bar{Y}^2(v(k-1,s), v(k,s)) \right) \\
&\geq \sum_{i=1}^{n} Y_i^2 - \sum_{s=1}^{m_0} (V(s)+1) \max_{\tau_{r-1}^0 \leq i < j \leq \tau_r^0} \left((j-i)\bar{Y}^2(i,j) \right).
\end{aligned}$$

Now by Yao's Lemma we have that the final term is $\mathcal{O}(\log n)$, so

$$\hat{S}_n(\hat{\tau}_1, \ldots, \hat{\tau}_m, \tau_1^0, \ldots, \tau_m, w_1, \ldots, w_\lambda) \geq \sum_{i=1}^{n} Y_i^2 - \mathcal{O}_p(\log n). \tag{B.1.7}$$

Hence the lemma follows from (B.1.5) and (B.1.7). □

Lemma B.1.5. *For every $m < m_0$ there exists $\epsilon > 0$ such that $\Pr(\sigma^2_{w,m} > \sigma^2 + \epsilon) \to 1$ as $n \to \infty$.*

Proof. Let $\delta > 0$ be such that $q_j + 2\delta < q_{j+1} - 2\delta$ for $j = 0, \ldots, m_0$ and $2[n\delta] < w$. Similarly let

$$B_j(n, \delta) = \left\{(\tau_1, \ldots, \tau_m) : 0 < \tau_1 < \cdots < \tau_m, \text{ and } |\tau_j^0 - \tau_s| > [n\delta] \text{ for } 1 \leq s \leq m\right\}.$$

Since $m < m_0$ we have that $(\hat{\tau}_1, \ldots, \hat{\tau}_m) \in B_j(n, \delta)$ for some $j = 1, \ldots, m$. Then we only need to demonstrate that for each $j = 1, \ldots, m$ we have that

$$\min_{(\tau_1, \ldots, \tau_m) \in B_j(n,\delta)} \frac{\hat{S}(\tau_1, \ldots, \tau_m)}{n} > \sigma^2 + \epsilon.$$

Let $w^1 = \tau_j^0 - [\omega/2]$ and $w^2 = \tau_j^0 + [\omega/2]$. Then choose $w_3, \ldots, w_\lambda + 1$ such that

$$\max_{0 \leq i \leq \lambda} w_{i+1} - w_i < \omega.$$

Then we have that

$$\hat{S}_n(\hat{\tau}_1, \ldots, \hat{\tau}_m) \geq \hat{S}_n(\hat{\tau}_1, \ldots, \hat{\tau}_m, \tau_1^0, \ldots, \tau_{j-1}^0, \tau_j^0 - [n\delta], \tau_j + [n\delta], \tau_{j+1}, \ldots, \tau_{m_0}, w_1, \ldots, w_{\lambda+1})$$

Note since every segment on the right hand side has max length w we have that

$$\begin{aligned} &\hat{S}_n(\hat{\tau}_1, \ldots, \hat{\tau}_m, \tau_1^0, \ldots, \tau_{j-1}^0, \tau_j^0 - [n\delta], \tau_j + [n\delta], \tau_{j+1}, \ldots, \tau_{m_0}, w_3, \ldots, w_{\lambda+1}) \\ = &S_n(\hat{\tau}_1, \ldots, \hat{\tau}_m, \tau_1^0, \ldots, \tau_{j-1}^0, \tau_j^0 - [n\delta], \tau_j + [n\delta], \tau_{j+1}, \ldots, \tau_{m_0}, w_3, \ldots, w_{\lambda+1}) \end{aligned}$$

We can break up the right hand side of this equation into segments of common mean along with one segment with two means. More explicitly it can be expressed as the sum of $T_1 + \cdots + T_{m_0+2}$ where T_s $(s = 1, \ldots j-1, j+2, \ldots, m_0+2)$ denotes the sum of squares relating to data $X_i(\tau_{s-1}^0 < i \leq \tau_s^0)$, T_j is the sum of squares involving $X_i(\tau_{j-1}^0 < i \leq \tau_j^0 - [n\delta])$, T_{j+1} is the sum of squares involving $X_i(\tau_j^0 + [n\delta] < i \leq \tau_{j+1}^0)$

and T_{m_0+2} is the sum relating to $X_i(\tau_j^0 - [n\delta] < i \leq \tau_j^0 + [n\delta])$. For the sum of squares involving homogenous segments i.e. $T_s(s = 1, \ldots j-1, j+2, \ldots, m_0+2)$ using the same argument as Lemma 2 we have that,

$$\begin{aligned}\sum_{i=\tau_s^0+1}^{\tau_s^0} Y_i^2 + \geq T_s \geq & \sum_{i=\tau_s^0+1}^{\tau_s^0} Y_i^2 - (m_0+\lambda+1) \max_{\tau_s^0+1<i,j\leq\tau_s^0} (j-i)\bar{Y}^2(i,j) \\ & \geq \sum_{i=\tau_s^0+1}^{\tau_s^0} Y_i^2 - \mathcal{O}_p(\log n).\end{aligned}$$

Hence we have that T_s converges uniformly to $\sigma^2(q_s)$ on $(\tau_1, \ldots, \tau_m) \in B_j(n, \delta)$. Similarly we have that T_j and T_{j+1} converge uniformly to $\sigma^2(q_j - \delta)$ and $\sigma^2(q_{j+1} - \delta)$ respectively. Then we only need to show that T_{m_0+2} converges to something larger then $2\delta\sigma^2$. Now

$$\begin{aligned}T_{m_0+2} &= \sum_{i=\tau_j^0-[n\delta]+1}^{\tau_j^0+[n\delta]} \left\{X_i - \bar{X}(\tau_j^0 - [n\delta], \tau_j^0 + [n\delta])\right\}^2 \\ &= \sum_{i=\tau_j^0-[n\delta]+1}^{\tau_j^0} \left\{Y_i - \bar{Y}(\tau_j^0 - [n\delta], \tau_j^0 + [n\delta]) + \frac{\mu_j^0 - \mu_{j+1}^0}{2}\right\}^2 \\ &+ \sum_{i=\tau_j^0+1}^{\tau_j^0+[n\delta]} \left\{Y_i - \bar{Y}(\tau_j^0 - [n\delta], \tau_j^0 + [n\delta]) + \frac{\mu_{j+1}^0 - \mu_j^0}{2}\right\}^2 \\ &= \sum_{i=\tau_j^0-[n\delta]+1}^{\tau_j^0+[n\delta]} \left\{Y_i - \bar{Y}(\tau_j^0 - [n\delta], \tau_j^0 + [n\delta])\right\}^2 + (2[n\delta])\left(\frac{\mu_{j+1}^0 - \mu_j^0}{2}\right)^2.\end{aligned}$$

So T_{m_0+2}/n converges to $2\delta\{\sigma^2 + (\mu_{j+1}^0 - \mu_j^0)^2/4\}$. Therefore

$$\begin{aligned}&\min_{(\tau_1,\ldots,\tau_m)\in B_j(n,\delta)} \frac{\hat{S}(\tau_1, \ldots, \tau_m)}{n} \\ \geq &\min_{(\tau_1,\ldots,\tau_m)\in B_j(n,\delta)} \frac{\hat{S}_n(\tau_1, \ldots, \tau_m, \tau_1^0, \ldots, \tau_{j-1}^0, w_1, w_2, \tau_{j+1}^0, \ldots, \tau_{m_0}^0, w_3, \ldots, w_{\lambda+1})}{n} \\ \to &\sigma^2 + \delta(\mu_{j+1}^0 - \mu_j^0)^2/2\end{aligned}$$

in probability, completing the proof. □

Lemma B.1.6. *Let*

$$\sum_{j=1}^{p} \hat{\mathcal{D}}_j(t, s) + \sum_{j=1}^{p} \hat{\mathcal{D}}_j(s, T) - p\alpha \leq \sum_{j=1}^{p} \hat{\mathcal{D}}_j(t, T)$$

Proof. For a segment beginning at time p and ending at time q, for each $\mathcal{S}_j$ we have two possibilities, $\mathcal{S}_j = 1$ or 0. If $\mathcal{S}_j = 1$ we have that

$$\hat{D}_j(p,q) = D_j(p,q) + \alpha.$$

On the other hand if $\mathcal{S}_j = 0$ we know that

$$\hat{D}_j(p,q) < D_j(p,q) + \alpha.$$

Case 1: (1,0,0)

$$\begin{aligned}\hat{D}_j(t,T) - \hat{D}_j(t,s) - \hat{D}_j(s,T) + \alpha &= D_j(t,s|\theta) + D_j(s,T|\theta) - D_j(t,s) - \alpha - D_j(s,T|\theta(t,s)) + \alpha \\ &= [D_j(t,s|\theta) - D_j(t,s) - \alpha] + [D_j(s,T|\theta) - D_j(s,T|\theta(t,s)) + \alpha] \\ &\geq 0\end{aligned}$$

Case 2: (0,1,0)

$$\begin{aligned}\hat{D}_j(t,T) - \hat{D}_j(t,s) - \hat{D}_j(s,T) + \alpha &= D_j(t,s|\theta) + D_j(s,T|\theta) - D_j(t,s|\theta) - D_j(s,T) - \alpha + \alpha \\ &= [D_j(t,s|\theta) - D_j(t,s|\theta)] + [D_j(s,T|\theta) - D_j(s,T)] \\ &\geq 0\end{aligned}$$

Case 3: (0,0,1)

$$\begin{aligned}&\hat{D}_j(t,T) - \hat{D}_j(t,s) - \hat{D}_j(s,T) + \alpha = \\ &\qquad D_j(t,s|\theta(t,T)) + D_j(s,T|\theta(t,T)) + \alpha - D_j(t,s|\theta) - D_j(s,T|\theta) + \alpha \\ &\qquad = [D_j(t,s|\theta(t,T)) + \alpha - D_j(t,s|\theta)] + [D_j(s,T|\theta(t,T)) + \alpha - D_j(s,T|\theta)] \\ &\qquad \geq [D_j(t,s|\theta(t,s)) + \alpha - D_j(t,s|\theta)] + [D_j(s,T|\theta(s,T)) + \alpha - D_j(s,T|\theta)] \\ &\qquad \geq 0\end{aligned}$$

Case 4: (1,1,0)

$$\begin{aligned}\hat{D}_j(t,T) - \hat{D}_j(t,s) - \hat{D}_j(s,T) + \alpha &= D_j(t,s|\theta) + D_j(s,T|\theta) - D_j(t,s) - \alpha - D_j(s,T) - \alpha + \alpha \\ &= [D_j(t,s|\theta) - D_j(t,s) - \alpha] + [D_j(s,T|\theta) - D_j(s,T)] \\ &\geq 0\end{aligned}$$

Case 5: (1,0,1)

$$\begin{aligned}\hat{D}_j(t,T) - \hat{D}_j(t,s) - \hat{D}_j(s,T) + \alpha = \\ & D_j(t,s|\theta(t,T)) + D_j(s,T|\theta(t,T)) + \alpha - D_j(t,s) - \alpha - D_j(s,T|\theta(t,s)) + \alpha \\ & = [D_j(t,s|\theta(t,T)) - D_j(t,s)] + [D_j(s,T|\theta(t,T)) + \alpha - D_j(s,T|\theta(t,s))] \\ & \geq [D_j(t,s) - D_j(t,s|\theta)] + [D_j(s,T) + \alpha - D_j(s,T|\theta(t,s))] \\ & \geq 0\end{aligned}$$

Case 6: (0,1,1)

$$\begin{aligned}\hat{D}_j(t,T) - \hat{D}_j(t,s) - \hat{D}_j(s,T) + \alpha = \\ & D_j(t,s|\theta(t,T)) + D_j(s,T|\theta(t,T)) + \alpha - D_j(t,s|\theta) - D_j(s,T) - \alpha + \alpha \\ & = [D_j(t,s|\theta(t,T)) + \alpha - D_j(t,s|\theta)] + [D_j(s,T|\theta(t,T)) - D_j(s,T)] \\ & \geq [D_j(t,s) + \alpha - D_j(t,s|\theta)] + [D_j(s,T|\theta(t,T)) - D_j(s,T)] \\ & \geq 0\end{aligned}$$

Case 7: (1,1,1)

$$\begin{aligned}\hat{D}_j(t,T) - \hat{D}_j(t,s) - \hat{D}_j(s,T) + \alpha = \\ & D_j(t,s|\theta(t,T)) + D_j(s,T|\theta(t,T)) + \alpha - D_j(t,s) - \alpha - D_j(s,T) - \alpha + \alpha \\ & \geq 0\end{aligned}$$

Case 8: (0,0,0)

$$\hat{D}_j(t,T) - \hat{D}_j(t,s) - \hat{D}_j(s,T) + \alpha = D_j(t,T|\theta) - D_j(t,T|\theta) + \alpha = 0$$

□

Lemma B.1.7. *Let E_k and $I_{t,k}$ be defined as in the proof of Theorem 4.3.4. Furthermore let*

$$L := \lim_{n\to\infty} L_n = \lim_{n\to\infty} 1 + \sum_{j=1}^{n-1} E_j$$

and assume that the conditions (A1)-(A4) defined in Theorem 4.3.4 hold. Then L is bounded.

Proof. Firstly by choosing $t = k$ we have that E_k is the probability that $I_{k,k} = 1$ i.e. the probability that a changepoint at time zero has not been pruned after observing the jth observation. For this to be true we require that,

$$C(0,k) - 2p\alpha \leq \hat{F}(k)$$

Let m_k denote the true number of changepoints prior to time k, and $\tau_1, \ldots, \tau_{m_k}$ their locations. Again for simplicity we have that $\tau_0 = 0$ and $\tau_{m_k+1} = j$. Now

$$\hat{F}(k) \leq \sum_{i=1}^{m_k+1} \sum_{j=1}^{p} \left[\mathcal{D}^j(\tau_{i-1}, \tau_i) + \alpha\right] + \beta,$$

so

$$E_j \leq \Pr\left(C(0,j) - 2p\alpha \leq \sum_{i=1}^{m_k+1} \sum_{j=1}^{p} \left[\mathcal{D}^j(\tau_{i-1}, \tau_i) + \alpha\right] + \beta\right).$$

Now define θ_i^j to be the value of the parameter associated with the true segment of observation i for variable j; and

$$\tilde{\theta}_i := \arg\max_{\theta^j} \sum_{r=\tau_{l-1}+1}^{\tau_l} \log f^j(y_k|\theta^j),$$

where l is such that $\tau_{l-1} \leq i \leq \tau_l$. Now $C(0,k) = -\sum_{j=1}^{p} \sum_{i=1}^{k} \log f^j(X_i^j|\hat{\theta}_k^j)$ where $\hat{\theta}_k^j$ is the maximum likelihood estimate for parameter θ^j given data $X_{1:k}^j$ under an assumption of a single segment. Similarly we have that

$$\sum_{i=1}^{m_k+1} \sum_{j=1}^{p} \mathcal{D}^j(\tau_{i-1}, \tau_i) = -\sum_{j=1}^{p} \sum_{i=1}^{k} \log f^j(y_k|\tilde{\theta}^j).$$

So we can write

$$\overbrace{C(0,k) - 2p\alpha - \sum_{i=1}^{m_k+1} \sum_{j=1}^{p} \left[\mathcal{D}^j(\tau_{i-1}, \tau_i) + \alpha\right] + \beta}^{A_k} = \overbrace{\left[\sum_{j=1}^{p} \sum_{i=1}^{k} \log f^j(y_i|\theta^{*j}) - \log f^j(y_i|\hat{\theta}^k)\right]}^{B_k} +$$

$$\overbrace{\left[\sum_{j=1}^{p} \sum_{i=1}^{k} \log f^j(y_i|\theta_i^j) - \log f^j(y_i|\theta^{*j})\right]}^{D_k} - (m_j + 1)(\beta + p\alpha) - 2p\alpha +$$

$$\overbrace{\left[\sum_{j=1}^{p} \sum_{i=1}^{k} \log f^j(y_i|\tilde{\theta}_i^j) - \log f^j(y_i|\theta^j)\right]}^{R_k}.$$

First note that $R_k \leq 0$. So $E_k = \Pr(A_k \leq 0) \leq \Pr(B_k + D_k \leq 0)$. We can bound this probability using Markov's inequality.

By (A1), and using that the expected number of changepoints is related to the expected segment length, $\mathbb{E}(M_k) = j/\mathbb{E}(S) + \mathcal{O}(j)$ (elementary renewal theorem), we have

$$\mathbb{E}(B_k + D_k) = \mathbb{E}\left(\left[\sum_{j=1}^{p}\sum_{i=1}^{k} \log f^j(y_i|\theta_i^j) - \log f^j(y_i|\theta^{*j})\right]\right) - (\beta + p\alpha)\frac{k}{\mathbb{E}(S)} + \mathcal{O}(k).$$

Thus, using (A4), we have that there exists $c > 0$ such that for sufficiently large k

$$\mathbb{E}(B_k + D_k) > ck.$$

Let $B_k^j = \log f^j(y_i|\theta^{*j}) - \log f^j(y_i|\hat{\theta}^k)$, $D_k^j = \log f^j(y_i|\theta_i^j) - \log f^j(y_i|\theta^{*j})$, $B_k^* = B_k - \mathbb{E}(B_k)$ and $D_k^* = D_k - \mathbb{E}(D_k)$. Then we have that

$$\mathbb{E}((B_k^* + D_k^*)^4) \leq \mathbb{E}\left(\left(\sum_{j=1}^{p} B_{k,j}^* + D_{k,j}^*\right)^4\right) \tag{B.1.8}$$

Killick et al. demonstrate that for any finite k there exists a constant $K_j < \infty$ such that $\mathbb{E}(B_{k,j} + D_{k,j})^4 < K_j k^2$ By Minkowski's inequality we have that

$$\mathbb{E}((B_k^* + D_k^*)^4) = \mathbb{E}\left(\left(\sum_{j=1}^{p} B_{k,j}^* + D_{k,j}^*\right)^4\right) \leq \mathbb{E}\left(\left(p\max_{1\leq j\leq p} B_{k,j}^* + D_{k,j}^*\right)^4\right) \leq p^4 \max_{1\leq j\leq p} K_j k^2.$$

Now using Markov's inequality we have, for k large enough that $\mathbb{E}(B_k + D_k) > ck$

$$E_k \leq \Pr(B_k + D_k \leq 0) \leq \Pr\left(|B_k^* + D_k^*| \geq \mathbb{E}(B_k + D_k)\right) \leq \frac{\mathbb{E}\left((B_k^* + D_k^*)^4\right)}{\mathbb{E}(B_k + D_k)^4} \leq \frac{Tk^2}{c^4k^4},$$

where $T = \max_{1\leq j\leq p} K_j$. Thus we have that $E_k = \mathcal{O}(k^{-2})$, and hence $L = \lim_{n\to\infty}\sum_{k=1}^{n} E_k$ is finite, as required. □

Appendix C

Appendix for Chapter 6

C.1 Auxillary Results

The results in this section are required for the proof of Theorem 6.3.1.

Lemma C.1.1. *Let* $\boldsymbol{\gamma} := (\gamma_1, \gamma_2)$ *and* f_1 *be the real valued function*

$$f_1(x) := (1-x)^2.$$

Then

$$\lim_{r\downarrow 1} \frac{1}{4\pi i} \oint_{|z|=1} f\left(\frac{|1+h\xi|^2}{(1-\gamma_2)^2}\right)\left[\frac{1}{\xi - r^{-1}} + \frac{1}{\xi + r^{-1}} - \frac{2}{\xi + \frac{\gamma_2}{h}}\right] d\xi = 2K_3\left(1 - \frac{\gamma_2^2}{h^2}\right) + \frac{2K_2\gamma_2}{h}$$

where

$$K_2 = \frac{2h(1+h^2)}{(1-\gamma_2)^4} - \frac{2h}{(1-\gamma_2)^2}, \;\; K_3 = \frac{h^2}{(1-\gamma_2)^4}.$$

Proof. Firstly we have that

$$\begin{aligned}
f_1\left(\frac{|1+h\xi|^2}{(1-\gamma_2)^2}\right) &= \left(1-\frac{(1+h\xi)(1+h\bar{\xi})}{(1-\gamma_2)^2}\right)^2 \\
&= 1-2\frac{(1+h\xi)(1+h\bar{\xi})}{(1-\gamma_2)^2}+\frac{(1+h\xi)^2(1+h\bar{\xi})^2}{(1-\gamma_2)^4} \\
&= 1-2\frac{1+h\xi+h\bar{\xi}+h}{(1-\gamma_2)^2} \\
&+\frac{(1+4h^2+h^4)+2h(1+h^2)\xi+2h(1+h^2)\bar{\xi}+h^2\xi^2+h^2\bar{\xi}^2}{(1-\gamma_2)^4} \\
&= \left(1-2\frac{1+h}{(1-\gamma_2)^2}+\frac{1+4h^2+h^4}{(1-\gamma_2)^4}\right)+\left(\frac{2h(1+h^2)}{(1-\gamma_2)^4}-\frac{2h}{(1-\gamma_2)^2}\right)\xi+ \\
&\left(\frac{2h(1+h^2)}{(1-\gamma_2)^4}-\frac{2h}{(1-\gamma_2)^2}\right)\bar{\xi}+\left(\frac{h^2}{(1-\gamma_2)^4}\right)\xi^2+\left(\frac{h^2}{(1-\gamma_2)^4}\right)\bar{\xi}^2 \\
&= K_1+K_2\xi+K_2\bar{\xi}+K_3\xi^2+K_3\bar{\xi}^2
\end{aligned}$$

Then

$$\begin{aligned}
&\frac{1}{2\pi i}\oint_{|\xi|=1} f_1\left(\frac{|1+h\xi|^2}{(1-\gamma_2)^2}\right)\left[\frac{1}{\xi-r^{-1}}+\frac{1}{\xi+r^{-1}}-\frac{2}{\xi+\frac{\gamma_2}{h}}\right] \\
&= \frac{1}{2\pi i}\oint_{|\xi|=1}\left(K_1+K_2\xi+K_3\bar{\xi}+K_4\xi^2+K_5\bar{\xi}^2\right)\left[\frac{1}{\xi-r^{-1}}+\frac{1}{\xi+r^{-1}}-\frac{2}{\xi+\frac{\gamma_2}{h}}\right]
\end{aligned}$$

where $\bar{x}$ is the conjugate of x. By linearity of the integral we can handle each term separately. We can now evaluate the integral using the Cauchy Residue theorem. Note the first term is a constant function with respect to ξ and thus cancels out. Then

$$\begin{aligned}
\frac{1}{2\pi i}\oint_{|\xi|=1} K_2\xi\left[\frac{1}{\xi-r^{-1}}+\frac{1}{\xi+r^{-1}}-\frac{2}{\xi+\frac{\gamma_2}{h}}\right] &= K_2\left(r^{-1}-r^{-1}+\frac{2\gamma_2}{h}\right)=\frac{2K_2\gamma_2}{h} \\
\frac{1}{2\pi i}\oint_{|\xi|=1} K_3\xi^2\left[\frac{1}{\xi-r^{-1}}+\frac{1}{\xi+r^{-1}}-\frac{2}{\xi+\frac{\gamma_2}{h}}\right] &= K_3(r^{-2}+r^{-2}-2\frac{\gamma_2^2}{h^2})=2K_3\left(1-\frac{\gamma_2^2}{h^2}\right) \\
\frac{1}{2\pi i}\oint_{|\xi|=1}\frac{K_2}{\xi}\left[\frac{1}{\xi-r^{-1}}+\frac{1}{\xi+r^{-1}}-\frac{2}{\xi+\frac{\gamma_2}{h}}\right] &= K_2(-r+r-\frac{2h}{\gamma_2}+r-r+\frac{2h}{\gamma_2})=0 \\
\frac{1}{2\pi i}\oint_{|\xi|=1}\frac{K_3}{\xi^2}\left[\frac{1}{\xi-r^{-1}}+\frac{1}{\xi+r^{-1}}-\frac{2}{\xi+\frac{\gamma_2}{h}}\right] &= K_3\left(-r^2-r^2+\frac{2h^2}{\gamma_2^2}+r^2+r^2-\frac{2h^2}{\gamma_2^2}\right)=0
\end{aligned}$$

Summing these gives

$$2K_3\left(1-\frac{\gamma_2^2}{h^2}\right)+\frac{2K_2\gamma_2}{h}. \tag{C.1.1}$$

□

Lemma C.1.2. *Let* $\boldsymbol{\gamma} := (\gamma_1, \gamma_2)$ *and* f_1 *be the real valued function*

$$f_1(x) := (1-x)^2.$$

Then

$$-\lim_{r\downarrow 1}\frac{2}{4\pi^2}\oint_{|\xi_1|=1}\oint_{|\xi_2|=1}\frac{1}{(\xi_1 - r\xi_2)^2}f_1\left(\frac{|1+h\xi_1|^2}{(1-\gamma_2)^2}\right)f_1\left(\frac{|1+h\xi_2|^2}{(1-\gamma_2)^2}\right)d\xi_2 d\xi_1 = K_2^2 + 2K_3^2$$

where

$$K_2 = \frac{2h(1+h^2)}{(1-\gamma_2)^4} - \frac{2h}{(1-\gamma_2)^2},\ \ K_3 = \frac{h^2}{(1-\gamma_2)^4}.$$

Proof. Using a similar strategy to the Lemma C.1.1 we have that

$$\begin{aligned}
&-\frac{1}{4\pi^2}\oint_{|\xi_1|=1}\oint_{|\xi_2|=1}\frac{f_1\left(\frac{|1+h\xi_1|^2}{(1-\gamma_2)^2}\right)f_1\left(\frac{|1+h\xi_2|^2}{(1-\gamma_2)^2}\right)}{(\xi_1 - r\xi_2)^2}d\xi_1 d\xi_2\\
&= -\frac{1}{4\pi^2}\oint_{|\xi_2|=1} f_1\left(\frac{|1+h\xi_2|^2}{(1-\gamma_2)^2}\right)\oint_{|\xi_1|=1}\frac{\left(K_1 + K_2\xi_1 + K_2\xi_1^{-1} + K_3\xi_1^2 + K_3\xi_1^{-2}\right)}{(\xi_1 - r\xi_2)^2}d\xi_1 d\xi_2\\
&= -\frac{2\pi i}{4\pi^2}\oint_{|\xi_2|=1} f_1\left(\frac{|1+h\xi_2|^2}{(1-\gamma_2)^2}\right)\left(\frac{K_2}{r^2\xi_2^2} + \frac{2K_3}{r^3\xi_2^3}\right)d\xi_2\\
&= -\frac{2\pi i}{4\pi^2}\oint_{|\xi_2|=1}\left(K_1 + K_2\xi_2 + K_2\xi_2^{-1} + K_3\xi_2^2 + K_3\xi_2^{-2}\right)\left(\frac{K_2}{r^2\xi_2^2} + \frac{2K_3}{r^3\xi_2^3}\right)d\xi_2\\
&= -\frac{2\pi i}{4\pi^2}\oint_{|\xi_2|=1}\left(K_1 + K_2\xi_2 + K_3\xi_2^2\right)\left(\frac{K_2}{r^2\xi_2^2} + \frac{2K_3}{r^3\xi_2^3}\right) + \left(K_2\xi_2^{-1} + K_3\xi_2^{-2}\right)\left(\frac{K_2}{r^2\xi_2^2} + \frac{2K_3}{r^3\xi_2^3}\right)d\xi_2\\
&= -\frac{2\pi i}{4\pi^2}\oint_{|\xi_2|=1}\left(K_1 + K_2\xi_2 + K_3\xi_2^2\right)\left(\frac{K_2}{r^2\xi_2^2} + \frac{2K_3}{r^3\xi_2^3}\right) + \left(\frac{K_2^2}{r^2\xi_2^3} + \frac{2K_2K_3}{r^2\xi_2^4} + \frac{K_2K_3}{r^2\xi_2^4} + \frac{2K_2K_3}{r^2\xi_2^5}\right)d\xi_2
\end{aligned}$$

Now by the Cauchy Residue Theorem, we have that

$$\oint_{|\xi_2|=1}\frac{K_2^2}{r^2\xi_2^3} + \frac{2K_2K_3}{r^2\xi_2^4} + \frac{K_2K_3}{r^2\xi_2^4} + \frac{2K_2K_3}{r^2\xi_2^5}d\xi_2 = 0 \text{ and } \oint_{|\xi_2|=1}\left(\frac{K_2}{r^2\xi_2^2} + \frac{2K_3}{r^3\xi_2^3}\right)d\xi_2 = 0,$$

as these expressions can be written as a constant function times a pole of order higher than two. Now

$$\begin{aligned}
&-\frac{2\pi i}{4\pi^2}\oint_{|\xi_2|=1}\left(K_2\xi_2 + K_3\xi_2^2\right)\left(\frac{K_2}{r^2\xi_2^2} + \frac{2K_3}{r^3\xi_2^3}\right)d\xi_2\\
&= \frac{2\pi i}{4\pi^2}\oint_{|\xi_2|=1}\left(\frac{K_2^2}{r^2\xi_2} + \frac{2K_3^2}{r^3\xi_2}\right)d\xi_2 + \frac{2\pi i}{4\pi^2}\oint_{|\xi_2|=1}\left(\frac{2K_2K_3}{r^3\xi_2^2} + \frac{K_2K_3}{r^2}\right)d\xi_2\\
&= \frac{2\pi i}{4\pi^2}\oint_{|\xi_2|=1}\left(\frac{K_2^2}{r^2\xi_2} + \frac{2K_3^2}{r^3\xi_2}\right)d\xi_2 = \frac{K_2^2}{r^2} + \frac{2K_3^2}{r^3}
\end{aligned}$$

Then taking the limit as $r \downarrow 1$ completes the proof. □

Lemma C.1.3. *Let* $\boldsymbol{\gamma} := (\gamma_1, \gamma_2)$ *and* f_1, f_2 *be the real valued function*

$$f_1(x) := (1-x)^2 \text{ and } f_2(x) := (1-\frac{1}{x})^2.$$

Then

$$-\lim_{r\downarrow 1}\frac{1}{4\pi^2}\oint_{|\xi_1|=1}\oint_{|\xi_2|=1}\frac{1}{(\xi_1 - r\xi_2)^2} f_1\left(\frac{|1+h\xi_1|^2}{(1-\gamma_2)^2}\right) f_2\left(\frac{|1+h\xi_2|^2}{(1-\gamma_2)^2}\right) d\xi_2 d\xi_1 = \frac{J_1K_2}{h} + \frac{J_1K_2}{h(h^2-1)} + \frac{-J_1K_3(h^2+1)}{h^2} + \frac{-J_1K_3}{h^2(h^2-1)} + \frac{J_2K_2 2h}{(h^2-1)^3} + \frac{J_2K_3}{h^2} + \frac{J_2K_3(1-3h^2))}{h^2(h^2-1)^3}$$

where

$$K_2 = \frac{2h(1+h^2)}{(1-\gamma_2)^4} - \frac{2h}{(1-\gamma_2)^2}, \quad K_3 = \frac{h^2}{(1-\gamma_2)^4}$$
$$J_1 = -2(1-\gamma_2)^2 \text{ and } J_2 = (1-\gamma_2)^4.$$

Proof. Firstly we have that,

$$\begin{aligned} f_2\left(\frac{|1+h\xi|^2}{(1-\gamma_2)^2}\right) &= \left(1 - \frac{(1-\gamma_2)^2}{(1+h\xi_2)(1+h\bar{\xi_2})}\right)^2 \\ &= 1 - 2\frac{(1-\gamma_2)^2}{(1+h\xi_2)(1+h\bar{\xi_2})} + \frac{(1-\gamma_2)^4}{(1+h\xi_2)^2(1+h\bar{\xi_2})^2} \\ &= 1 + \frac{J_1}{(1+h\xi_2)(1+h\bar{\xi_2})} + \frac{J_2}{(1+h\xi_2)^2(1+h\bar{\xi_2})^2} \\ &= 1 + \frac{J_1\xi_2}{(1+h\xi_2)(\xi_2+h)} + \frac{J_2\xi_2^2}{(1+h\xi_2)^2(\xi_2+h)^2} \end{aligned}$$

Using the same constants as in Lemmas C.1.1 and C.1.2 we have the following,

$$
\begin{aligned}
&-\frac{1}{4\pi^2}\oint_{|\xi_1|=1}\oint_{|\xi_2|=1}\frac{\left(K_1+K_2\xi_1+K_2\xi_1^{-1}+K_3\xi_1^2+K_3\xi_1^{-2}\right)\left(1+\frac{J_1\xi_2}{(1+h\xi_2)(\xi_2+h)}+\frac{J_2\xi_2^2}{(1+h\xi_2)^2(\xi_2+h)^2}\right)}{(\xi_1-r\xi_2)^2}d\xi_1 d\xi_2\\
&=-\frac{1}{4\pi^2}\oint_{|\xi_1|=1}\oint_{|\xi_2|=1}\frac{\left(K_1+K_2\xi_1+K_2\xi_1^{-1}+K_3\xi_1^2+K_3\xi_1^{-2}\right)}{(\xi_1-r\xi_2)^2}d\xi_1\\
&\quad\times\left(1+\frac{J_1\xi_2}{(1+h\xi_2)(\xi_2+h)}+\frac{J_2\xi_2^2}{(1+h\xi_2)^2(\xi_2+h)^2}\right)d\xi_2\\
&=-\frac{2\pi i}{4\pi^2}\oint_{|\xi_2|=1}\left(\frac{K_2}{r^2\xi_2^2}+\frac{K_3}{r^3\xi_2^3}\right)\left(1+\frac{J_1\xi_2}{(1+h\xi_2)(\xi_2+h)}+\frac{J_2\xi_2^2}{(1+h\xi_2)^2(\xi_2+h)^2}\right)d\xi_2\\
&=-\frac{2\pi i}{4\pi^2}\oint_{|\xi_2|=1}\frac{K_2}{r^2\xi_2^2}+\frac{K_3}{r^3\xi_2^3}+\frac{J_1K_2}{r^2\xi_2(1+h\xi_2)(\xi_2+h)}+\frac{J_1K_3}{\xi_2^2(1+h\xi_2)(\xi_2+h)}+\\
&\quad\frac{J_2K_2}{r^2(1+h\xi_2)^2(\xi_2+h)^2}+\frac{J_2K_3}{r^2\xi(1+h\xi_2)^2(\xi_2+h)^2}d\xi_2\\
&=-\frac{2\pi i}{4\pi^2}\oint_{|\xi_2|=1}\frac{J_1K_2}{r^2\xi_2(1+h\xi_2)(\xi_2+h)}+\frac{J_1K_3}{\xi_2^2(1+h\xi_2)(\xi_2+h)}+\\
&\quad\frac{J_2K_2}{r^2(1+h\xi_2)^2(\xi_2+h)^2}+\frac{J_2K_3}{r^2\xi_2(1+h\xi_2)^2(\xi_2+h)^2}d\xi_2\\
&=-\frac{2\pi i}{4\pi^2}\oint_{|\xi_2|=1}((i)+(ii)+(iii)+(iv))d\xi_2
\end{aligned}
$$

These values can be calculated using the residue theorem.

Term	(i)	(ii)	(iii)	(iv)
Residue Locations	0, -h	0, -h	-h	0, -h
Orders	1,1	2, 1	2	1, 2

Then the integral is given by the following,

$$\frac{J_1K_2}{h}+\frac{J_1K_2}{h(h^2-1)}+\frac{-J_1K_3(h^2+1)}{h^2}+\frac{-J_1K_3}{h^2(h^2-1)}+ \tag{C.1.2}$$

$$\frac{J_2K_2 2h}{(h^2-1)^3}+\frac{J_2K_3}{h^2}+\frac{J_2K_3(1-3h^2)}{h^2(h^2-1)^3}. \tag{C.1.3}$$

□

C.2 Proof of Main Results

In this section, we provide proofs for the main results in the chapter.

Proof of Proposition 6.2.1. Firstly

$$R(\boldsymbol{X}_1,\boldsymbol{X}_2)=(\boldsymbol{X}_2^T\boldsymbol{X}_2)^{-1}\boldsymbol{X}_1^T\boldsymbol{X}_1=\left((\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_2^T\boldsymbol{X}_2\right)^{-1}=(R(\boldsymbol{X}_2,\boldsymbol{X}_1))^{-1},$$

which implies that

$$\lambda_j(R(\boldsymbol{X}_1, \boldsymbol{X}_2)) = \lambda_j^{-1}(R(\boldsymbol{X}_2, \boldsymbol{X}_1))$$

Then

$$\begin{aligned} R(\boldsymbol{X}_1, \boldsymbol{X}_2) &= \sum_{j=1}^{p} \left(1 - \lambda_j(R(\boldsymbol{X}_1, \boldsymbol{X}_2))\right)^2 + \left(1 - \lambda_j^{-1}(R(\boldsymbol{X}_1, \boldsymbol{X}_2))\right)^2 \\ &= \sum_{j=1}^{p} \left(1 - \lambda_j^{-1}(R(\boldsymbol{X}_2, \boldsymbol{X}_1))\right)^2 + \left(1 - \lambda_j(R(\boldsymbol{X}_2, \boldsymbol{X}_1))\right)^2. \end{aligned}$$

Now the final term is the definition of the test statistic $T(\boldsymbol{X}_2, \boldsymbol{X}_1)$. Thus

$$T(\boldsymbol{X}_1, \boldsymbol{X}_2) = \sum_{j=1}^{p} \left(1 - \lambda_j^{-1}(R(\boldsymbol{X}_2, \boldsymbol{X}_1))\right)^2 + \left(1 - \lambda_j(R(\boldsymbol{X}_2, \boldsymbol{X}_1))\right)^2 = T(\boldsymbol{X}_2, \boldsymbol{X}_1)$$

proving symmetry.

We can write $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ as $\Sigma_1^{\frac{1}{2}}\boldsymbol{Z}_1$ and $\Sigma_1^{\frac{1}{2}}\boldsymbol{Z}_2$ respectively, where $\mathbb{E}(\boldsymbol{Z}_\cdot^T\boldsymbol{Z}_\cdot) = \boldsymbol{I}_p$. Then

$$\begin{aligned} \lambda_j(R(\boldsymbol{X}_1, \boldsymbol{X}_2)) &= \lambda_j\left(\left(\Sigma_1^{\frac{1}{2}}\boldsymbol{Z}_2^T\boldsymbol{Z}_2\Sigma_1^{\frac{1}{2}}\right)^{-1}\Sigma_1^{\frac{1}{2}}\boldsymbol{Z}_1^T\boldsymbol{Z}_1\Sigma_1^{\frac{1}{2}}\right) \\ &= \lambda_j\left(\Sigma_1^{-\frac{1}{2}}\left(\boldsymbol{Z}_2^T\boldsymbol{Z}_2\right)^{-\frac{1}{2}}\Sigma_1^{-\frac{1}{2}}\Sigma_1^{\frac{1}{2}}\boldsymbol{Z}_1^T\boldsymbol{Z}_1\Sigma_1^{\frac{1}{2}}\right) = \\ &= \lambda_j\left(\Sigma_1^{-\frac{1}{2}}\left(\boldsymbol{Z}_2^T\boldsymbol{Z}_2\right)^{-\frac{1}{2}}\boldsymbol{Z}_1^T\boldsymbol{Z}_1\Sigma_1^{\frac{1}{2}}\right) = \\ = \lambda_j\left(\left(\boldsymbol{Z}_2^T\boldsymbol{Z}_2\right)^{-\frac{1}{2}}\boldsymbol{Z}_1^T\boldsymbol{Z}_1\Sigma_1^{\frac{1}{2}}\Sigma_1^{-\frac{1}{2}}\right) &= \lambda_j\left(\left(\boldsymbol{Z}_2^T\boldsymbol{Z}_2\right)^{-\frac{1}{2}}\boldsymbol{Z}_1^T\boldsymbol{Z}_1\right) = \lambda_j(R(\boldsymbol{Z}_1, \boldsymbol{Z}_2)). \end{aligned}$$

We can write $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ as $\Sigma_2^{\frac{1}{2}}\boldsymbol{Z}_3$ and $\Sigma_2^{\frac{1}{2}}\boldsymbol{Z}_4$, where $\boldsymbol{Z}_1 \overset{D}{=} \boldsymbol{Z}_3$ and $\boldsymbol{Z}_2 \overset{D}{=} \boldsymbol{Z}_4$. Then by a similar argument as before we have that

$$T(\boldsymbol{Y}_1, \boldsymbol{Y}_2) = T(\boldsymbol{Z}_3, \boldsymbol{Z}_4) \overset{D}{=} T(\boldsymbol{Z}_1, \boldsymbol{Z}_2) = T(\boldsymbol{X}_1, \boldsymbol{X}_2),$$

completing the proof □

The proof of Theorem 6.3.1 requires the application of Theorem 3.1 Zheng, 2012. For completeness, we state the this result in full below.

Theorem C.2.1. *Zheng, 2012 Let $X \in \mathbb{R}^{n_1 \times p}$ and $Y \in \mathbb{R}^{n_2 \times p}$ be random matrices satisfying Assumption 6.3.1, and $f_1, \ldots, f_s$ (s is a fixed integer) be functions analytic*

in an open region in the complex plane containing the interval $[a_\gamma, b_\gamma]$. *Then, as* $\boldsymbol{n} \to \infty$, *the random vector*

$$\left[\int f_k(x) dG_n(x)\right] 1 \leq k \leq s$$

converges weakly to a Gaussian vector $(X_{f_1}, \ldots X_{f_s})$ *with means,* μ_{f_k}, *and variances,* $\sigma^2_{f_k}$,

$$\mathbb{E}_{f_k}(\boldsymbol{\gamma}) := \lim_{r \downarrow 1} \frac{1}{4\pi i} \oint_{|z|=1} f\left(\frac{|1+h\xi|^2}{(1-\gamma_2)^2}\right) \left[\frac{1}{\xi - r^{-1}} + \frac{1}{\xi + r^{-1}} - \frac{2}{\xi + \frac{\gamma_2}{h}}\right] d\xi \tag{C.2.1}$$

$$Cov_{f_k,f_j}(\boldsymbol{\gamma}) := -\lim_{r \downarrow 1} \frac{2}{4\pi^2} \oint_{|\xi_1|=1} \oint_{|\xi_2|=1} \frac{1}{(\xi_1 - r\xi_2)^2} f\left(\frac{|1+h\xi_1|^2}{(1-\gamma_2)^2}\right) f\left(\frac{|1+h\xi_2|^2}{(1-\gamma_2)^2}\right) d\xi_2 d\xi_1. \tag{C.2.2}$$

We now use the above result, and the results in the previous section to prove the main result of the chapter.

Proof. Proof of Theorem 6.3.1 Let $t_1(x) = (1-x)^2$ and $t_2(x) = (1 - \frac{1}{x})^2$. Then by Theorem C.2.1 the vector $\boldsymbol{t}_n(x) := (\int f_1(x) dF_n(x), \int f_2(x) dF_n(x))$ converges to a Normal vector with mean and covariance given by equations (C.2.1) and (C.2.2). Now our test statistic (at a single time point) can be expressed as $\mathbf{1}^T \boldsymbol{t}_n(x)$ and thus by the continuous mapping theorem converges weakly to a Normal random variable with moments

$$\mathbb{E}_{f_1}(\gamma) + \mathbb{E}_{f_2}(\gamma) \text{ and } Cov_{f_1,f_1}(\gamma) + 2Cov_{f_1,f_2}(\gamma) + Cov_{f_2,f_2}(\gamma). \tag{C.2.3}$$

We also have the following relationship between t_1 and t_2,

$$t_1(\lambda_j(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2)) = (1 - \lambda_j(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2))^2 = (1 - \lambda_j(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1))^2 = t_2(\lambda_j(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1)).$$

By Theorem C.2.1, the limiting distributions of f_1 and f_2 depend on $\boldsymbol{\gamma}$ which implies that

$$\mathbb{E}_{t_1}(\gamma_1, \gamma_2) = \mathbb{E}_{t_2}(\gamma_2, \gamma_1) \text{ and } Cov^2_{t_1,t_1}(\gamma_1, \gamma_2) = Cov^2_{t_2,t_2}(\gamma_2, \gamma_1). \tag{C.2.4}$$

By Lemma C.1.1 we have that

$$\mathbb{E}_{t_1}(\gamma) = 2K_{3,1}\left(1 - \frac{\gamma_2^2}{h^2}\right) + \frac{2K_{2,1}\gamma_2}{h}$$

where

$$K_{2,1} = \frac{2h(1+h^2)}{(1-\gamma_2)^4} - \frac{2h}{(1-\gamma_2)^2}, \;\; K_{3,1} = \frac{h^2}{(1-\gamma_2)^4}.$$

By symmetry

$$\mathbb{E}_{t_2}(\gamma) = 2K_{3,2}\left(1 - \frac{\gamma_1^2}{h^2}\right) + \frac{2K_{2,2}\gamma_1}{h}$$

where

$$K_{2,2} = \frac{2h(1+h^2)}{(1-\gamma_1)^4} - \frac{2h}{(1-\gamma_1)^2}, \;\; K_{3,2} = \frac{h^2}{(1-\gamma_1)^4}.$$

Combining these values gives the expectation.

By Lemma C.1.2 we have that

$$Cov_{t_1,t_1}(\gamma) = 2\left(K_{2,1}^2 + 2K_{3,1}^2\right)$$

and by symmetry we have that

$$Cov_{t_2,t_2}(\gamma) = 2\left(K_{2,2}^2 + 2K_{3,2}^2\right).$$

Finally by Lemma C.1.3 we have that

$$Cov_{t_1,t_2}(\gamma) = 2\Bigg(\frac{J_1K_{2,1}}{h} + \frac{J_1K_{2,1}}{h(h^2-1)} + \frac{-J_1K_{3,1}(h^2+1)}{h^2} + \frac{-J_1K_{3,1}}{h^2(h^2-1)} + \frac{J_2K_{2,1}2h}{(h^2-1)^3} + \frac{J_2K_{3,1}}{h^2} + \frac{J_2K_{3,1}(1-3h^2)}{h^2(h^2-1)^3}\Bigg)$$

where

$$J_1 = -2(1-\gamma_2)^2 \text{ and } J_2 = (1-\gamma_2)^4.$$

Plugging these values into (C.2.3) gives the required result. □

Bibliography

Anderson, G. W., Guionnet, A., & Zeitouni, O. (2010). *An introduction to random matrices*. Cambridge University Press.

Arlot, S., Celisse, A., & Harchaoui, Z. (2019). A kernel multiple change-point algorithm via model selection. *Journal of Machine Learning Research*, *20*(162), 1–56.

Aston, J. A. D., & Kirch, C. (2012a). Detecting and estimating changes in dependent functional data. *Journal of Multivariate Analysis*, *109*, 204–220.

Aston, J. A. D., & Kirch, C. (2012b). Evaluating stationarity via change-point alternatives with applications to fMRI data. *The Annals of Applied Statistics*, *6*(4), 1906–1948.

Aue, A., Gabrys, R., Horváth, L., & Kokoszka, P. (2009). Estimation of a change-point in the mean function of functional data. *Journal of Multivariate Analysis*, *100*(10), 2254–2269.

Aue, A., Hörmann, S., Horváth, L., Reimherr, M. et al. (2009). Break detection in the covariance structure of multivariate time series models. *The Annals of Statistics*, *37*(6B), 4046–4087.

Aue, A., Rice, G., & Sönmez, O. (2018). Detecting and dating structural breaks in functional data without dimension reduction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *80*(3), 509–529.

Aue, A., Rice, G., & Sönmez, O. (2020). Structural break analysis for spectrum and trace of covariance operators. *Environmetrics*, *31*(1), e2617.

Auger, I. E., & Lawrence, C. E. (1989). Algorithms for the optimal identification of segment neighborhoods. *Bulletin of Mathematical Biology*, *51*(1), 39–54.

Avanesov, V., & Buzun, N. (2018). Change-point detection in high-dimensional covariance structure. *Electronic Journal of Statistics*, *12*(2), 3254–3294.

Bai, J. (2010). Common breaks in means and variances for panel data. *Journal of Econometrics*, *157*(1), 78–92.

Bai, Z., & Silverstein, J. W. (2004). Clt for linear spectral statistics of large-dimensional sample covariance matrices. *Annals of Probability*, 553–605.

Bardwell, L., Fearnhead, P., Eckley, I. A., Smith, S., & Spott, M. (2018). Most recent changepoint detection in panel data. *Technometrics*, 1–11.

Barnett, I., & Onnela, J.-P. (2016). Change point detection in correlation networks. *Scientific Reports*, *6*, 18893.

Berens, T., Weiß, G. N., & Wied, D. (2015). Testing for structural breaks in correlations: Does it improve value-at-risk forecasting? *Journal of Empirical Finance*, *32*, 135–152.

Berkes, I., Gabrys, R., Horváth, L., & Kokoszka, P. (2009). Detecting changes in the mean of functional observations. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *71*(5), 927–946.

Bleakley, K., & Vert, J.-P. (2011). The group fused lasso for multiple change-point detection. *arXiv preprint arXiv:1106.4199*.

Brault, V., Ouadah, S., Sansonnet, L., & Lévy-Leduc, C. (2018). Nonparametric multiple change-point estimation for analyzing large hi-c data matrices. *Journal of Multivariate Analysis*, *165*, 143–165.

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., & Vandergheynst, P. (2017). Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, *34*(4), 18–42.

Carr, J. R., Bell, H., Killick, R., & Holt, T. (2017). Exceptional retreat of novaya zemlya's marine-terminating outlet glaciers between 2000 and 2013. *The Cryosphere*, *11*(5), 2149–2174.

Celisse, A., Marot, G., Pierre-Jean, M., & Rigaill, G. (2018). New efficient algorithms for multiple change-point detection with reproducing kernels. *Computational Statistics & Data Analysis*, *128*, 200–220.

Chen, H., & Zhang, N. (2015). Graph-based change-point detection. *The Annals of Statistics*, *43*(1), 139–176.

Chen, J., & Gupta, A. (1997). Testing and locating variance changepoints with application to stock prices. *Journal of the American Statistical Association*, *92*(438), 739–747.

Chen, J., & Gupta, A. (2004). Statistical inference of covariance change points in gaussian model. *Statistics*, *38*(1), 17–28.

Chen, K., Cohen, A., & Sackrowitz, H. (2011). Consistent multiple testing for change points. *Journal of Multivariate Analysis*, *102*(10), 1339–1343.

Cho, H. (2016). Change-point detection in panel data via double cusum statistic. *Electronic Journal of Statistics*, *10*(2), 2000–2038.

Cho, H., & 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.

Chu, L., & Chen, H. (2019). Asymptotic distribution-free change-point detection for multivariate and non-euclidean data. *The Annals of Statistics*, *47*(1), 382–414.

Cribben, I., Haraldsdottir, R., Atlas, L. Y., Wager, T. D., & Lindquist, M. A. (2012). Dynamic connectivity regression: Determining state-related changes in brain connectivity. *NeuroImage*, *61*(4), 907–920.

Cribben, I., Wager, T., & Lindquist, M. (2013). Detecting functional connectivity change points for single-subject fMRI data. *Frontiers in Computational Neuroscience*, *7*, 143.

Cribben, I., & Yu, Y. (2017). Estimating whole-brain dynamics by using spectral clustering. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *66*(3), 607–627.

Csorgo, M., & Horváth, L. (1997). *Limit theorems in change-point analysis*. John Wiley & Sons Chichester.

Davis, R. A., Lee, T. C. M., & Rodriguez-Yam, G. A. (2006). Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association*, *101*(473), 223–239.

Davis, R. A., Zang, P., & Zheng, T. (2016). Sparse vector autoregressive modeling. *Journal of Computational and Graphical Statistics*, *25*(4), 1077–1096.

Dette, H., & Gösmann, J. (2018). Relevant change points in high dimensional time series. *Electronic Journal of Statistics*, *12*(2), 2578–2636.

Dette, H., & Kutta, T. (2019). Detecting structural breaks in eigensystems of functional time series. *arXiv preprint arXiv:1911.07580*.

Dette, H., Pan, G., & Yang, Q. (2018). Estimating a change point in a sequence of very high-dimensional covariance matrices. *arXiv preprint arXiv:1807.10797*.

Dubey, P., & Müller, H.-G. (2019). Fréchet change point detection. *arXiv preprint arXiv:1911.11864*.

Eichinger, B., & Kirch, C. (2018). A MOSUM procedure for the estimation of multiple random change points. *Bernoulli*, *24*(1), 526–564.

Enikeeva, F., & Harchaoui, Z. (2019). High-dimensional change-point detection under sparse alternatives. *The Annals of Statistics*, *47*(4), 2051–2079.

Fan, J., Lv, J., & Qi, L. (2011). Sparse high-dimensional models in economics. *Annual Review of Economics*, *3*(1), 291–317.

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

Finn, J. D. (1974). *A general model for multivariate analysis*. Holt, Rinehart & Winston.

Franke, J., Kirch, C., & Kamgaing, J. T. (2012). Changepoints in times series of counts. *Journal of Time Series Analysis*, *33*(5), 757–770.

Frick, K., Munk, A., & Sieling, H. (2014). Multiscale change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(3), 495–580.

Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, *42*(6), 2243–2281.

Fryzlewicz, P. (2020). Detecting possibly frequent change-points: Wild binary segmentation 2 and steepest-drop model selection. *Journal of the Korean Statistical Society*, 1–44.

Fujita, A., Sato, J. R., Garay-Malpartida, H. M., Yamaguchi, R., Miyano, S., Sogayar, M. C., & Ferreira, C. E. (2007). Modeling gene expression regulatory networks with the sparse vector autoregressive model. *BMC Systems Biology*, *1*(1), 39.

Garreau, D., Arlot, S. et al. (2018). Consistent change-point detection with kernels. *Electronic Journal of Statistics*, *12*(2), 4440–4486.

Gibberd, A. J., & Nelson, J. (2014). High dimensional changepoint detection with a dynamic graphical lasso. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2684–2688.

Gibberd, A. J., & Nelson, J. (2017). Regularized estimation of piecewise constant gaussian graphical models: The group-fused graphical lasso. *Journal of Computational and Graphical Statistics*, *26*(3), 623–634.

Grattarola, D., Zambon, D., Livi, L., & Alippi, C. (2019). Change detection in graph streams by learning graph embeddings on constant-curvature manifolds. *IEEE Transactions on Neural Networks and Learning Systems*, 1–14.

Grundy, T., Killick, R., & Mihaylov, G. (2020). High-dimensional changepoint detection via a geometrically inspired mapping. *arXiv preprint arXiv:2001.05241*.

Guha-Sapir, D., Schlüter, B., Rodriguez-Llanes, J. M., Lillywhite, L., & Hicks, M. H.-R. (2018). Patterns of civilian and child deaths due to war-related violence in syria: A comparative analysis from the violation documentation center dataset, 2011–16. *The Lancet Global Health*, *6*(1), e103–e110.

Guionnet, A., Zeitouni, O. et al. (2000). Concentration of the spectral measure for large matrices. *Electronic Communications in Probability*, *5*, 119–136.

Haynes, K., Eckley, I. A., & Fearnhead, P. (2017). Computationally efficient change-point detection for a range of penalties. *Journal of Computational and Graphical Statistics*, *26*(1), 134–143.

Hernandez-Lopez, F. J., & Rivera, M. (2014). Change detection by probabilistic segmentation from monocular view. *Machine Vision and Applications*, *25*(5), 1175–1195.

Hillel, D. (2003). *Introduction to environmental soil physics*. Elsevier.

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

Hocking, T. D., Schleiermacher, G., Janoueix-Lerosey, I., Boeva, V., Cappo, J., Delattre, O., ... Vert, J.-P. (2013). Learning smoothing models of copy number profiles using breakpoint annotations. *BMC bioinformatics*, *14*(1), 164.

Horváth, L., & Hušková, M. (2012). Change-point detection in panel data. *Journal of Time Series Analysis*, *33*(4), 631–648.

Inclan, C., & 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.

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

James, N. A., & Matteson, D. S. (2015). Ecp: An R package for nonparametric multiple change point analysis of multivariate data. *Journal of Statistical Software*, *62*(i07).

Jarušková, D. (2013). Testing for a change in covariance operator. *Journal of Statistical Planning and Inference*, *143*(9), 1500–1511.

Jirak, M. (2012). Change-point analysis in increasing dimension. *Journal of Multivariate Analysis*, *111*, 136–159.

Jirak, M. et al. (2015). Uniform change point tests in high dimension. *The Annals of Statistics*, *43*(6), 2451–2483.

Killick, R., Eckley, I. A., Ewans, K., & Jonathan, P. (2010). Detection of changes in variance of oceanographic time-series using changepoint analysis. *Ocean Engineering*, *37*(13), 1120–1126.

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

Kirch, C., Muhsal, B., & Ombao, H. (2015). Detection of changes in multivariate time series with application to EEG data. *Journal of the American Statistical Association*, *110*(511), 1197–1216.

Kovács, S., Li, H., Bühlmann, P., & Munk, A. (2020). Seeded binary segmentation: A general methodology for fast and optimal change point detection. *arXiv preprint arXiv:2002.06633*.

Kowal, D. R. (2020). *Dsp: Dynamic shrinkage processes*. R package version 0.1.0.

Kowal, D. R., Matteson, D. S., & Ruppert, D. (2019). Dynamic shrinkage processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *81*(4), 781–804.

Kwon, D., Ko, K., Vannucci, M., Reddy, A. N., & Kim, S. (2006). Wavelet methods for the detection of anomalies and their application to network traffic analysis. *Quality and Reliability Engineering International*, *22*(8), 953–969.

Lavielle, M., & Teyssiere, G. (2006). Detection of multiple change-points in multivariate time series. *Lithuanian Mathematical Journal*, *46*(3), 287–306.

Lawley, D. N. (1938). A generalization of fisher's z test. *Biometrika*, *30*(1/2), 180–187.

Li, J. (2020). Asymptotic distribution-free change-point detection based on interpoint distances for high-dimensional data. *Journal of Nonparametric Statistics*, *0*(0), 1–28.

Li, J., Xu, M., Zhong, P.-S., & Li, L. (2019). Change point detection in the mean of high-dimensional time series data under dependence. *arXiv preprint arXiv:1903.07006*.

Li, S., Xie, Y., Dai, H., & Song, L. (2015). M-statistic for kernel change-point detection. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems 28* (pp. 3366–3374). Curran Associates, Inc.

Londschien, M., Kovács, S., & Bühlmann, P. (2019). Change point detection for graphical models in presence of missing values. *arXiv preprint arXiv:1907.05409*.

Lung-Yut-Fong, A., Lévy-Leduc, C., & Cappé, O. (2015). Homogeneity and change-point detection tests for multivariate data using rank statistics. *Journal de la Société Française de Statistique*, *156*(4), 133–162.

Ma, T. F., & Yau, C. Y. (2016). A pairwise likelihood-based approach for changepoint detection in multivariate time series models. *Biometrika*, *103*(2), 409–421.

Mahmoud, M. A., Parker, P. A., Woodall, W. H., & Hawkins, D. M. (2007). A change point method for linear profile data. *Quality and Reliability Engineering International*, *23*(2), 247–268.

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

Matteson, D. S., & 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.

Nicholson, W. B., Matteson, D. S., & Bien, J. (2017). Varx-l: Structured regularization for large vector autoregressions with exogenous variables. *International Journal of Forecasting*, *33*(3), 627–651.

Olshen, A. B., Venkatraman, E. S., Lucito, R., & Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, *5*(4), 557–572.

Padilla, O. H. M., Yu, Y., & Priebe, C. E. (2019). Change point localization in dependent dynamic nonparametric random dot product graphs. *arXiv preprint arXiv:1911.07494*.

Padilla, O. H. M., Yu, Y., Wang, D., & Rinaldo, A. (2019). Optimal nonparametric multivariate change point detection and localization. *arXiv preprint arXiv:1910.13289*.

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

Pickering, B. (2016). *Changepoint detection for acoustic sensing signals* (Doctoral dissertation, Lancaster University).

Potthoff, R. F., & Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, *51*(3/4), 313–326.

Prabuchandran, K., Singh, N., Dayama, P., & Pandit, V. (2019). Change point detection for compositional multivariate data. *arXiv preprint arXiv:1901.04935*.

Ramsay, J. O. (2004). Functional data analysis. *Encyclopedia of Statistical Sciences*, *4*.

Rigaill, G. (2010). Pruned dynamic programming for optimal multiple change-point detection. *arXiv preprint arXiv:1004.0887*, *17*.

Rigaill, G. (2015). A pruned dynamic programming algorithm to recover the best segmentations with 1 to k_max change-points. *Journal de la Société Française de Statistique*, *156*(4), 180–205.

Rubin-Delanchy, P., Lawson, D. J., & Heard, N. A. (2016). Anomaly detection for cyber security applications. In *Dynamic networks and cyber-security* (pp. 137–156). World Scientific.

Safikhani, A., & Shojaie, A. (2020). Joint structural break detection and parameter estimation in high-dimensional non-stationary VAR models. *Journal of the American Statistical Association*.

Scott, A. J., & Knott, M. (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, *30*(3), 507–512.

Silverstein, J. W. (1985). The limiting eigenvalue distribution of a multivariate f matrix. *SIAM Journal on Mathematical Analysis*, *16*(3), 641–646.

Steland, A. (2020). Testing and estimating change-points in the covariance matrix of a high-dimensional time series. *Journal of Multivariate Analysis*, 104582.

Stoehr, C., Aston, J. A. D., & Kirch, C. (2020). Detecting changes in the covariance structure of functional time series with application to fmri data. *Econometrics and Statistics*.

Storchi, R., Rodgers, J., Gracey, M., Martial, F. P., Wynne, J., Ryan, S., ... Lucas, R. J. (2019). Measuring vision using innate behaviours in mice with intact and impaired retina function. *Scientific reports*, *9*(1), 1–16.

Tao, T. (2012). *Topics in random matrix theory*. American Mathematical Soc.

Tartakovsky, A., Nikiforov, I., & Basseville, M. (2014). *Sequential analysis: Hypothesis testing and changepoint detection*. Chapman and Hall/CRC.

Tickle, S., Eckley, I., Fearnhead, P., & Haynes, K. (2020). Parallelization of a common changepoint detection method. *Journal of Computational and Graphical Statistics*, *29*(1), 149–161.

Truong, C., Oudre, L., & Vayatis, N. (2019). Greedy kernel change-point detection. *IEEE Transactions on Signal Processing*, *67*(24), 6204–6214.

Venkatraman, E. S. (1993). Consistency results in multiple change-point problems.

Violations Documentation Center in Syria. (2019). About us. [Online; accessed 23-May-2019].

Wagner, A. B., Hill, E. L., Ryan, S. E., Sun, Z., Deng, G., Bhadane, S., ... Matteson, D. S. (2020). Social distancing has merely stabilized covid-19 in the us. *medRxiv*.

Wang, D., Yu, Y., & Rinaldo, A. (2017). Optimal covariance change point localization in high dimension. *arXiv preprint arXiv:1712.09912*.

Wang, D., Yu, Y., & Rinaldo, A. (2018). Optimal change point detection and localization in sparse dynamic networks. *arXiv preprint arXiv:1809.09602*.

Wang, D., Yu, Y., Rinaldo, A., & Willett, R. (2019). Localizing changes in high-dimensional vector autoregressive processes. *arXiv preprint arXiv:1909.06359*.

Wang, R., Volgushev, S., & Shao, X. (2019). Inference for change points in high dimensional data. *arXiv preprint arXiv:1905.08446*.

Wang, T., & Samworth, R. J. (2016). Inspectchangepoint: High-dimensional changepoint estimation via sparse projection. R package.

Wang, T., & 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.

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York.

Wied, D., Ziggel, D., & Berens, T. (2013). On the application of new tests for structural changes on global minimum-variance portfolios. *Statistical Papers*, *54*(4), 955–975.

Wigner, E. (1967). Random matrices in physics. *SIAM Review*, *9*(1), 1–23.

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika*, *24*(3/4), 471–494.

Yao, Y.-C. (1988). Estimating the number of change-points via schwarz' criterion. *Statistics and Probability Letters*, *6*(3), 181–189.

Yin, Y., Bai, Z., & Krishnaiah, P. (1983). Limiting behavior of the eigenvalues of a multivariate f matrix. *Journal of Multivariate Analysis*, *13*(4), 508–516.

Zhang, N., Siegmud, D. O., Ji, H., & Li, J. Z. (2010). Detecting simultaneous change-points in multiple sequences. *Biometrika*, *97*(3), 631–645.

Zhao, Z., Ma, T. F., Ng, W. L., & Yau, C. Y. (2019). A composite likelihood-based approach for change-point detection in spatio-temporal process. *arXiv preprint arXiv:1904.06340*.

Zheng, S. (2012). Central limit theorems for linear spectral statistics of large dimensional f -matrices. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, *48*(2), 444–476.