MULTIPLE CHANGE POINT ANALYSIS OF MULTIVARIATE DATA VIA ENERGY STATISTICS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Nicholas James

August 2015

MULTIPLE CHANGE POINT ANALYSIS OF MULTIVARIATE DATA VIA ENERGY STATISTICS

Nicholas James, Ph.D.

Cornell University 2015

In this dissertation we consider the offline multiple change point problem. More specifically we are interested in estimating both the number of change points, and their locations within a given multivariate time series. Many current works in this area assume that the time series observations follow a known parametric model, or that there is at most one change point. This work examines the change point problem in a more general setting, where both the observation distributions and number of change points are unknown. Our goal is to develop methods for identifying change points, while making as few unrestrictive assumptions as possible.

The following chapters are a collections of works that introduced new nonparametric change point algorithms. These new algorithms are based upon E-Statistics and have the ability to detect *any* type of distributional change. The theoretical properties of these new algorithms are studied, and conditions under which consistent estimates for the number of change point and change point locations are presented. These newly proposed algorithms are used to analyze various dataset, ranging from financial time series to emergency medical service data. Efficient implementations of these algorithms are provided by the `R` package `ecp`. A portion of this dissertation is devoted to the discussion of the implementation of these algorithms, as well as the use of the software package.

BIOGRAPHICAL SKETCH

Nicholas was born in Georgetown, Guyana on May 4, 1988. In 1995 he moved to Tallahassee, Florida along with his family. After completing high school Nicholas went on to attend the University of Florida in Gainesville, Florida. While at the University of Florida he majored in mathematics and minored in computer science. After graduating from the University of Florida with a bachelors of science he was admitted to the MS/PhD Operations Research and Information Engineering program at Cornell University.

Nicholas was advised by David S. Matteson during his time at Cornell University. His PhD dissertation is on performing change point analysis on multivariate time series, where he developed theoretical results for newly created algorithms, as well as the creation of various software packages. Nicholas joined Google Inc. after graduating with a PhD from Cornell.

ACKNOWLEDGEMENTS

First and foremost I would like to thank my advisor Professor David S. Matteson for his constant support and interest in my work. I would also like to thank the National Physical Science Consortium for their financial support throughout my PhD program. Finally, I would like to thank everyone who made my time at Cornell University an enjoyable and rewarding experience.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Change point analysis is the process of detecting distributional changes within time-ordered observations. For a given observed time series, the instance where distributional changes occur are referred to as change points. Change points have implications in both applied and theoretical statistics. For instance, it has been shown that when fitting either a location shift or linear model, ignoring the existence of change points can lead to inconsistent parameter estimates, thus potentially causing practitioners to draw incorrect conclusions. Change point analysis has also more recently become a valuable tool in bioinformatics, where variations in DNA copy number data can indicate the presence of certain kinds of cancers.

Generally speaking, change point analysis can be performed in one of four ways. Analysis can be performed in either an online or offline fashion, through the use of either parametric or nonparametric models. Parametric change point analysis assumes that observations are drawn from a class of distribution functions, who's members can be uniquely identified by parameter value. In this setting analysis focuses on detecting changes in the parameter of interest. While parametric approaches lead to many useful theoretical results about their performance, they can not always be used in real world settings, since adherence to the underlying model is not guaranteed. In such situations using a nonparametric change point algorithm would be more appropriate. Compared to their parametric counterparts these methods sacrifice some statistical power but are suitable for more areas of application.

This dissertation focuses on performing offline nonparametric change point analysis. The methods developed in this dissertation are able to detect *any* type

of distributional change. Unlike many nonparametric approaches to change point analysis, the existence of a density function for the observed distributions is not assumed. Therefore, the use of any density estimation tools become inappropriate. Instead, the change point algorithms introduced in this dissertation require the existence of certain absolute moments. This is accomplished through the use of E-Statistics [73]; a class of statistical divergence measures that are indexed by a parameter α. For $0 < \alpha < 2$, it can be shown that E-Statistics are able to detect any type of distributional change, while a choice of $\alpha = 2$ allows for detecting only change in expectation.

The following chapters in this dissertation are a select collection of works published while attending Cornell University. Chapter 2 introduces E-Statistics. Using this tool, two change point algorithms, E-Divisive and E-Agglomerative, are developed. Consistency results are shown for the E-Divisive method, while applications to real world datasets show that both perform quite well. Chapter 3 is dedicated to the `ecp` package. This is an `R` software package that allows its user to perform change point analysis through the use of E-Statistics. This chapter focuses on the methods that implement the E-Divisive and E-Agglomerative procedures. Finally, Chapter 4 introduces a new probabilistic pruning procedure, `cp3o`, which can be used to increase the speed of many change point algorithms. Combining this pruning procedure with E-Statistics provides us with the `e-cp3o` algorithm, which is also part of the `ecp` package. This chapter shows that the `e-cp3o` algorithm addresses some of the weaknesses of both the E-Agglomerative and E-Divisive algorithms.

CHAPTER 2

A NONPARAMETRIC APPROACH FOR MULTIPLE CHANGE POINT ANALYSIS OF MULTIVARIATE DATA

2.1 Introduction

Change point analysis is the process of detecting distributional changes within time-ordered observations. This arises in financial modeling [74], where correlated assets are traded and models are based on historical data represented as multivariate time series. It is applied in bioinformatics [57] to identify genes that are associated with specific cancers and other diseases. Change point analysis is also used to detect credit card fraud [9] and other anomalies [69, 1]; and for data classification in data mining [51]. Applications can also be found in signal processing, where change point analysis can be used to detect significant changes within a stream of images [46].

While change point analysis is important in a variety of fields, the methodologies that have been developed to date often assume a single or known number of change points. This assumption is often unrealistic, as seen in Section 2.5. Increasingly, applications also require detecting changes in multivariate data, for which traditional methods have limited applicability. To address these shortcomings, we propose a new methodology, based on U-statistics, that is capable of consistently estimating an unknown number of multiple change point locations. The proposed methods are broadly defined for observations from an arbitrary, but fixed dimension.

In general, change point analysis may be performed in either parametric and

nonparametric settings. Parametric analysis necessarily assumes that the underlying distributions belong to some known family, and the likelihood function plays a major role. For example, in [11] and [48] analysis is performed by maximizing a log-likelihood function, while [61] examines the ratio of log-likelihood functions to estimate change points. Additionally, [17] combine the log-likelihood, the minimum description length, and a genetic algorithm in order to identify change points. Nonparametric alternatives are applicable in a wider range of applications than are parametric ones [32]. Nonparametric approaches often rely heavily on the estimation of density functions [43], though they have also been performed using rank statistics [49]. We propose a nonparametric approach based on Euclidean distances between sample observations. It is simple to calculate and avoids the difficulties associated with multivariate density estimation.

Change point methods are often directly motivated by specific fields of study. For example, [42] discusses an approach that is rooted in information theory, and ideas from model selection are applied for determining both the number and location of change points in [79] and [82]. The proposed approach is motivated by methods from cluster analysis [73].

Change point algorithms either estimate all change points concurrently or hierarchically. Concurrent methods generally optimize a single objective function. For example, given that there are k change points, [34] estimates change point locations by maximizing a likelihood function. [48] accomplish the same task by minimizing a loss function. Sequential methods generally estimate change points one at a time [29], although some have the ability to estimate two or more at any given stage [59]. Such approaches are often characterized as bisection procedures. The proposed method utilizes a bisection approach for its computational efficiency.

We propose a new method that can detect any distributional change within an independent sequence, and which does not make any distributional assumptions beyond the existence of the αth absolute moment, for some $\alpha \in (0, 2)$. Estimation is performed in a manner that simultaneously identifies both the number and locations of change points. In Section 2.2 we describe our methodology; its properties are discussed in Section 2.3. In Sections 2.4 and 2.5 we present the results of our procedure when applied to simulated and real data, respectively. In Section 2.6 we propose an alternative algorithm and illustrate its use on a novel spatio-temporal application. Concluding remarks are in Section 2.7 and technical details are stated in the Appendix.

2.2 Methodology

To highlight the generality of the proposed method, we briefly summarize the different conditions under which analysis may be performed, in increasing complexity. Let $Z_1, Z_2, \ldots, Z_T \in \mathbb{R}^d$ be an independent sequence of time-ordered observations. Throughout this manuscript, the time between observations is assumed positive; it may be fixed or randomly distributed. The time index simply denotes the time order. In the simplest case, there is a single hypothesized change point location τ. Specifically, $Z_1, \ldots, Z_\tau \overset{iid}{\sim} F_1$ and $Z_{\tau+1}, \ldots, Z_T \overset{iid}{\sim} F_2$, in which F_1 and F_2 are unknown probability distributions. Here we test for homogeneity in distribution, $H_0 : F_1 = F_2$ verses $H_A : F_1 \neq F_2$. For univariate observations with continuous distributions the familiar Kolmogorov-Smirnov test may be applied, and in the general case the approach in [67] may be applied. If H_0 is rejected we conclude there is a change point at τ, otherwise we conclude there is no distributional change in the observations.

A slight modification of the above setting assumes instead that the change point location is unknown, but assumes that at most only one change point exists. A natural way to proceed is to choose τ as the most likely location for a change point, based on some criterion. Here, τ is chosen from some subset of $\{1, 2, \ldots, T-1\}$, then a test for homogeneity is performed. This should necessarily incorporate the fact that τ is unknown.

Now, suppose there is a known number of change points k in the series, but with unknown locations. Thus, there exist change points $0 < \tau_1 < \cdots < \tau_k < T$, that partition the sequence into $k+1$ clusters, such that observations within clusters are identically distributed, and observations between adjacent clusters are not. A naive approach for estimating the best of all $O(T^k)$ change point locations quickly becomes computationally intractable for $k \geq 3$. One remedy is to instead maximize the objective function through the use of dynamic programming as in [31], [65] and [49].

Finally, in the most general case, both the number of change points as well as their locations are unknown. Here, the naive approach to concurrent estimation becomes infeasible. As such, bisection [76, 14] and model selection procedures [48, 4] are popular under these conditions.

We now present a nonparametric technique, which we call E-Divisive, for performing multiple change point analysis of a sequence of multivariate observations. The E-Divisive method combines bisection [76] with a multivariate divergence measure from [73]. We first discuss measuring differences in multivariate distributions. We then propose a procedure for hierarchically estimating change point locations. We conclude this section by discussing the hierarchical statistical testing used to determine the number of change points.

2.2.1 Measuring Differences in Multivariate Distributions

For complex-valued functions $\phi(\cdot)$, the complex conjugate of ϕ is denoted by $\overline{\phi}$, and the absolute square $|\phi|^2$ is defined as $\phi\overline{\phi}$. The Euclidean norm of $x \in \mathbb{R}^d$ is $|x|_d$, or simply $|x|$ when there is no ambiguity. A primed variable such as X' is an independent copy of X; that is, X and X' are independent and identically distributed (iid).

For random variables $X, Y \in \mathbb{R}^d$, let ϕ_x and ϕ_y denote the characteristic functions of X and Y, respectively. A divergence measure between multivariate distributions may be defined as

$$\int_{\mathbb{R}^d} |\phi_x(t) - \phi_y(t)|^2 \, w(t) \, dt, \tag{2.1}$$

in which $w(t)$ denotes an arbitrary positive weight function, for which the above integral exists. In consideration of Lemma 7 (see Appendix), we use the following weight function

$$w(t;\alpha) = \left(\frac{2\pi^{d/2}\Gamma(1-\alpha/2)}{\alpha 2^{\alpha}\Gamma((d+\alpha)/2)} |t|^{d+\alpha} \right)^{-1}, \tag{2.2}$$

for some fixed constant $\alpha \in (0,2)$. Then, if $E|X|^{\alpha}, E|Y|^{\alpha} < \infty$, a characteristic function based divergence measure may be defined as

$$\mathcal{D}(X,Y;\alpha) = \int_{\mathbb{R}^d} |\phi_x(t) - \phi_y(t)|^2 \left(\frac{2\pi^{d/2}\Gamma(1-\alpha/2)}{\alpha 2^{\alpha}\Gamma((d+\alpha)/2)} |t|^{d+\alpha} \right)^{-1} dt. \tag{2.3}$$

Suppose $X, X' \overset{iid}{\sim} F_x$ and $Y, Y' \overset{iid}{\sim} F_y$, and that X, X', Y, and Y' are mutually independent. If $E|X|^{\alpha}, E|Y|^{\alpha} < \infty$, then we may employ an alternative divergence measure based on Euclidean distances, defined by [73] as

$$\mathcal{E}(X,Y;\alpha) = 2E|X-Y|^{\alpha} - E|X-X'|^{\alpha} - E|Y-Y'|^{\alpha}. \tag{2.4}$$

Lemma 1. *For any pair of independent random vectors $X, Y \in \mathbb{R}^d$, and for any $\alpha \in (0,2)$, if $E(|X|^\alpha + |Y|^\alpha) < \infty$, then $\mathcal{E}(X,Y;\alpha) = \mathcal{D}(X,Y;\alpha)$, $\mathcal{E}(X,Y;\alpha) \in [0,\infty)$, and $\mathcal{E}(X,Y;\alpha) = 0$ if and only if X and Y are identically distributed.*

A proof is given in the Appendix, and for a more general setting in [73].

The equivalence established in Lemma 1 motivates a remarkably simple empirical divergence measure for multivariate distributions based on U-statistics. Let $\boldsymbol{X}_n = \{X_i : i = 1,\dots,n\}$ and $\boldsymbol{Y}_m = \{Y_j : j = 1,\dots,m\}$ be independent iid samples from the distribution of $X, Y \in \mathbb{R}^d$, respectively, such that $E|X|^\alpha, E|Y|^\alpha < \infty$ for some $\alpha \in (0,2)$. Then an empirical divergence measure analogous to Equation (2.4) may be defined as

$$\widehat{\mathcal{E}}(\boldsymbol{X}_n, \boldsymbol{Y}_m; \alpha) = \frac{2}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}|X_i - Y_j|^\alpha - \binom{n}{2}^{-1}\sum_{1\le i<k\le n}|X_i - X_k|^\alpha - \binom{m}{2}^{-1}\sum_{1\le j<k\le m}|Y_j - Y_k|^\alpha. \quad (2.5)$$

This measure is based on Euclidean distances between sample elements and is $O(m^2 \vee n^2)$, whereas the sample counterpart of Equation (2.3) requires d-dimensional integration to evaluate.

Under the assumptions above, $\widehat{\mathcal{E}}(\boldsymbol{X}_n, \boldsymbol{Y}_m; \alpha) \to \mathcal{E}(X,Y;\alpha)$ almost surely as $m \wedge n \to \infty$ by the Strong Law of Large Numbers for U-statistics [35] and the continuity theorem. Additionally, under the null hypothesis of equal distributions, i.e., $\mathcal{E}(X,Y;\alpha) = 0$, we note that $\frac{mn}{m+n}\widehat{\mathcal{E}}(\boldsymbol{X}_n, \boldsymbol{Y}_m; \alpha)$ converges in distribution to a non-degenerate random variable as $m \wedge n \to \infty$. Further, under the alternative hypothesis of unequal distributions, i.e., $\mathcal{E}(X,Y;\alpha) > 0$, we note that $\frac{mn}{m+n}\widehat{\mathcal{E}}(\boldsymbol{X}_n, \boldsymbol{Y}_m; \alpha) \to \infty$ almost surely as $m \wedge n \to \infty$. These asymptotic results motivate the statistical tests described in Section 2.2.4.

2.2.2 Estimating the Location of a Change Point

Let

$$\widehat{\mathcal{Q}}(\boldsymbol{X}_n, \boldsymbol{Y}_m; \alpha) = \frac{mn}{m+n}\widehat{\mathcal{E}}(\boldsymbol{X}_n, \boldsymbol{Y}_m; \alpha) \tag{2.6}$$

denote the scaled sample measure of divergence discussed above. This statistic leads to a consistent approach for estimating change point locations. Let $Z_1, \ldots, Z_T \in \mathbb{R}^d$ be an independent sequence of observations and let $1 \leq \tau < \kappa \leq T$ be constants. Now define the following sets, $\boldsymbol{X}_\tau = \{Z_1, Z_2, \ldots, Z_\tau\}$ and $\boldsymbol{Y}_\tau(\kappa) = \{Z_{\tau+1}, Z_{\tau+2}, \ldots, Z_\kappa\}$. A change point location $\hat{\tau}$ is then estimated as

$$(\hat{\tau}, \hat{\kappa}) = \underset{(\tau,\kappa)}{\operatorname{argmax}}\ \widehat{\mathcal{Q}}(\boldsymbol{X}_\tau, \boldsymbol{Y}_\tau(\kappa); \alpha). \tag{2.7}$$

It is possible to calculate the argmax in Equation (2.7) in $O(T^2)$ by observing that $\widehat{\mathcal{Q}}(\boldsymbol{X}_\tau, \boldsymbol{Y}_\tau(\kappa); \alpha)$ can be derived directly from $\widehat{\mathcal{Q}}(\boldsymbol{X}_{\tau-1}, \boldsymbol{Y}_{\tau-1}(\kappa); \alpha)$ and the distances $\{|Z_\tau - Z_j|^\alpha : 1 \leq j < \tau\}$.

If it is known that at most one change point exists, we fix $\kappa = T$. Otherwise, the variable κ is introduced to alleviate a weakness of bisection, as mentioned in [75], in which it may be more difficult to detect certain types of distributional changes in the multiple change point setting using only bisection. For example, if we fix $\kappa = T$ and the set $\boldsymbol{Y}_\tau(T)$ contains observations across multiple change points (e.g., distinct distributions), then it is possible that the resulting mixture distribution in $\boldsymbol{Y}_\tau(T)$ is indistinguishable from the distribution of the observations in $\boldsymbol{X}_\tau$, even when τ corresponds to a valid change point. We avoid this confounding by allowing κ to vary, with minimal computational cost by storing the distances mentioned above. This modification to bisection is similar to that taken in [59].

2.2.3 Hierarchically Estimating Multiple Change Points

To estimate multiple change points we iteratively apply the above technique as follows. Suppose that $k-1$ change points have been estimated at locations $0 < \hat{\tau}_1 < \cdots < \hat{\tau}_{k-1} < T$. This partitions the observations into k clusters $\widehat{C}_1, \widehat{C}_2, \ldots, \widehat{C}_k$, such that $\widehat{C}_i = \{Z_{\hat{\tau}_{i-1}+1}, \ldots, Z_{\hat{\tau}_i}\}$, in which $\hat{\tau}_0 = 0$ and $\hat{\tau}_k = T$. Given these clusters, we then apply the procedure for finding a single change point to the observations *within* each of the k clusters. Specifically, for the ith cluster $\widehat{C}_i$ denote a proposed change point location as $\hat{\tau}(i)$ and the associated constant $\hat{\kappa}(i)$, as defined by Equation (2.7). Now, let

$$i^* = \underset{i \in \{1,\ldots,k\}}{\operatorname{argmax}} \; \hat{Q}(\boldsymbol{X}_{\hat{\tau}(i)}, \boldsymbol{Y}_{\hat{\tau}(i)}(\hat{\kappa}(i)); \alpha),$$

in which $\boldsymbol{X}_{\hat{\tau}(i)}$ and $\boldsymbol{Y}_{\hat{\tau}(i)}(\hat{\kappa}(i))$ are defined with respect to $\widehat{C}_i$, and denote a corresponding test statistic as

$$\hat{q}_k = \hat{Q}(\boldsymbol{X}_{\hat{\tau}_k}, \boldsymbol{Y}_{\hat{\tau}_k}(\hat{\kappa}_k); \alpha), \tag{2.8}$$

in which $\hat{\tau}_k = \hat{\tau}(i^*)$ denotes the kth estimated change point, located within cluster $\widehat{C}_{i^*}$, and $\hat{\kappa}_k = \hat{\kappa}(i^*)$ the corresponding constant. This iterative procedure has running time $O(kT^2)$, in which k is the unknown number of change points.

2.2.4 Hierarchical Significance Testing

The previous sections have proposed a method for estimating the locations of change points. We now propose a testing procedure to determine the statistical significance of a change point, conditional on previously estimated change points. For hierarchical estimation, this test may be used as a stopping criterion for the proposed iterative estimation procedure.

As above, suppose that $k-1$ change points have been estimated, resulting in k clusters, and that conditional on $\{\hat{\tau}_1, \ldots, \hat{\tau}_{k-1}\}$, $\hat{\tau}_k$ and $\hat{q}_k$ are the newly proposed change point location and the associated test statistic, respectively. Large values of $\hat{q}_k$ correspond to a significant change in distribution within one of the existing clusters, however, calculating a precise critical value requires knowledge of the underlying distributions, which are generally unknown. Therefore, we propose a permutation test to determine the significance of $\hat{q}_k$.

Under the null hypothesis of no additional change points, we conduct a permutation test as follows. First, the observations *within* each cluster are permuted to construct a new sequence of length T. Then, we reapply the estimation procedure as described in Sections 2.2.2 and 2.2.3 to the permuted observations. This process is repeated and after the rth permutation of the observations we record the value of the test statistic $\hat{q}_k^{(r)}$.

This permutation test will result in an exact p-value if we consider all possible permutations. This is not computationally tractable, in general; instead we obtain an approximate p-value by performing a sequence of R *random* permutations. In our implementation we fix the significance level $p_0 \in (0, 1)$ of the conditional test, as well as the the number of permutations R, and the approximate p-value is defined as $\#\{r : \hat{q}_k^{(r)} \geq \hat{q}_k\}/(R+1)$. In our analysis we fix $p_0 = 0.05$ and use $R = 499$ permutations for all of our testing. Determining a suitably large R to obtain an adequate approximation depends on the distribution of the observations, as well as the number and size of clusters. As an alternative, a sequential implementation of the random permutations may be implemented with a uniformly bounded resampling risk, see [24].

The permutation test may be performed at each stage in the iterative estima-

tion algorithm. The kth change point is deemed significant, given $\{\hat{\tau}_1, \ldots, \hat{\tau}_{k-1}\}$, if the approximate p-value is less than p_0, and the procedure then estimates an additional location. Otherwise, we are unable to reject the null hypothesis of no additional change points and the algorithm terminates. The permutation test may be performed after the E-Divisive procedure reaches a predetermined number of clusters to quickly provide initial estimates. The independent calculations of the permuted observations may be performed in parallel to easily reduce computation time.

2.3 Consistency

We now present results pertaining to the consistency of the estimated change point locations that are returned by the proposed procedure. It is assumed throughout that the dimension of the observations is arbitrary, but constant, and that the unknown number of change points is also constant. Below, we consider the case of a single change point, and demonstrate that we obtain a strongly consistent estimator in a rescaled time setting. We then do the same for the more general case of multiple change points.

2.3.1 Single Change Point

In Section 2.2.1 we have stated that in the case of a single change point, at a given location, the two-sample test is statistically consistent against all alternatives. We now show that $\hat{\tau}$ is a strongly consistent estimator for a single change point location within the setting described.

Assumption 1. *Suppose that we have a heterogeneous sequence of independent observations from two different distributions. Specifically, let* $\gamma \in (0,1)$ *denote the fraction of the observations belonging to one of the distributions, such that* $Z_1, \ldots, Z_{\lfloor \gamma T \rfloor} \sim F_x$ *and* $Z_{\lfloor \gamma T \rfloor + 1}, \ldots, Z_T \sim F_y$ *for every sample of size* T*. Let* $r = \lfloor \gamma T \rfloor$ *and* $s = T - r$*. Also, let* $\mu_X^\alpha = E|X - X'|^\alpha$, $\mu_Y^\alpha = E|Y - Y'|^\alpha$, *and* $\mu_{XY}^\alpha = E|X - Y|^\alpha$, *in which* $X, X' \overset{iid}{\sim} F_x$, $Y, Y' \overset{iid}{\sim} F_y$, *and* X, X', Y, *and* Y' *are mutually independent. Further, suppose* $E(|X|^\alpha + |Y|^\alpha) < \infty$ *for some* $\alpha \in (0,2)$*; hence,* $\mu_X^\alpha, \mu_Y^\alpha, \mu_{XY}^\alpha, \mathcal{E}(X, Y; \alpha) < \infty$*. Finally, let* $\{\delta_T\}$ *be a sequence of positive numbers such that* $\delta_T \to 0$ *and* $T\delta_T \to \infty$*, as* $T \to \infty$*.*

Lemma 2. *Suppose Assumption 1 holds, then*

$$\sup_{\gamma \in [\delta_T, 1-\delta_T]} \left| \binom{T}{2}^{-1} \sum_{i<j} |Z_i - Z_j|^\alpha \; - \left[\gamma^2 \mu_X^\alpha + (1-\gamma)^2 \mu_Y^\alpha + 2\gamma(1-\gamma)\mu_{XY}^\alpha \right] \right| \overset{a.s.}{\to} 0, \quad as \;\; T \to \infty.$$

Proof. Let $\epsilon > 0$. Define the following disjoint sets: $\Pi_1 = \{(i,j) : i < j, Z_i, Z_j \sim F_x\}$; $\Pi_2 = \{(i,j) : Z_i \sim F_x, Z_j \sim F_y\}$; and $\Pi_3 = \{(i,j) : i < j, Z_i, Z_j \sim F_y\}$. By the Strong Law of Large Numbers for U-statistics, we have that with probability 1, $\exists N_1 \in \mathbb{N}$ such that

$$\left| \binom{\#\Pi_1}{2}^{-1} \sum_{\Pi_1} |Z_i - Z_j|^\alpha - \mu_X^\alpha \right| < \epsilon$$

whenever $\#\Pi_1 > N_1$. By the same argument we can similarly define $N_2, N_3 \in \mathbb{N}$. Furthermore, $\exists N_4 \in \mathbb{N}$ such that $\frac{1}{T-1} < \epsilon/2$ for $T > N_4$. Let $N = N_1 \vee N_2 \vee N_3 \vee N_4$, such that for any $T\delta_T > N$, and every $\gamma \in [\delta_T, 1 - \delta_T]$, we have $\#\Pi_1 = \lfloor \gamma T \rfloor > N_1$, $\#\Pi_2 = \lfloor \gamma T \rfloor (T - \lfloor \gamma T \rfloor) > N_2$, $\#\Pi_3 = (T - \lfloor \gamma T \rfloor) > N_3$, and the quantities $|\frac{r}{T} - \gamma|$, $|\frac{r-1}{T-1} - \gamma|$, $|\frac{s}{T} - (1-\gamma)|$, $|\frac{s-1}{T-1} - (1-\gamma)|$ are each less than ϵ.

Now, considering the nature of the summands, $\frac{2}{T(T-1)} \sum_{\Pi_1} |Z_i - Z_j|^\alpha$ may be rewritten as

$$\binom{r}{2}^{-1} \left(\frac{r}{T} \right) \left(\frac{r-1}{T-1} \right) \sum_{\Pi_1} |Z_i - Z_j|^\alpha.$$

For $T > N$, we have

$$P\left(\left|\binom{r}{2}^{-1}\left(\frac{r}{T}\right)\left(\frac{r-1}{T-1}\right)\sum_{\Pi_1}|Z_i - Z_j|^\alpha - \gamma^2\mu_X^\alpha\right| < \epsilon^3 + \epsilon^2(2 + 3\mu_X^\alpha) + \epsilon\right) = 1.$$

The last inequality is obtained from noting that $\left|\frac{r}{T} - \gamma\right|\left|\frac{r-1}{T-1} - \gamma\right| < \epsilon^2$ implies $\left|\left(\frac{r}{T}\right)\left(\frac{r-1}{T-1}\right) - \gamma^2\right| < \epsilon^2 + 2\gamma\epsilon$. Therefore, $\left|\left(\frac{r}{T}\right)\left(\frac{r-1}{T-1}\right) - \gamma^2\right|\left|\binom{r}{2}^{-1}\sum_{\Pi_1}|Z_i - Z_j|^\alpha - \mu_X^\alpha\right| < \epsilon^3 + 2\gamma\epsilon^2$; rearranging terms, and using the previous inequality yields

$$\left|\binom{r}{2}^{-1}\left(\frac{r}{T}\right)\left(\frac{r-1}{T-1}\right)\sum_{\Pi_1}|Z_i - Z_j|^\alpha - \gamma^2\mu_X^\alpha\right| < \epsilon^3 + (2\gamma + (1+2\gamma)\mu_X^\alpha)\epsilon + \gamma^2\epsilon < \epsilon^3 + \epsilon^2(2+3\mu_X^\alpha) + \epsilon.$$

By applying the same approach, we have similar expressions for both $\frac{2}{T(T-1)}\sum_{\Pi_2}|Z_i - Z_j|^\alpha$ and $\frac{2}{T(T-1)}\sum_{\Pi_3}|Z_i - Z_j|^\alpha$. Finally, applying the triangle inequality establishes the claim, since ϵ is arbitrary. □

In order to establish the uniform convergence above, it is assumed that γ is bounded away from 0 and 1, such that $r \wedge s \to \infty$ as $T \to \infty$. In application, we impose a minimum size for each cluster when estimating the location of a change point. This minimum cluster size should be specified *a priori*; in our examples we primarily use 30 as the minimum size, but larger sizes may be needed when $\mathcal{E}(X, Y; \alpha)$ is relatively small.

Theorem 3. *Suppose Assumption 1 holds. Let $\hat{\tau}_T$ denote the estimated change point location for a sample of size T, as defined in Equation (2.7), here with $\kappa = T$; i.e., using an unmodified bisection approach. Then for T large enough $\gamma \in [\delta_T, 1 - \delta_T]$, and furthermore, for all $\epsilon > 0$*

$$P\left(\lim_{T\to\infty}\left|\gamma - \frac{\hat{\tau}_T}{T}\right| < \epsilon\right) = 1.$$

Proof. Let T be such that $\gamma \in [\delta_T, 1 - \delta_T]$, then for any $\tilde{\gamma} \in [\delta_T, 1 - \delta_T]$, let $\boldsymbol{X}_T(\tilde{\gamma}) = \{Z_1, \ldots, Z_{\lfloor\tilde{\gamma}T\rfloor}\}$ and $\boldsymbol{Y}_T(\tilde{\gamma}) = \{Z_{\lfloor\tilde{\gamma}T\rfloor+1}, \ldots, Z_T\}$ for all T. Then

$$\widehat{\mathcal{E}}(\boldsymbol{X}_T(\tilde{\gamma}), \boldsymbol{Y}_T(\tilde{\gamma}); \alpha) \xrightarrow{a.s.} \left(\frac{\gamma}{\tilde{\gamma}}\mathbb{1}_{\tilde{\gamma}\geq\gamma} + \frac{1-\gamma}{1-\tilde{\gamma}}\mathbb{1}_{\tilde{\gamma}<\gamma}\right)^2 \mathcal{E}(X, Y; \alpha) = h(\tilde{\gamma}; \gamma)\mathcal{E}(X, Y; \alpha) \qquad (2.9)$$

as $T \to \infty$, uniformly in $\tilde{\gamma}$. The maximum of $h(\tilde{\gamma};\gamma)$ is attained when $\tilde{\gamma} = \gamma$. Now, note that $\frac{1}{T}\widehat{Q}(\boldsymbol{X}_T(\tilde{\gamma}), \boldsymbol{Y}_T(\tilde{\gamma}); \alpha) \xrightarrow{a.s.} \tilde{\gamma}(1-\tilde{\gamma})h(\tilde{\gamma};\gamma)\mathcal{E}(X,Y;\alpha)$ as $T \to \infty$, uniformly in $\tilde{\gamma}$. Additionally, the maximum value of $\tilde{\gamma}(1-\tilde{\gamma})h(\tilde{\gamma};\gamma)$ is also attained when $\tilde{\gamma} = \gamma$. Define

$$\hat{\tau}_T = \underset{\tau \in \{\lceil T\delta_T \rceil, \lceil T\delta_T \rceil + 1, \ldots, \lfloor T(1-\delta_T) \rfloor\}}{\operatorname{argmax}} \widehat{Q}(\boldsymbol{X}_\tau, \boldsymbol{Y}_\tau(T); \alpha),$$

and the interval $\hat{\Gamma}_T = \underset{\tilde{\gamma} \in [\delta_T, 1-\delta_T]}{\operatorname{argmax}} \widehat{Q}(\boldsymbol{X}_T(\tilde{\gamma}), \boldsymbol{Y}_T(\tilde{\gamma}); \alpha)$, then $\frac{\hat{\tau}_T}{T} \in \hat{\Gamma}_T$. Since

$$\frac{1}{T}\widehat{Q}\left(\boldsymbol{X}_T\left(\hat{\tau}_T/T\right), \boldsymbol{Y}_T\left(\hat{\tau}_T/T\right); \alpha\right) > \frac{1}{T}\widehat{Q}(\boldsymbol{X}_T(\gamma), \boldsymbol{Y}_T(\gamma); \alpha) - o(1),$$

we have

$$\frac{1}{T}\widehat{Q}(\boldsymbol{X}_T(\hat{\tau}_T/T), \boldsymbol{Y}_T(\hat{\tau}_T/T); \alpha) \geq \gamma(1-\gamma)h(\gamma;\gamma)\mathcal{E}(X,Y;\alpha) - o(1),$$

by the almost sure uniform convergence. Letting $\hat{\gamma} = \hat{\tau}_T/T$, it follows that

$$\begin{aligned} 0 &\leq \gamma(1-\gamma)h(\gamma;\gamma)\mathcal{E}(X,Y;\alpha) - \hat{\gamma}(1-\hat{\gamma})h(\hat{\gamma};\gamma)\mathcal{E}(X,Y;\alpha) \\ &\leq \frac{1}{T}\widehat{Q}(\boldsymbol{X}_T(\hat{\gamma}), \boldsymbol{Y}(\hat{\gamma}); \alpha) - \hat{\gamma}(1-\hat{\gamma})h(\hat{\gamma};\gamma)\mathcal{E}(X,Y;\alpha) + o(1) \\ &\to 0, \end{aligned}$$

as $T \to \infty$. For every $\epsilon > 0$, there exists η such that

$$\tilde{\gamma}(1-\tilde{\gamma})h(\tilde{\gamma};\gamma)\mathcal{E}(X,Y;\alpha) < \gamma(1-\gamma)h(\gamma;\gamma)\mathcal{E}(X,Y;\alpha) - \eta$$

for all $\tilde{\gamma}$ with $|\tilde{\gamma} - \gamma| \geq \epsilon$. Therefore,

$$\begin{aligned} P\left(\lim_{T\to\infty} |\hat{\gamma}_T - \gamma| \geq \epsilon\right) &\leq P\Bigg(\lim_{T\to\infty} \hat{\gamma}_T(1-\hat{\gamma}_T)h(\hat{\gamma}_T;\gamma)\mathcal{E}(X,Y;\alpha) < \\ &\qquad \gamma(1-\gamma)h(\gamma;\gamma)\mathcal{E}(X,Y;\alpha) - \eta\Bigg) \\ &= 0. \end{aligned}$$

□

Consistency only requires that each cluster's size increase, but not necessarily at the same rate. To consider rates of convergence, additional information about the distribution of the estimators, which depends on the unknown distributions of the data, is also necessary.

2.3.2 Multiple Change Points

The consistency result presented in [76] cannot be applied in this general situation because it assumes that the expectation of the observed sequence consists of a piecewise linear function, making it only suitable for estimating change points resulting from breaks in expectation.

Assumption 2. *Suppose that we have a heterogeneous sequence of independent observations from $k+1$ distributions, denoted $\{F_i\}_{i=0}^{k}$. Specifically, let $0 = \gamma^{(0)} < \gamma^{(1)} < \cdots < \gamma^{(k)} < \gamma^{(k+1)} = 1$. Then, for $i = 0, 1, \ldots, k$ we have $Z_{\lfloor T\gamma^{(i)} \rfloor + 1}, \ldots, Z_{\lfloor T\gamma^{(i+1)} \rfloor} \overset{iid}{\sim} F_i$, such that $F_i \neq F_{i+1}$. Let $\mu_{ii}^{\alpha} = E|X_i - X_i'|^{\alpha}$ and $\mu_{ij}^{\alpha} = E|X_i - X_j|^{\alpha}$, in which $X_i, X_i' \overset{iid}{\sim} F_i$, independent of $X_j \sim F_j$. Furthermore, suppose that $\sum_{i=0}^{k} E|X_i|^{\alpha} < \infty$ for some $\alpha \in (0, 2)$; hence μ_{ii}^{α}, $\mu_{ij}^{\alpha}, \mathcal{E}(X_i, X_j; \alpha) < \infty$, for all i and j. Let $\{\delta_T\}$ be a sequence of positive numbers such that $\delta_T \to 0$ and $T\delta_T \to \infty$, as $T \to \infty$.*

Under Assumption 2, analysis of multiple change points can be reduced to the analysis of only two change points. For any $i \in \{1, \ldots, k-1\}$, consider $\gamma^{(i)}$ and $\gamma^{(i+1)}$. The observations $\{Z_j : j \leq \lfloor T\gamma^{(i)} \rfloor\}$ can be seen as a random sample from a mixture of distributions $\{F_j : j \leq i\}$, denoted here as F. Similarly, observations $\{Z_j : j \geq \lfloor T\gamma^{(i+1)} \rfloor + 1\}$ are a sample from a mixture of distributions $\{F_j : j > i+1\}$, denoted here as H. The remaining observations are distributed according to some distribution G. Furthermore, $F \neq G$ and $G \neq H$, if not, we refer to the single

change point setting. For notation, we simply consider $\gamma^{(1)}$ and $\gamma^{(2)}$.

Let X, Y, U be random variables such that $X \sim F$, $Y \sim H$, and $U \sim G$. Consider any $\tilde{\gamma}$ such that, $\gamma^{(1)} \leq \tilde{\gamma} \leq \gamma^{(2)}$, then this choice of $\tilde{\gamma}$ will create two mixture distributions. One with component distributions F and G, and the other with component distributions H and G. Then the divergence measure in Equation (2.3) between these two mixture distributions is equal to

$$\int_{\mathbb{R}^d} \left| \frac{\gamma^{(1)}}{\tilde{\gamma}} \phi_x(t) + \left(\frac{\tilde{\gamma} - \gamma^{(1)}}{\tilde{\gamma}} \right) \phi_u(t) - \left(\frac{1 - \gamma^{(2)}}{1 - \tilde{\gamma}} \right) \phi_y(t) - \left(\frac{\gamma^{(2)} - \tilde{\gamma}}{1 - \tilde{\gamma}} \right) \phi_u(t) \right|^2 w(t; \alpha)\, dt \quad (2.10)$$

Lemma 4. *Suppose that Assumption 2 holds for some $\alpha \in (0, 2)$, then the divergence measure in Equation (2.10) is maximized when either $\tilde{\gamma} = \gamma^{(1)}$ or $\tilde{\gamma} = \gamma^{(2)}$.*

Proof. Equation (2.10) can be rewritten as

$$f(\tilde{\gamma}) = \int_{\mathbb{R}^d} \left| \frac{\gamma^{(1)}}{\tilde{\gamma}} [\phi_x(t) - \phi_u(t)] + \frac{1 - \gamma^{(2)}}{1 - \tilde{\gamma}} [\phi_u(t) - \phi_y(t)] \right|^2 w(t; \alpha)\, dt. \quad (2.11)$$

We then express the above integral as the sum of the following three integrals:

$$\left(\frac{\gamma^{(1)}}{\tilde{\gamma}} \right)^2 \int_{\mathbb{R}^d} |\phi_x(t) - \phi_u(t)|^2 w(t; \alpha)\, dt;$$

$$\frac{2\gamma^{(1)}(1 - \gamma^{(2)})}{\gamma(1 - \tilde{\gamma})} \int_{\mathbb{R}^d} |\phi_x(t) - \phi_u(t)||\phi_u(t) - \phi_y(t)|\, w(t; \alpha)\, dt; \quad \text{and}$$

$$\left(\frac{1 - \gamma^{(2)}}{1 - \tilde{\gamma}} \right)^2 \int_{\mathbb{R}^d} |\phi_u(t) - \phi_y(t)|^2 w(t; \alpha)\, dt.$$

Each of these is a strictly convex positive function of $\tilde{\gamma}$, and therefore so is their sum. Since $\gamma^{(1)} \leq \tilde{\gamma} \leq \gamma^{(2)}$, the maximum value is attained when either $\tilde{\gamma} = \gamma^{(1)}$ or $\tilde{\gamma} = \gamma^{(2)}$. □

Lemma 5. *Suppose that Assumption 2 holds for some $\alpha \in (0, 2)$, then*

$$\sup_{\tilde{\gamma} \in [\gamma^{(1)}, \gamma^{(2)}]} \left| \widehat{\mathcal{E}}(\boldsymbol{X}_T(\tilde{\gamma}), \boldsymbol{Y}_T(\tilde{\gamma}); \alpha) - f(\tilde{\gamma}) \right| \xrightarrow{a.s.} 0, \quad as \quad T \to \infty.$$

Proof. Let $p(\tilde{\gamma};\gamma) = \frac{\gamma^{(1)}}{\tilde{\gamma}}$ and $q(\tilde{\gamma};\gamma) = \frac{1-\gamma^{(2)}}{1-\tilde{\gamma}}$. Using methods from the proof of Lemma 1, Equation (2.11) is equal to

$$\begin{aligned} p(\tilde{\gamma};\gamma)^2\mathcal{E}(X,U;\alpha) &+ q(\tilde{\gamma};\gamma)^2\mathcal{E}(Z,U;\alpha) \\ &+ 2pq(\tilde{\gamma};\gamma)\left(E|X-U|^\alpha + E|Y-U|^\alpha - E|X-Y|^\alpha - E|U-U'|^\alpha\right). \end{aligned}$$

Since $\min\left(\frac{\gamma^{(1)}}{\gamma^{(2)}}, \frac{1-\gamma^{(2)}}{1-\gamma^{(1)}}\right) > 0$, by Lemma 2 the within distances for $X_T(\tilde{\gamma})$ and $Y_T(\tilde{\gamma})$ converge uniformly to

$$p(\tilde{\gamma};\gamma)^2E|X-X'|^\alpha + (1-p(\tilde{\gamma};\gamma))^2E|U-U'|^\alpha + 2p(\tilde{\gamma};\gamma)(1-p(\tilde{\gamma};\gamma))E|X-U|^\alpha \quad \text{and}$$

$$q(\tilde{\gamma};\gamma)^2E|Y-Y'|^\alpha + (1-q(\tilde{\gamma};\gamma))^2E|U-U'|^\alpha + 2q(\tilde{\gamma};\gamma)(1-q(\tilde{\gamma};\gamma))E|Y-U|^\alpha,$$

respectively. Similarly, it can be shown that the between distance converges uniformly to

$$\begin{aligned} &pq(\tilde{\gamma};\gamma)E|X-Y|^\alpha + p(\tilde{\gamma};\gamma)(1-q(\tilde{\gamma};\gamma))E|X-U|^\alpha + \\ &(1-p(\tilde{\gamma};\gamma))(1-q(\tilde{\gamma};\gamma))E|U-U'|^\alpha + (1-p(\tilde{\gamma};\gamma))q(\tilde{\gamma};\gamma)E|Y-U|^\alpha. \end{aligned}$$

Combining twice the between less the within distances provides the desired quantity. □

Under Assumption 2, for each $i = 0, 1, \ldots, k$, there exist distributions F_i, G_i, and H_i such that for $\gamma^{(i)} \le \tilde{\gamma} \le \gamma^{(i+1)}$, Equation (2.11) holds; otherwise $f_i(\tilde{\gamma}) = 0$. By Lemmas 4 and 5, $f_i(\tilde{\gamma})$ is maximized when $\tilde{\gamma} = \gamma^{(i)}$ or $\tilde{\gamma} = \gamma^{(i+1)}$ for $i = 1, 2, \ldots, k-1$. By Theorem 3, $f_0(\tilde{\gamma})$ and $f_k(\tilde{\gamma})$ are maximized at $\gamma^{(1)}$ and $\gamma^{(k)}$, respectively.

Theorem 6. *Suppose that Assumption 2 holds for some $\alpha \in (0,2)$. For $\mathscr{A}_T \subset (\delta_T, 1-\delta_T)$ and $x \in \mathbb{R}$, define $d(x,\mathscr{A}_T) = \inf\{|x-y| : y \in \mathscr{A}_T\}$. Additionally, define $f(\gamma) = \gamma(1-\gamma)\sum_{i=0}^{k} f_i(\gamma)$. Let $\hat{\tau}_T$ be the estimated change point as defined by Equation (2.7), and $\mathscr{A}_T = \{y \in [\delta_T, 1-\delta_T] : f(y) \ge f(\gamma),\ \forall\gamma\}$. Then $d(\hat{\tau}_T/T, \mathscr{A}_T) \xrightarrow{a.s.} 0$ as $T \to \infty$.*

Proof. First we observe that $\frac{1}{T}\widehat{\mathcal{Q}}(\boldsymbol{X}_T(\tilde{\gamma}), \boldsymbol{Y}_T(\tilde{\gamma}); \alpha) \xrightarrow{a.s.} f(\tilde{\gamma})$ as $T \to \infty$, uniformly in $\tilde{\gamma}$ by Lemma 5. Also, for each i, $\tilde{\gamma}(1-\tilde{\gamma})f_i(\tilde{\gamma})$ is a strictly convex function. Therefore, for T large enough, $\delta_T < \gamma^{(1)}$ and $\gamma^{(k)} < 1 - \delta_T$, so that $\mathscr{A}_T \neq \emptyset$. Since $\tilde{\gamma}(1-\tilde{\gamma})f_i(\tilde{\gamma})$ is continuously differentiable and strictly convex, there exists a $c_i > 0$, such that for any $\tilde{\gamma}_1, \tilde{\gamma}_2 \in [\gamma^{(i)}, \gamma^{(i+1)}]$,

$$|\tilde{\gamma}_1(1-\tilde{\gamma}_1)f_i(\tilde{\gamma}_1) - \tilde{\gamma}_2(1-\tilde{\gamma}_2)f_i(\tilde{\gamma}_2)| > c_i|\tilde{\gamma}_1 - \tilde{\gamma}_2| + o(|\tilde{\gamma}_1 - \tilde{\gamma}_2|). \tag{2.12}$$

Let $\epsilon > 0$. By Equation (2.12), there exists $\eta(\epsilon) > 0$ such that if $d(\tilde{\gamma}, \mathscr{A}_T) > \eta(\epsilon)$, then $|f(\tilde{\gamma}) - f(x)| > \epsilon$, for all $x \in \mathscr{A}_T$. Now, let $\hat{\gamma}_T = \hat{\tau}_T/T$ and $\gamma^* = \operatorname{argmin}_{x \in \mathscr{A}_T} |\hat{\gamma}_T - x|$, then

$$f(\hat{\gamma}_T) + \frac{\epsilon}{2} > \frac{1}{T}\widehat{\mathcal{Q}}(\boldsymbol{X}_T(\hat{\gamma}_T), \boldsymbol{Y}_T(\hat{\gamma}_T); \alpha) \geq \frac{1}{T}\widehat{\mathcal{Q}}(\boldsymbol{X}_T(\gamma^*), \boldsymbol{Y}_T(\gamma^*); \alpha) > f(\gamma^*) - \frac{\epsilon}{2},$$

with probability 1. Combining the first and last terms in the above expression provides us with $f(\gamma^*) - f(\hat{\gamma}_T) < \epsilon$. Therefore, $P\left(\lim_{T\to\infty} d(\hat{\tau}_T/T, \mathscr{A}_T) \leq \eta(\epsilon)\right) = 1$, and since ϵ was arbitrary, we have established the claim. □

2.4 Simulation Study

In this section we present simulation results from the E-Divisive procedure using various univariate and multivariate distributions. We compare performance with the MultiRank procedure [49], which is based on a generalization of a Wilcoxon/Mann-Whitney (marginal) rank based approach, the parametric Pruned Exact Linear Time (PELT) procedure [44], and the nonparametric Kernel Change Point (KCP) procedure [4]. Each simulation applies these methods to a set of 1,000 independent sequences with two change points, and computes the average Rand index [21, 39], defined below, and approximate standard errors. All computation was completed using the statistical software `R` [63], using the `ecp` package [41].

Throughout this section the E-Divisive procedure was implemented with $\alpha = 1$; results for $\alpha = 0.5, 1.5$ were similar, and within the margin of error. We used $R = 499$ iterations when performing the permutation test, which was conducted at the marginal $p_0 = 0.05$ significance level. Furthermore, we set the minimum cluster size for the E-Divisive procedure to 30. The MultiRank and KCP procedure require upper limits on the number of change points, these were set to $\frac{T}{30} - 1$, in which T is the length of the sequence.

2.4.1 Comparing Sets of Change Point Estimates

To measure the performance of a particular method we calculate the Rand index [64] as well as Morey and Agresti's Adjusted Rand index [55]. These indices represent a measure of similarity between two different partitions of the same observations. The first is most suitable for comparing an estimated set of change points to a baseline or known set of locations, while the second is tailored to compare two sets of estimated change points. In both cases, the number of change points in each set need not be equal.

Suppose that the two clusterings of T observations are given by $U = \{U_1, \ldots, U_a\}$ and $V = \{V_1, \ldots, V_b\}$, with a and b clusters, respectively. For these two clusterings, the Rand index is calculated by noting the relative cluster membership for all *pairs* of observations. Consider the pairs of observation that fall into one of the following two sets: $\{A\}$ pairs of observation in same cluster under U and in same cluster under V; $\{B\}$ pairs of observation in different cluster under U and in different cluster under V. Let $\#A$ and $\#B$ denote the number of pairs of observation in each of these two

sets, respectively. The Rand index is then defined as

$$\text{Rand} = \frac{\#A + \#B}{\binom{T}{2}}.$$

One shortcoming of the Rand index is that it is difficult to compare two different estimated sets of clusterings, since it does not measure the departure from a given baseline model. As mentioned in [39], the Rand index, as well as other similarity indices, are not adjusted for chance (e.g., the index does not take on a constant value when comparing two random clusterings) for a given model of randomness. A common model of randomness, used in [39] and [21], is the hypergeometric model, which conditions on both the number of clusters and their sizes. Under this model, the adjustment for chance requires the expected index value and its maximum value. An Adjusted Rand index is then defined as

$$\text{Adjusted Rand} = \frac{\text{Rand} - \text{Expected Rand}}{1 - \text{Expected Rand}},$$

in which 1 corresponds to the maximum Rand index value.

2.4.2 Univariate Analysis

In this section we compare the simulation performance of the E-Divisive, Multi-Rank, and the PELT algorithms on various univariate sequences. Within these simulations, we attempt to identify change points that resulted because of a distributional change in mean, variance, or tail shape. The magnitude of these respective changes was also varied, as shown in Table 2.1.

For detecting changes in mean and variance, the E-Divisive procedure compares favorably with the parametric PELT procedure. Since the PELT procedure is

specifically designed to only identify changes in mean or variance, we compare the E-Divisive and MultiRank procedures when considering changes in tail shape. The sample size was also varied $T = 150, 300, 600$, while the three clusters maintained equal sizes of $T/3$, with distributions $N(0,1), G, N(0,1)$, respectively. We note that the Rand index values for the E-Divisive procedure tend towards 1 as the sample size increases. This follows from the consistency established in Theorem 6.

	Change in Mean			**Change in Variance**			**Change in Tail**		
T	μ	E-Divisive	PELT	σ^2	E-Divisive	PELT	ν	E-Divisive	MultiRank
150	1	$0.950_{0.001}$	$0.945_{0.002}$	2	$0.907_{0.003}$	$0.935_{0.002}$	16	$0.835_{0.017}$	$0.631_{0.005}$
	2	$0.992_{4.6\times10^{-4}}$	$0.990_{4.1\times10^{-4}}$	5	$0.973_{0.001}$	$0.987_{4.7\times10^{-4}}$	8	$0.836_{0.020}$	$0.648_{0.005}$
	4	$1.000_{3.7\times10^{-5}}$	$0.999_{9.3\times10^{-5}}$	10	$0.987_{7.1\times10^{-4}}$	$0.994_{2.7\times10^{-4}}$	2	$0.841_{0.011}$	$0.674_{0.004}$
300	1	$0.972_{9.1\times10^{-4}}$	$0.973_{8.9\times10^{-4}}$	2	$0.929_{0.003}$	$0.968_{0.001}$	16	$0.791_{0.015}$	$0.624_{0.007}$
	2	$0.996_{2.2\times10^{-4}}$	$0.994_{2.3\times10^{-4}}$	5	$0.990_{5.1\times10^{-4}}$	$0.994_{2.1\times10^{-4}}$	8	$0.729_{0.018}$	$0.639_{0.006}$
	4	$1.000_{1.0\times10^{-5}}$	$1.000_{4.5\times10^{-5}}$	10	$0.994_{3.2\times10^{-4}}$	$0.998_{1.2\times10^{-4}}$	2	$0.815_{0.006}$	$0.682_{0.006}$
600	1	$0.987_{1.5\times10^{-5}}$	$0.987_{4.1\times10^{-4}}$	2	$0.968_{0.001}$	$0.984_{5.1\times10^{-4}}$	16	$0.735_{0.019}$	$0.647_{0.016}$
	2	$0.998_{3.9\times10^{-6}}$	$0.997_{1.1\times10^{-4}}$	5	$0.995_{2.2\times10^{-4}}$	$0.997_{1.1\times10^{-4}}$	8	$0.743_{0.025}$	$0.632_{0.016}$
	4	$1.000_{3.1\times10^{-7}}$	$1.000_{2.3\times10^{-5}}$	10	$0.998_{1.5\times10^{-4}}$	$0.999_{6.4\times10^{-5}}$	2	$0.817_{0.006}$	$0.708_{0.010}$

Table 2.1: Results for E-Divisive univariate simulations

Average Rand index and approximate standard errors from 1,000 simulations for the E-Divisive, PELT and MultiRank methods. Each sample has $T = 150, 300$ or 600 observations, consisting of three equally sized clusters, with distributions $N(0,1), G, N(0,1)$, respectively. For changes in mean $G = N(\mu, 1)$, with $\mu = 1, 2$, and 4; for changes in variance $G = N(0, \sigma^2)$, with $\sigma^2 = 2, 5$, and 10; and for changes in tail shape $G = t_\nu(0,1)$, with $\nu = 16, 8$, and 2.

2.4.3 Multivariate Analysis

We next compare the results of running the E-Divisive, KCP and MultiRank methods on bivariate observations. In these simulations the distributional differences

are either a change in mean or correlation. The results of these simulations can be found in Table 2.2. Let $N_2(\boldsymbol{\mu}, \Sigma_\rho)$ denote the bivariate normal distribution with mean vector $\boldsymbol{\mu} = (\mu, \mu)'$ and covariance matrix $\Sigma_\rho = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ for $\rho \in (-1, 1)$, or simply the identity I for $\rho = 0$. We use the same setup as in the previous section, with observations from $N_2(\mathbf{0}, I), G, N_2(\mathbf{0}, I)$ distributions, respectively.

For a simultaneous change in mean, with $G = N_2(\boldsymbol{\mu}, I)$, all methods performed similarly. When detecting changes in correlation, with $G = N_2(\mathbf{0}, \Sigma_\rho)$, the KCP approach performed best when the sample size was sufficiently large for it to detect any changes. However, its computational time was about three times longer than E-Divisive, for these simulations. The MultiRank method was not reliable for detecting changes in correlation.

The final multivariate simulation examines the performance of the E-Divisive method as the dimension of the data increases. In this simulation we consider two scenarios. *With noise*: in which added components are independent, and do not have a change point. *No noise*: in which the added dimensions are correlated, and all marginal and joint distributions have common change point locations. The setting is similar to above; each sample of $T = 300, 600,$ or 900 observations consist of three equally sized clusters, with distributions $N_d(\mathbf{0}, I), G, N_d(\mathbf{0}, I)$, respectively, in which d denotes the dimension, for which we consider $d = 2, 5$ or 9.

For the no noise case, we consider $G = N_d(\mathbf{0}, \Sigma_{0.9})$, in which the diagonal elements of $\Sigma_{0.9}$ are 1 and the off-diagonal elements are 0.9. For the with noise case, we consider $G = N_d(\mathbf{0}, \Sigma_{0.9}^{noise})$, in which the diagonal elements of $\Sigma_{0.9}^{noise}$ are 1 and *only* the $(1, 2)$ and $(2, 1)$ elements are 0.9, the others are zero, such that a change in distribution occurs in the correlation of only the first two components. The results

	Change in Mean				Change in Correlation			
T	μ	E-Divisive	KCP	MultiRank	ρ	E-Divisive	KCP	MultiRank
300	1	$0.987_{4.7\times10^{-4}}$	$0.985_{6.6\times10^{-4}}$	$0.983_{4.8\times10^{-4}}$	0.5	$0.712_{0.018}$	$0.331_{N/A}$	$0.670_{0.006}$
	2	$0.992_{8.9\times10^{-5}}$	$0.998_{1.1\times10^{-4}}$	$0.991_{1.1\times10^{-4}}$	0.7	$0.758_{0.021}$	$0.331_{N/A}$	$0.723_{0.004}$
	3	$1.000_{1.3\times10^{-5}}$	$1.000_{3.9\times10^{-5}}$	$0.991_{5.1\times10^{-5}}$	0.9	$0.769_{0.017}$	$0.331_{N/A}$	$0.748_{0.002}$
600	1	$0.994_{2.2\times10^{-4}}$	$0.993_{2.3\times10^{-4}}$	$0.992_{2.1\times10^{-4}}$	0.5	$0.652_{0.022}$	$0.331_{N/A}$	$0.712_{0.011}$
	2	$1.000_{4.3\times10^{-5}}$	$0.999_{5.2\times10^{-5}}$	$0.995_{5.3\times10^{-5}}$	0.7	$0.650_{0.017}$	$0.848_{0.073}$	$0.741_{0.006}$
	3	$1.000_{3.3\times10^{-6}}$	$1.000_{2.2\times10^{-5}}$	$0.996_{2.7\times10^{-5}}$	0.9	$0.806_{0.019}$	$0.987_{0.001}$	$0.748_{0.002}$
900	1	$0.996_{1.6\times10^{-4}}$	$0.995_{1.6\times10^{-4}}$	$0.995_{1.3\times10^{-4}}$	0.5	$0.658_{0.024}$	$0.778_{0.048}$	$0.666_{0.044}$
	2	$1.000_{3.0\times10^{-5}}$	$0.999_{4.0\times10^{-5}}$	$0.997_{3.5\times10^{-5}}$	0.7	$0.633_{0.022}$	$0.974_{0.002}$	$0.764_{0.021}$
	3	$1.000_{5.2\times10^{-6}}$	$1.000_{1.4\times10^{-5}}$	$0.997_{1.8\times10^{-5}}$	0.9	$0.958_{0.004}$	$0.992_{0.004}$	$0.741_{0.006}$

Table 2.2: Results for E-Divisive multivariate simulations

Average Rand index and approximate standard errors from 1,000 simulations for the E-Divisive, MCP and MultiRank methods. Each sample has $T = 300, 600$ or 900 observations, consisting of three equally sized clusters, with distributions $N_2(\mathbf{0}, I), G, N_2(\mathbf{0}, I)$, respectively. For changes in mean $G = N_2(\boldsymbol{\mu}, I)$, with $\boldsymbol{\mu} = (1,1)', (2,2)'$, and $(3,3)'$; for changes in correlation $G = N(\mathbf{0}, \Sigma_\rho)$, in which the diagonal elements of Σ_ρ are 1 and the off-diagonal are ρ, with $\rho = 0.5, 0.7$, and 0.9.

are shown in Table 2.3. The performance of the E-Divisive method improves with increasing dimension when all components of the observed vectors are related, i.e., no noise, even when the number of observations T is fixed. However, the opposite is true when the additional components are independent with no change points. We conjecture that our method performs better when there are simultaneous changes within the components, and in the presence of noise, dimension reduction may be necessary to obtain comparable performance.

T	d	**No Noise**	**With Noise**
300	2	$0.723_{0.019}$	$0.751_{0.018}$
	5	$0.909_{0.010}$	$0.706_{0.019}$
	9	$0.967_{0.003}$	$0.710_{0.026}$
600	2	$0.930_{0.018}$	$0.822_{0.019}$
	5	$0.994_{5.4\times10^{-4}}$	$0.653_{0.023}$
	9	$0.997_{3.3\times10^{-4}}$	$0.616_{0.021}$
900	2	$0.967_{0.003}$	$0.966_{0.003}$
	5	$0.998_{1.8\times10^{-4}}$	$0.642_{0.018}$
	9	$0.999_{1.0\times10^{-4}}$	$0.645_{0.021}$

Table 2.3: Results for E-Divisive multivariate simulations with growing dimension

Average Rand index and approximate standard errors from 1,000 simulations for the E-Divisive method. Each sample has $T = 300, 600$ or 900 observations, consisting of three equally sized clusters, with distributions $N_d(\mathbf{0}, I), G, N_d(\mathbf{0}, I)$, respectively, in which $d = 2, 5$ or 9 denotes the dimension. For the no noise case, $G = N_d(\mathbf{0}, \Sigma_{0.9})$, in which the diagonal elements of $\Sigma_{0.9}$ are 1 and the off-diagonal are 0.9. For the with noise case, $G = N_d(\mathbf{0}, \Sigma_{0.9}^{noise})$, in which the diagonal elements of $\Sigma_{0.9}^{noise}$ are 1 and *only* the $(1, 2)$ and $(2, 1)$ elements are 0.9, the others are zero.

2.5 Applications

We now present results from applying the proposed E-Divisive procedure, and others, to genetics and financial datasets.

2.5.1 Genetics Data

We first consider the genome data from [8]. Genome samples for 57 individuals with a bladder tumor are scanned for variations in DNA copy number using array comparative genomic hybridization (aCGH). The relative hybridization intensity

with respect to a normal genome reference signal is recorded. These observations were normalized so that the modal ratio is zero on a logarithmic scale.

The approach in [8] assumes that each sequence is constant between change points, with additive noise. Thus, this approach is primarily concerned with finding a distributional change in the mean. In order to directly apply the procedures we first account for missing values in the data; for simplicity, we imputed the missing values as the average of their neighboring values. We removed all series that had more than 7% of values missing; leaving genome samples of 43 individuals for analysis.

When applied to the 43-dimension joint series of individuals, the MultiRank algorithm found 43 change points, while the E-Divisive algorithm found 97 change points, using $\alpha = 1$, a minimum cluster size of 10 observations, $R = 499$ permutations and $p_0 = 0.05$ in our significance testing. Estimated change point locations, for individual 10, under four methods are shown in Figure 2.1. MultiRank estimated 17 change points, with adjusted Rand values of 0.572 (Kernel CP), 0.631 (PELT), 0.677 (E-Divisive), respectively. KCPA estimated 41 change points, with adjusted Rand values of 0.678 (PELT), 0.658 (E-Divisive), respectively. PELT estimated 47 change points, with adjusted Rand value of 0.853 (E-Divisive), and E-Divisive estimated 35 change points.

2.5.2 Financial Data

Here we apply the E-Divisive algorithm to the 262 monthly log returns for Cisco Systems Inc. stock, an industry leader in the design and manufacturing of networks, from April 1990 through January 2012. In our analysis we specified $\alpha = 1$,

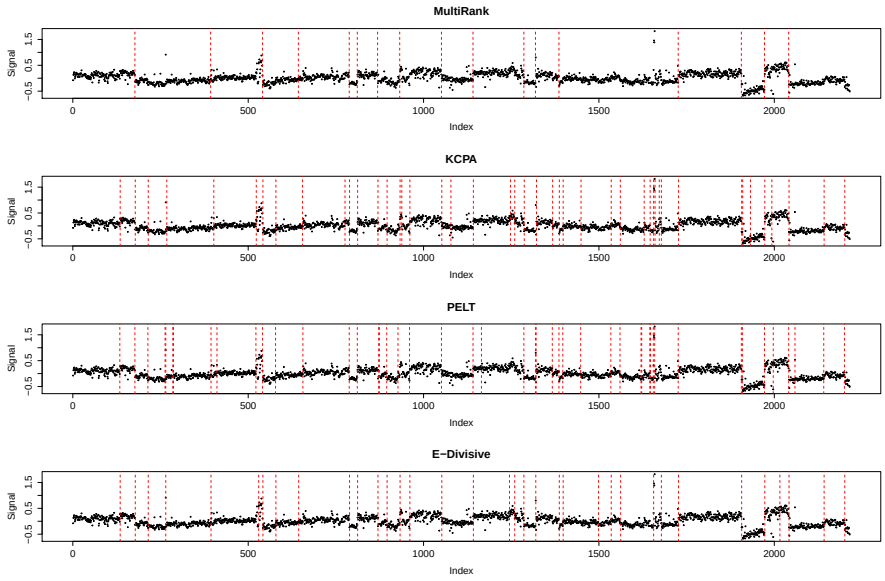

Figure 2.1: Comparison of change point procedures on aCGH data

The normalized relative aCGH signal for the tenth individual with a bladder tumor; the estimated change point locations for the MultiRank, KCPA, PELT and E-Divisive methods are indicated by the dashed vertical lines.

a minimum cluster size of 30 observations, and used $R = 499$ permutations with a level of $p_0 = 0.05$ in our significance testing. We estimated two significant change points, both with approximate p-values below 0.03. The series is shown in Figure 2.2 with vertical lines to denote the estimated change point locations at April 2000 and October 2002.

The change point in April of 2000 corresponds to the company's acquisition of Pirelli Optical Systems to counter rising competitors Nortel and Lucent. The acquisition allowed Cisco to provide its customers with lower network costs and a more complete network infrastructure. The October 2002 change point represents the end of a period of highly aggressive ventures in emerging markets, during which Cisco was chosen to develop a multi-billion dollar network for Shanghai,

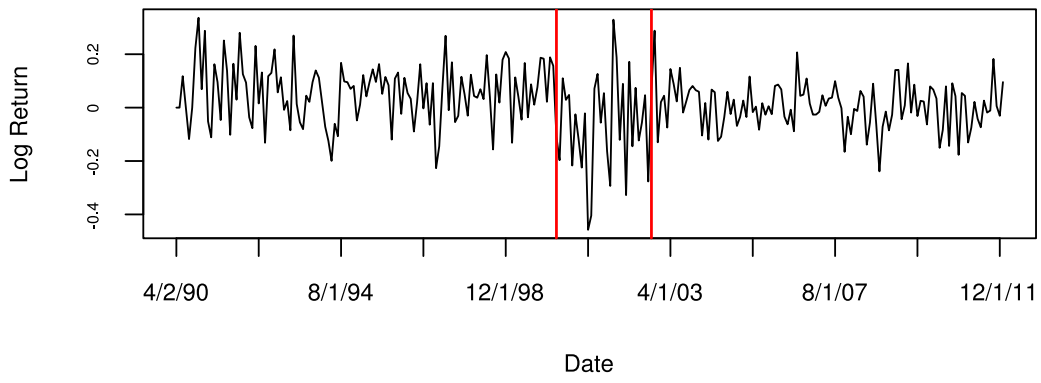

Figure 2.2: Monthly log returns for Cisco

Monthly log returns for Cisco Systems Inc. stock, from April 1990 through January 2012; the E-Divisive procedure estimates significant changes in distribution at the vertical lines April 2000 and October 2002.

which became China's largest urban communications network.

Figure 2.3 shows distributional comparisons between the three time periods. Quantile-quantile plots between adjacent time periods are shown in the first two plots and kernel density estimates for each of the three periods are shown in the third plot. Included with the kernel density estimates are 95% point-wise confidence bands, which were created by applying a bootstrap procedure to each of the three time periods. The second time period is relatively more volatile and skewed than either of its neighboring time periods.

To graphically support the assumption of independent observations within clusters, Figure 2.4 shows several lags of the sample auto-correlation function (ACF) for the returns (top row) and the squared returns (bottom row), for the entire

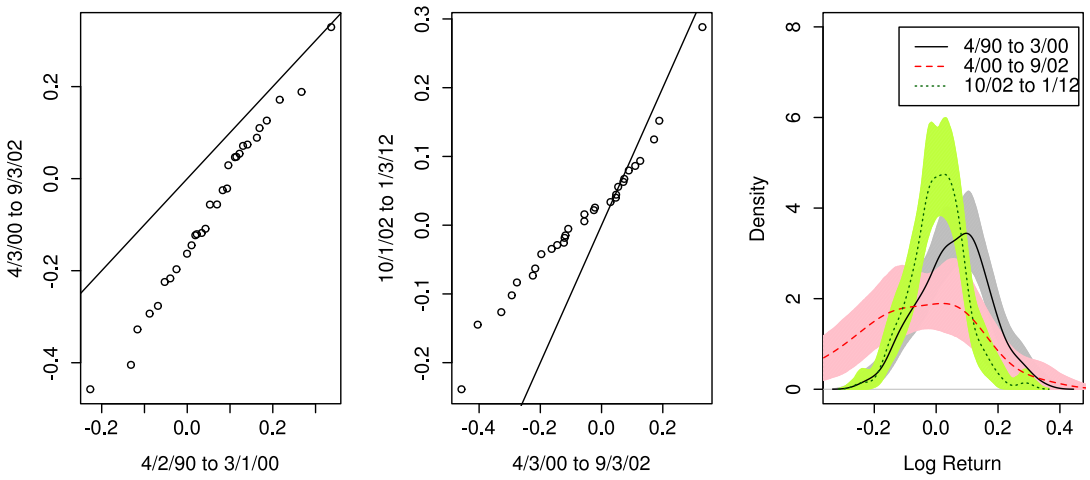

Figure 2.3: Kernel density and QQ plots for Cisco time series

Distributional comparisons between the estimated change points from the E-Divisive procedure: (a,b) quantile-quantile plots between adjacent time periods; and (c) kernel density estimates for each period with 95% confidence bands.

period (first column) and each sub-period (later columns). The dashed horizontal lines represent approximate 95% confidence intervals about zero, suggesting that the lagged correlation statistics are not significant. Within sub-periods there is no significant serial correlation or conditional heteroskedasticity. Although there appears to be minor serial dependence when studying the entire series, this is an artifact of the distributional changes over time.

2.6 An Agglomerative Algorithm

Our hierarchical approach up to this point has only considered the use of a divisive algorithm. However, we may also consider an agglomerative approach.

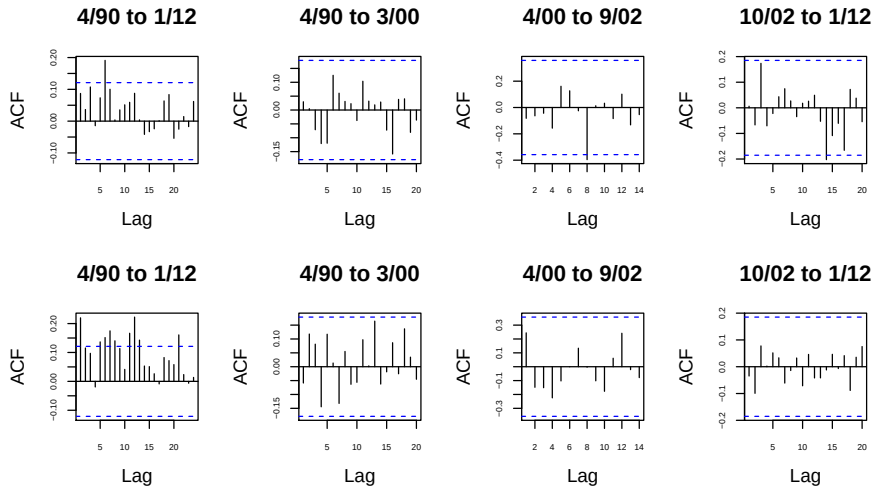

Figure 2.4: Sample auto-correlation plots for Cisco

Sample auto-correlation function for the returns (top row) and the squared returns (bottom row), for the entire period (first column) and each estimated sub-period (later columns). The dashed horizontal lines represent approximate 95% confidence intervals about zero.

2.6.1 Overview

Suppose the sequence of observations $Z_1, Z_2, \ldots, Z_T$ are independent, each with finite αth absolute moment, for some $\alpha \in (0, 2)$. Unlike most general purpose agglomerative clustering algorithms, the proposed procedure will preserve the time ordering of the observations. The number of change points will be estimated by the maximization of a goodness-of-fit statistic.

Suppose that we are initially provided a clustering $\mathcal{C} = \{C_1, C_2, \ldots, C_n\}$ of n clusters. These clusters need not consist of a single observation. We then impose the following restriction on which clusters are allowed to be merged. Suppose that $C_i = \{Z_k, Z_{k+1}, \ldots, Z_{k+t}\}$ and $C_j = \{Z_\ell, Z_{\ell+1}, \ldots, Z_{\ell+s}\}$. To preserve the time ordering, we allow C_i and C_j to merge if either $k + t + 1 = \ell$ or $\ell + s + 1 = k$, that is, if C_i and

C_j are adjacent.

To identify which adjacent pair of clusters to merge we use a goodness-of-fit statistic, defined below. We greedily optimize this statistic by merging the pair of adjacent clusters that results in either the largest increase or smallest decrease of the statistic's value. This process is repeated, recording the goodness-of-fit statistic at each step, until all observations belong to a single cluster. Finally, the estimated number of change points is estimated by the clustering that maximizes the goodness-of-fit statistic over the entire merging sequence.

2.6.2 Goodness-of-Fit

The goodness-of-fit statistic we employ is the between-within distance among adjacent clusters. Suppose that $\mathcal{C} = \{C_1, C_2, \ldots, C_n\}$, then

$$\widehat{\mathcal{S}}_n(\mathcal{C}; \alpha) = \sum_{i=1}^{n-1} \widehat{\mathcal{Q}}(C_i, C_{i+1}; \alpha), \tag{2.13}$$

in which C_i and C_{i+1} are adjacent, arranged by relabeling the clusters as necessary, and $\widehat{\mathcal{Q}}$ is defined analogous to Equation (2.6).

Initialization of the merging sequence $\{\widehat{\mathcal{S}}_k : k = n, \ldots, 2\}$ is performed by calculating $\widehat{\mathcal{Q}}$ for *all* pairs of clusters, similar to any agglomerative algorithm. We additionally note that once a pair of clusters has been merged, the statistic $\widehat{\mathcal{S}}_k$ can be updated to $\widehat{\mathcal{S}}_{k-1}$ in $\mathcal{O}(1)$; hence, the overall complexity of this approach is $\mathcal{O}(T^2)$.

2.6.3 Toronto EMS Data

In this section we apply the agglomerative algorithm to a spatio-temporal point process dataset. Data was collected during 2007 in the city of Toronto for all high priority emergency medical services (EMS) that required at least one ambulance. For each of these events a time rounded to the nearest second and a spatial location latitude and longitude were recorded. The hourly city-wide emergency event arrival rate was modeled in [52]; exploratory analysis immediately reveals that the spatial distribution also changes with time. This is largely driven by the relative changes in population density as individuals move throughout the city.

After removing data from holidays and special events, we found significant distributional changes across the course of a week, but little variation from week to week. Here we investigate the intra-week changes by pooling all of the approximately 200,000 events from 2007 into a single weekly period, in which time indicates seconds since midnight Saturday. Because of the large number of observations, we initialize the agglomerative algorithm by first partitioning the week into 672 equally spaced 15 minute periods.

The results from running the algorithm with $\alpha = 1$ are shown in the top of Figure 2.5. The goodness-of-fit measure in Equation (2.13) was maximized at 31 change points. The estimated change point locations occur everyday, primarily in the evening. Several changes occur after little duration, indicating times when the spatial distribution is quickly changing. Density estimates from observation in three adjacent cluster periods are shown, on the square-root scale, in the bottom of Figure 2.5. We note a persistently large density in the downtown region and various shape changes in the outlying regions.

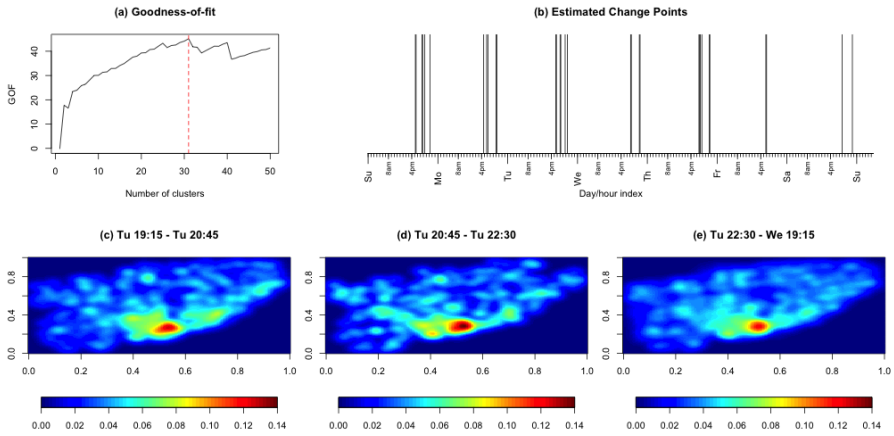

Figure 2.5: Toronto EMS data representation

Results from application of the proposed agglomerative algorithm on the Toronto EMS ambulance data: (a) the goodness-of-fit measure of Equation (2.13); (b) the 31 estimated change point locations; and spatial density estimates, on the square-root scale, from observation in three adjacent cluster periods (c) Tuesday 19:15 - 20:45, (d) Tuesday 20:45 - 22:30, and (e) Tuesday 22:30 - Wednesday 19:15.

2.7 Conclusion

We have presented a method to perform multiple change point analysis of an independent sequence of multivariate observations. We are able to consistently detect *any* type of distributional change, and do not make any assumptions beyond the existence of the αth absolute moment, for some $\alpha \in (0, 2)$. The proposed methods are able to estimate both the number of change points and their locations, thus eliminating the need for prior knowledge or supplementary analysis, unlike the methods presented in [34], [48], or [49]. Furthermore, this advantage does not come at the expense of additional computational complexity; similar to the previously mentioned methods, the proposed approach is $O(kT^2)$.

Both divisive and agglomerative versions of this method have been presented. The divisive version hierarchically tests the statistical significance of each hierarchically estimated change point, while the agglomerative version proceeds by optimizing a goodness-of-fit statistic. Because we have established consistency for the divisive procedure we prefer it in practice, even though its computation is dependent on the number of change points that are estimated.

2.8 Appendix

Let $\langle t, x \rangle$ denote the scalar product of vectors $t, x \in \mathbb{R}^d$. The following lemma is crucial to establishing a link between characteristic functions and Euclidean distances.

Lemma 7. *If* $\alpha \in (0, 2)$, *then* $\forall x \in \mathbb{R}^d$

$$\int_{\mathbb{R}^d} \frac{1 - \cos\langle t, x \rangle}{|t|^{d+\alpha}} \, dt \quad = \quad \frac{2\pi^{d/2}\Gamma(1 - \alpha/2)}{\alpha 2^\alpha \Gamma((d + \alpha)/2)} |x|^\alpha,$$

in which $\Gamma(\cdot)$ *is the complete gamma function.*

Proof. See page 177 in [73]. □

Proof of Lemma 1.

Lemma 1. *For any pair of independent random vectors* $X, Y \in \mathbb{R}^d$, *and for any* $\alpha \in (0, 2)$, *if* $E(|X|^\alpha + |Y|^\alpha) < \infty$, *then* $\mathcal{E}(X, Y; \alpha) = \mathcal{D}(X, Y; \alpha)$, $\mathcal{E}(X, Y; \alpha) \in [0, \infty)$, *and* $\mathcal{E}(X, Y; \alpha) = 0$ *if and only if* X *and* Y *are identically distributed.*

Proof. Let $w(t)$ denote any arbitrary positive weight function and note that X and Y are identically distributed if and only if Equation (2.1) is equal to zero. Take $w(t)$ equal to $w(t;\alpha)$, as defined in Equation (2.2). By definition

$$\begin{aligned}
|\phi_x(t) - \phi_y(t)|^2 &= [\phi_x(t) - \phi_y(t)]\overline{[\phi_x(t) - \phi_y(t)]} \\
&= [\phi_x(t) - \phi_y(t)][\overline{\phi_x(t)} - \overline{\phi_y(t)}] \\
&= \phi_x(t)\overline{\phi_x(t)} + \phi_y(t)\overline{\phi_y(t)} - \phi_x(t)\overline{\phi_y(t)} - \phi_y(t)\overline{\phi_x(t)}.
\end{aligned}$$

By the boundedness property of characteristic functions, Fubini's theorem implies the following equalities

$$\begin{aligned}
\phi_x(t)\overline{\phi_x(t)} &= E\left(e^{i\langle t,X\rangle}\right)E\left(e^{-i\langle t,X\rangle}\right) = E\left(e^{i\langle t,X-X'\rangle}\right) = E(\cos\langle t, X - X'\rangle), \\
\phi_y(t)\overline{\phi_y(t)} &= E\left(e^{i\langle t,Y\rangle}\right)E\left(e^{-i\langle t,Y\rangle}\right) = E\left(e^{i\langle t,Y-Y'\rangle}\right) = E(\cos\langle t, Y - Y'\rangle), \\
\phi_x(t)\overline{\phi_y(t)} &= E\left(e^{i\langle t,X\rangle}\right)E\left(e^{-i\langle t,Y\rangle}\right) = E\left(e^{i\langle t,X-Y'\rangle}\right) = E(\cos\langle t, X - Y'\rangle) + E(i\sin\langle t, X - Y'\rangle), \\
\phi_y(t)\overline{\phi_x(t)} &= E\left(e^{i\langle t,Y\rangle}\right)E\left(e^{-i\langle t,X\rangle}\right) = E\left(e^{i\langle t,Y-X'\rangle}\right) = E(\cos\langle t, Y - X'\rangle) + E(i\sin\langle t, Y - X'\rangle).
\end{aligned}$$

Note that $E(i\sin\langle t, X - Y'\rangle) + E(i\sin\langle t, Y - X'\rangle) = 0, \forall t$. Then, applying the algebraic identity

$$a + b - c - d = (1 - c) + (1 - d) - (1 - a) - (1 - b)$$

we have

$$\begin{aligned}
|\phi_x(t) - \phi_y(t)|^2 &= [1 - E(\cos\langle t, X - Y'\rangle)] + [1 - E(\cos\langle t, Y - X'\rangle)] \\
&\quad - [1 - E(\cos\langle t, X - X'\rangle)] - [1 - E(\cos\langle t, Y - Y'\rangle)],
\end{aligned}$$

hence

$$\begin{aligned}
\int |\phi_x(t) - \phi_y(t)|^2 \, w(t;\alpha)\, dt &= \int E(1 - \cos\langle t, X - Y'\rangle)\, w(t;\alpha)\, dt \\
&\quad + \int E(1 - \cos\langle t, Y - X'\rangle)\, w(t;\alpha)\, dt \\
&\quad - \int E(1 - \cos\langle t, X - X'\rangle)\, w(t;\alpha)\, dt \\
&\quad - \int E(1 - \cos\langle t, Y - Y'\rangle)\, w(t;\alpha)\, dt.
\end{aligned}$$

For any $\alpha \in (0,2)$, if $E(|X|^\alpha + |Y|^\alpha) < \infty$, then the triangle inequality implies $E|X - X'|^\alpha, E|Y - Y'|^\alpha, E|X - Y'|^\alpha, E|Y - X'|^\alpha < \infty$. Therefore, by Fubini's theorem and Lemma 7 it follows that

$$\begin{aligned}
\mathcal{D}(X,Y;\alpha) &= \int |\phi_x(t) - \phi_y(t)|^2 \, w(t;\alpha)\, dt \\
&= E\left[\int (1 - \cos\langle t, X - Y'\rangle)\left(\frac{2\pi^{d/2}\Gamma(1-\alpha/2)}{\alpha 2^\alpha \Gamma(d/2+\alpha/2)}|t|^{d+\alpha}\right)^{-1} dt\right] \\
&\quad + E\left[\int (1 - \cos\langle t, Y - X'\rangle)\left(\frac{2\pi^{d/2}\Gamma(1-\alpha/2)}{\alpha 2^\alpha \Gamma(d/2+\alpha/2)}|t|^{d+\alpha}\right)^{-1} dt\right] \\
&\quad - E\left[\int (1 - \cos\langle t, X - X'\rangle)\left(\frac{2\pi^{d/2}\Gamma(1-\alpha/2)}{\alpha 2^\alpha \Gamma(d/2+\alpha/2)}|t|^{d+\alpha}\right)^{-1} dt\right] \\
&\quad - E\left[\int (1 - \cos\langle t, Y - Y'\rangle)\left(\frac{2\pi^{d/2}\Gamma(1-\alpha/2)}{\alpha 2^\alpha \Gamma(d/2+\alpha/2)}|t|^{d+\alpha}\right)^{-1} dt\right] \\
&= E|X - Y'|^\alpha + E|Y - X'|^\alpha - E|X - X'|^\alpha - E|Y - Y'|^\alpha \\
&= \mathcal{E}(X,Y;\alpha).
\end{aligned}$$

Finally, $\mathcal{E}(X,Y;\alpha) \geq 0$ since the integrand in Equation (2.3) is non-negative. □

CHAPTER 3

ECP: AN R PACKAGE FOR NONPARAMETRIC MULTIPLE CHANGE POINT ANALYSIS OF MULTIVARIATE DATA

3.1 Introduction

Change point analysis is the process of detecting distributional changes within time-ordered observations. This arises in financial modeling [74], where correlated assets are traded and models are based on historical data. It is applied in bioinformatics [57] to identify genes that are associated with specific cancers and other diseases. Change point analysis is also used to detect credit card fraud [9] and other anomalies [1, 69]; and for data classification in data mining [51].

We introduce the `ecp` **R** package for multiple change point analysis of multivariate time series [53]. The `ecp` package provides methods for change point analysis that are able to detect *any* type of distributional change within a time series. Determination of the number of change points is also addressed by these methods as they estimate both the number and locations of change points simultaneously. The only assumptions placed on distributions are that the absolute αth moment exists, for some $\alpha \in (0, 2]$, and that observations are independent over time. Distributional changes are identified by making use of the energy statistic of [73, 67].

There are a number of freely available **R** packages that can be used to perform change point analysis, each making its own assumptions about the observed time series. For instance, the `changepoint` package [45] provides many methods for performing change point analysis of univariate time series. Although the package only

considers the case of independent observations, the theory behind the implemented methods allows for certain types of serial dependence [44]. For specific methods, the expected computational cost can be shown to be linear with respect to the length of the time series. Currently, the `changepoint` package is only suitable for finding changes in mean or variance. This package also estimates multiple change points through the use of penalization. The drawback to this approach is that it requires a user specified penalty term.

The `cpm` package [68] similarly provides a variety of methods for performing change point analysis of univariate time series. These methods range from those to detect changes in independent Gaussian data to fully nonparametric methods that can detect general distributional changes. Although this package provides methods to perform analysis of univariate time series with arbitrary distributions, these methods cannot be easily extended to detect changes in the full joint distribution of multivariate data.

Unlike the `changepoint` and `cpm` packages, the `bcp` package [19] is designed to perform Bayesian single change point analysis of univariate time series. It returns the posterior probability of a change point occurring at each time index in the series. Recent versions of this package have reduced the computational cost from quadratic to linear with respect to the length of the series. However, all versions of this package are only designed to detect changes in the mean of independent Gaussian observations.

The `strucchange` package [81] provides a suite of tools for detecting changes within linear regression models. Many of these tools however, focus on detecting at most one change within the regression model. This package also contains methods that perform online change detection, thus allowing it to be used in settings where

there are multiple changes. Additionally, if the number of changes is known *a priori* then the `breakpoints` method [80] can be used to perform retrospective analysis. For a given number of changes, this method returns the change point estimates which minimize the residual sum of squares.

In Section 3.2 we introduce the energy statistic of [73, 67], which is the fundamental divergence measure applied for change point analysis. Sections 3.3 and 3.4 provide examples of the package's methods applied to simulated data and real datasets. In the Appendix we include an outline of the algorithms used by this package's methods. Finally, the `ecp` package can be obtained at `http://cran.r-project.org/web/packages/ecp/`.

3.2 The ecp package

The `ecp` package is designed to address many of the limitations of the currently available change point packages. It is able to perform multiple change point analysis for both univariate and multivariate time series. The methods are able to estimate multiple change point locations, without *a priori* knowledge of the number of change points. The procedures assume that observations are independent with finite αth absolute moments, for some $\alpha \in (0, 2]$.

3.2.1 Measuring differences in multivariate distributions

[73, 67] introduce a divergence measure that can determine whether two independent random vectors are identically distributed. Suppose that $X, Y \in \mathbb{R}^d$ are such that, $X \sim F$ and $Y \sim G$, with characteristic functions $\phi_x(t)$ and $\phi_y(t)$, respectively.

A divergence measure between the two distributions may be defined as

$$\int_{\mathbb{R}^d} |\phi_x(t) - \phi_y(t)|^2 \, w(t) \, dt,$$

in which $w(t)$ is any positive weight function, for which the above integral is defined. Following [53] we employ the following weight function,

$$w(t;\alpha) = \left(\frac{2\pi^{d/2}\Gamma(1-\alpha/2)}{\alpha 2^\alpha \Gamma[(d+\alpha)/2]} |t|^{d+\alpha} \right)^{-1},$$

for some fixed constant $\alpha \in (0,2)$. Thus our divergence measure is

$$\mathcal{D}(X,Y;\alpha) = \int_{\mathbb{R}^d} |\phi_x(t) - \phi_y(t)|^2 \left(\frac{2\pi^{d/2}\Gamma(1-\alpha/2)}{\alpha 2^\alpha \Gamma[(d+\alpha)/2]} |t|^{d+\alpha} \right)^{-1} dt.$$

An alternative divergence measure based on Euclidean distances may be defined as follows

$$\mathcal{E}(X,Y;\alpha) = 2E|X-Y|^\alpha - E|X-X'|^\alpha - E|Y-Y'|^\alpha.$$

In the above equation, X' and Y' are independent copies of X and Y, respectively. Then given our choice of weight function, we have the following result.

Lemma 8. *For any pair of independent random variables $X, Y \in \mathbb{R}^d$ and for any $\alpha \in (0,2)$, if $E(|X|^\alpha + |Y|^\alpha) < \infty$, then $\mathcal{D}(X,Y;\alpha) = \mathcal{E}(X,Y;\alpha)$, $\mathcal{E}(X,Y;\alpha) \in [0,\infty)$, and $\mathcal{E}(X,Y;\alpha) = 0$ if and only if X and Y are identically distributed.*

Proof. A proof is given in the appendices of [73] and [53]. □

Thus far we have always assumed that $\alpha \in (0,2)$, because in this setting $\mathcal{E}(X,Y;\alpha) = 0$ if and only if X and Y are identically distributed. However, if we allow for $\alpha = 2$ a weaker result of equality in mean is obtained.

Lemma 9. *For any pair of independent random variables $X, Y \in \mathbb{R}^d$, if $E(|X|^2 + |Y|^2) < \infty$, then $\mathcal{D}(X,Y;2) = \mathcal{E}(X,Y;2)$, $\mathcal{E}(X,Y;2) \in [0,\infty)$, and $\mathcal{E}(X,Y;2) = 0$ if and only if $EX = EY$.*

Proof. See [73]. □

3.2.2 A sample divergence for multivariate distributions

Let $X \sim F$ and $Y \sim G$ for arbitrary distributions F and G. Additionally, select $\alpha \in (0, 2)$ such that $E|X|^\alpha$, $E|Y|^\alpha < \infty$. Let $\boldsymbol{X}_n = \{X_i : i = 1, 2, \ldots, n\}$ be n independent observations with $X_i \sim F$, and $\boldsymbol{Y}_m = \{Y_j : j = 1, \ldots, m\}$ are m independent observations with $Y_j \sim G$. Furthermore, we assume full mutual independence between all observations, $\boldsymbol{X}_n \perp\!\!\!\perp \boldsymbol{Y}_m$. Then Lemmas 8 and 9 allow for the construction of the following sample divergence measure.

$$\widehat{\mathcal{E}}(\boldsymbol{X}_n, \boldsymbol{Y}_m; \alpha) = \frac{2}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} |X_i - Y_j|^\alpha - \binom{n}{2}^{-1} \sum_{1 \le i < k \le n} |X_i - X_k|^\alpha - \binom{m}{2}^{-1} \sum_{1 \le j < k \le m} |Y_j - Y_k|^\alpha. \quad (3.1)$$

By the strong law of large numbers for U-statistics [35] $\widehat{\mathcal{E}}(\boldsymbol{X}_n, \boldsymbol{Y}_m; \alpha) \xrightarrow{a.s.} \mathcal{E}(X, Y; \alpha)$ as $n \wedge m \to \infty$. Equation 3.1 allows for an estimation of $\mathcal{D}(X, Y; \alpha)$ without performing high dimensional integration. Furthermore, let

$$\widehat{\mathcal{Q}}(\boldsymbol{X}_n, \boldsymbol{Y}_m; \alpha) = \frac{mn}{m+n} \widehat{\mathcal{E}}(\boldsymbol{X}_n, \boldsymbol{Y}_m; \alpha)$$

denote the scaled empirical divergence. Under the null hypothesis of equal distributions, i.e., $\mathcal{E}(X, Y; \alpha) = 0$, [67] show that $\widehat{\mathcal{Q}}(\boldsymbol{X}_n, \boldsymbol{Y}_m; \alpha)$ converges in distribution to a non-degenerate random variable $\mathcal{Q}(X, Y; \alpha)$ as $m \wedge n \to \infty$. Specifically,

$$\mathcal{Q}(X, Y; \alpha) = \sum_{i=1}^{\infty} \lambda_i Q_i$$

in which $\lambda_i \ge 0$ are constants that depend on α and the distributions of X and Y, and the Q_i are iid chi-squared random variables with one degree of freedom. Under the alternative hypothesis of unequal distributions, i.e., $\mathcal{E}(X, Y; \alpha) > 0$, $\widehat{\mathcal{Q}}(\boldsymbol{X}_n, \boldsymbol{Y}_m; \alpha) \to \infty$ almost surely as $m \wedge n \to \infty$.

Using these facts we are able to develop two hierarchical methods for performing change point analysis, which we present in Sections 3.3 and 3.4.

3.3 Hierarchical divisive estimation

We first present the method for performing hierarchical divisive estimation of multiple change points. Here multiple change points are estimated by iteratively applying a procedure for locating a single change point. At each iteration a new change point location is estimated so that it divides an existing segment. As a result, the progression of this method can be diagrammed as a binary tree. In this tree, the root node corresponds to the case of no change points, and thus contains the entire time series. All other non-root nodes are either a copy of their parent, or correspond to one of the new segments created by the addition of a change point to their parent.

Let $Z_1, \ldots, Z_T \in \mathbb{R}^d$ be an independent sequence of observations and let $1 \leq \tau < \kappa \leq T$ be constants. Now define the following sets, $\boldsymbol{X}_\tau = \{Z_1, Z_2, \ldots, Z_\tau\}$ and $\boldsymbol{Y}_\tau(\kappa) = \{Z_{\tau+1}, \ldots, Z_\kappa\}$. A change point location $\hat{\tau}$ is estimated as

$$(\hat{\tau}, \hat{\kappa}) = \underset{(\tau,\kappa)}{\operatorname{argmax}} \ \widehat{\mathcal{Q}}(\boldsymbol{X}_\tau, \boldsymbol{Y}_\tau(\kappa); \alpha).$$

In [75] it is shown that binary segmentation procedures may not be able to detect some change points in a multiple change point setting. The variable κ is introduced in an attempt to overcome this problem by allowing for the examination of smaller segments within the series. Within these smaller segments, binary segmentation will be able to detect all change points.

The statistical significance of a change point is determined through a permuta-

tion test, since the distribution of $\hat{Q}(X, Y; \alpha)$ depends on the *unknown* distributions of the observations. In the case of independent observations, [53] show that this procedure generates strongly consistent change point estimates. A more complete outline of the divisive approach is given in the Appendix.

The signature of the method used to perform analysis based on this divisive approach is

```
e.divisive(X, sig.lvl = 0.05, R = 199, eps = 1e-3,
           half = 1000, k = NULL, min.size = 30, alpha = 1)
```

The arguments of this function are:

- `X` - A $T \times d$ matrix representation of a length T time series, with d-dimensional observations.
- `sig.lvl` - The marginal significance level used for the sequence of permutation tests.
- `R` - The maximum number of permutations to perform in the permutation test. The estimated p-value is calculated using the method outlined in [25].
- `eps` - The uniform error bound on the resampling risk [25].
- `half` - A constant used to control the epsilon spending rate, see [25] for further details.
- `k` - The number of change points to return. If this is `NULL` only the statistically significant estimated change points are returned.
- `min.size` - The minimum number of observations between change points.
- `alpha` - The index for the test statistic, as described in Section 3.2.

The returned value is a list with the following components:

- `estimates` - A vector containing the estimated change point locations.
- `cluster` - The estimated cluster membership vector.
- `k.hat` - The number of segments created by the estimated change points.
- `order.found` - The estimated change point locations in the order in which they were estimated.
- `considered.last` - The location of the last estimated change point that was not deemed statistically significant.
- `p.values` - The approximate p-values returned by the sequence of permutation tests.
- `permutations` - The number of permutations performed by each of the sequential permutation test.

The time complexity of this method is $O(kT^2)$, where k is the number of estimated change points, and T is the number of observations in the series.

3.3.1 Examples

We present some examples which illustrate the use of the `e.divisive` method.

Change in univariate normal distribution

We begin with the simple case of identifying changes in univariate normal distributions. The following example provides the output when using the method with

different values of α. As can be seen, if $\alpha = 2$ the `e.divisive` method can only identify changes in mean. For this reason, it is recommended that α is selected so as to lie in the interval $(0, 2)$. Figure 3.1 depicts the example time series, along with the change points associated with the results obtained by using $\alpha = 1$.

```
> set.seed(250)
> library(ecp)
> period1 <- rnorm(100)
> period2 <- rnorm(100,0,3)
> period3 <- rnorm(100,2,1)
> period4 <- rnorm(100,2,4)
> X.3.1.1 <- matrix(c(period1,period2,period3,period4),ncol=1)
> output1 <- e.divisive(X.3.1.1, R = 499, alpha = 1)
> output2 <- e.divisive(X.3.1.1, R = 499, alpha = 2)
> output2$estimates

[1]   1 201 358 401

> output1$k.hat

[1] 4

> output1$order.found

[1]   1 401 201 308 108

> output1$estimates
```

```
[1]   1 108 201 308 401
```

```
> output1$considered.last
```

```
[1] 358
```

```
> output1$p.values
```

```
[1] 0.002 0.002 0.010 1.000
```

```
> output1$permutations
```

```
[1] 499 499 499   5
```

```
> ts.plot(X.3.1.1,ylab='Value',
+         main='Change in a Univariate Gaussian Sequence')
> abline(v=c(101,201,301),col='blue')
> abline(v=output1$estimates[c(-1,-5)],col='red',lty=2)
```

Multivariate change in covariance

Next we apply the `e.divisive` method to multivariate data. In this example the marginal distributions remain the same, while the joint distribution changes. Therefore, applying a univariate change point procedure to each margin, such as those implemented by the `changepoint`, `cmp`, and `bcp` packages, will not detect the change. The observations in this example are drawn from trivariate normal distributions with differing correlation matrices. Observations are generated by using the `mvtnorm` package [26].

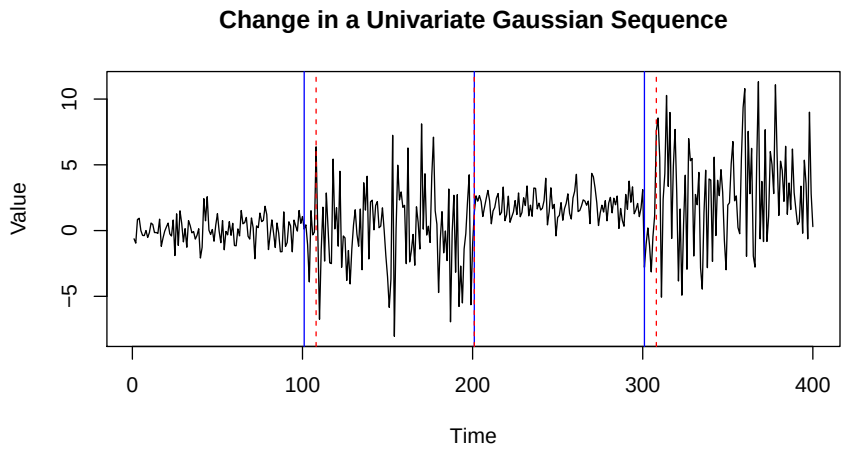

Figure 3.1: Simulated Gaussian data with 3 change points

Simulated independent Gaussian observations with changes in mean or variance. Dashed vertical lines indicate the estimated change point locations. Solid vertical lines indicate the true change point locations.

```
> set.seed(200)
> library(ecp)
> library(mvtnorm)
> mu <- rep(0,3)
> covA <- matrix(c(1,0,0,0,1,0,0,0,1),3,3)
> covB <- matrix(c(1,0.9,0.9,0.9,1,0.9,0.9,0.9,1),3,3)
> period1 <- rmvnorm(250, mu, covA)
> period2 <- rmvnorm(250, mu, covB)
> period3 <- rmvnorm(250, mu, covA)
> X.3.1.2 <- rbind(period1, period2, period3)
> output <- e.divisive(X.3.1.2, R = 499, alpha = 1)
> output$estimates
```

```
[1]   1 250 502 751
```

Multivariate change in tails

In this section we provide a second multivariate example. In this case, the change in distribution is caused by a change in tail behavior. Data points are drawn from a bivariate normal distribution and a bivariate Student's t-distribution with 2 degrees of freedom. Figure 3.2 depicts the different samples within the time series.

```
> set.seed(100)
> library(ecp)
> library(mvtnorm)
> mu <- rep(0,2)
> period1 <- rmvnorm(250, mu, diag(2))
> period2 <- rmvt(250, sigma = diag(2), df = 2)
> period3 <- rmvnorm(250, mu, diag(2))
> X.3.1.3 <- rbind(period1, period2, period3)
> output <- e.divisive(X.3.1.3, R = 499, alpha = 1)
> output$estimates
```

```
[1]   1 257 504 751
```

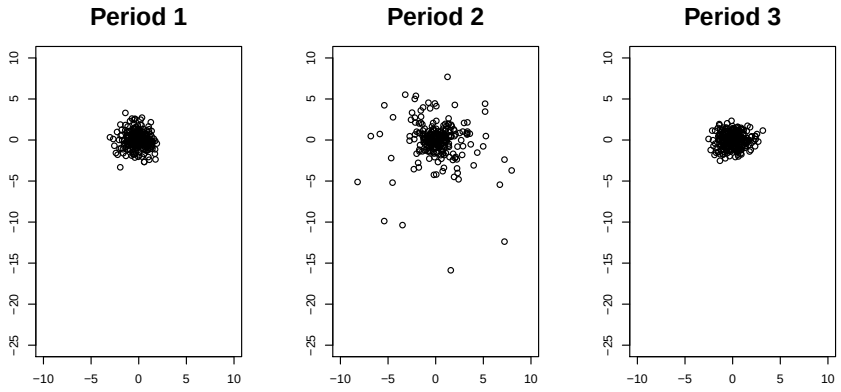

Figure 3.2: Simulated multivariate data with 2 changes in tail behavior

Data set used for example in Section 3.3.1. Periods 1 and 3 contain independent bivariate Gaussian observations with mean vector $(0, 0)^\top$ and identity covariance matrix. The second time period contains independent observations from a bivariate Student's t-distribution with 2 degrees of freedom and identity covariance matrix.

3.3.2 Real data

In this section we analyze the results obtained by applying the `e.divisive` method to two real datasets. We first apply the `e.divisive` method to the micro-array aCGH data from [8]. In this data set we are provided with records of the copy-number variations for multiple individuals. Next we apply the `e.divisive` method to a set of financial time series. For this we consider weekly log returns of the companies which compose the Dow Jones Industrial Average.

Micro-array data

This dataset consists of micro-array data for 57 different individuals with a bladder tumor. Since all individuals have the same disease, we would expect the change point locations to be almost identical on each micro-array set. The approach taken by [8] assumes that each micro-array can be modeled by a piecewise constant function, and is thus focused on changes in mean. To contrast, both the MultiRank and E-Divisive approaches are able to detect changes in mean, but can also detect other changes such as changes in variability.

The original dataset from [8] contained missing values, and thus our procedure could not be directly applied. We therefore, removed all individuals for which more than 7% of the value were missing. The remaining missing values we replaced by the average of their neighboring values. After performing this cleaning process, we were left with a sample of $d = 43$ individuals, which can be accessed by `data(ACGH)`. When applied to the full 43 dimensional series, the MultiRank procedure estimated 43 changes points, while the E-Divisive algorithm estimated 97. Figures 3.3 and 3.4 provide the results of applying the `e.divisive` and MultiRank methods to a subsample of two individuals (persons 10 and 15). The `e.divisive` procedure was run with `alpha=1`, `min.size=15`, and `R=499`. The marginal series are plotted, and the dashed lines are the estimated change point locations.

Financial data

Next we consider weekly log returns for the companies which compose the Dow Jones Industrial Average (DJIA). The time period under consideration is April 1990 to January 2012, thus providing us with 1140 observations. Since the time

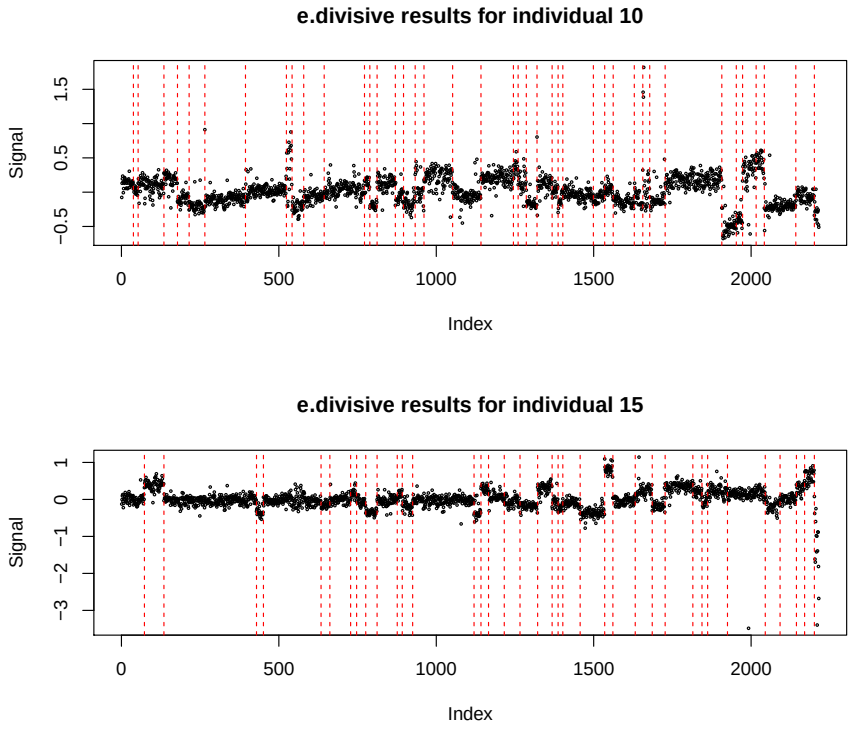

Figure 3.3: E-Divisive applied to two aCGH datasets

The aCGH data for individuals 10 and 15. The `e.divisive` procedure was run with the minimum segment size set to 15. Estimated change points are indicated by dashed vertical lines.

series for Kraft Foods Inc. does not span this entire period, it is not included in our analysis. This dataset is accessible by running `data(DJIA)`.

When applied to the 29 dimensional series, the `e.divisive` method identified change points at 7/13/98, 3/24/03, 9/15/08, and 5/11/09. The change points at 5/11/09 and 9/15/08 correspond to the release of the Supervisory Capital Asset Management program results, and the Lehman Brothers bankruptcy filing, respectively.

For comparison we also considered the univariate time series for the DJIA In-

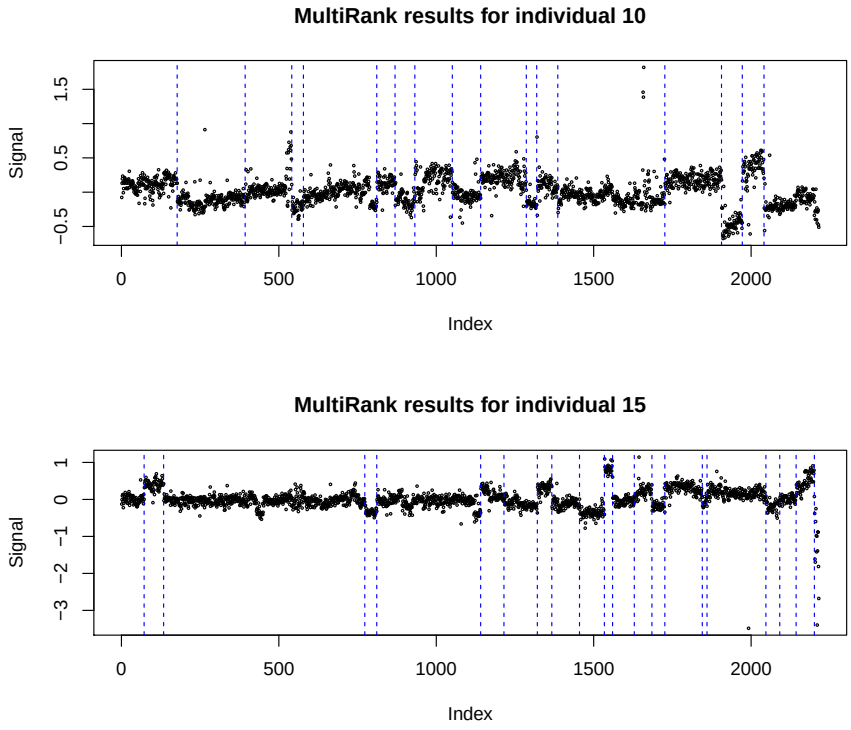

Figure 3.4: MultieRank applied to two aCGH datasets

The aCGH data for individuals 10 and 15. The MultiRank procedure was run with the ability to estimate at most 147 change points. Estimated change point locations are indicated by dashed vertical lines.

dex weekly log returns. In this setting, change points were identified at 10/21/96, 3/31/03, 10/15/07, and 3/9/09. Once again, some of these change points correspond to major financial events. The change point at 3/9/09 can be attributed to Moody's rating agency threatening to downgrade Wells Fargo & Co., JP Morgan Chase & Co., and Bank of America Corp. The 10/15/07 change point is located around the time of the financial meltdown caused by subprime mortgages. In both the univariate and multivariate cases the change point in March 2003 is around the time of the 2003 U.S. invasion of Iraq. A plot of the DJIA weekly log returns is provided in Figure 3.5 along with the locations of the estimated change points.

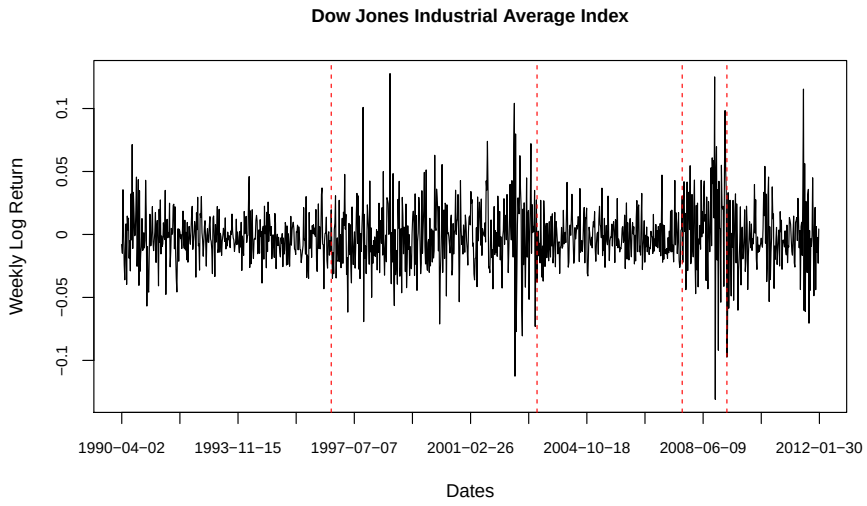

Figure 3.5: Weekly log returns for the Dow Jones Industrial Average

Weekly log returns for the Dow Jones Industrial Average index from April 1990 to January 2012. The dashed vertical lines indicate the locations of estimated change points. The estimated change points are located at 10/21/96, 3/31/03, 10/15/07, and 3/9/09.

3.4 Hierarchical agglomerative estimation

We now present a method for performing hierarchical agglomerative estimation of multiple change points. This method requires that an initial segmentation of the data be provided. This initial segmentation can help to reduce the computational time of the procedure. It also allows for the inclusion of *a priori* knowledge of possible change point locations, however if no such assumptions are made, then each observation can be assigned to its own segment. Neighboring segments are then sequentially merged to maximize a goodness-of-fit statistic. The estimated change point locations are determined by the iteration which maximized the penalized goodness-of-fit statistic. When using the `e.agglo` procedure it is assumed

that there is at least one change point present within the time series.

The goodness-of-fit statistic used in [53] is the between-within distance [73] among adjacent segments. Let $C = \{C_1, \ldots, C_n\}$ be a segmentation of the T observations into n segments, such that each segment C_i contains m_i contiguous observations. The goodness-of-fit statistic is defined as

$$\widehat{\mathcal{S}}_n(C;\alpha) = \sum_{i=1}^{n} \widehat{Q}(C_i, C_{i+1};\alpha)$$

which is equivalent to

$$\sum_{i=1}^{n}\left[\left(\frac{2}{m_i+m_{i+1}}\sum_{\substack{Z_j\in C_i\\ Z_k\in C_{i+1}}} Z_{jk}^{\alpha}\right) - \left(\frac{2m_{i+1}}{(m_i-1)(m_i+m_{i+1})}\sum_{\substack{Z_j,Z_k\in C_i\\ j<k}} Z_{jk}^{\alpha}\right) - \left(\frac{2m_i}{(m_{i+1}-1)(m_i+m_{i+1})}\sum_{\substack{Z_j,Z_k\in C_{i+1}\\ j<k}} Z_{jk}^{\alpha}\right)\right],$$

in which C_i and C_{i+1} are adjacent segments, $C_{n+1} = C_1$, and $Z_{jk}^{\alpha} = |Z_j - Z_k|^{\alpha}$. At each stage of the agglomerative process, the adjacent segments that are merged are those that result in the greatest increase (or smallest decrease) of the goodness-of-fit statistic. Therefore, for an initial segmentation with n segments, this procedure generates a sequence of $n-1$ goodness-of-fit statistics $\widehat{\mathcal{S}}_k$.

If overfitting is a concern, it is possible to penalize the sequence of goodness-of-fit statistics. This is accomplished through the use of the `penalty` argument, which generates a penalty based upon change point locations. Thus, the change point locations are estimated by maximizing

$$\widetilde{\mathcal{S}}_k = \widehat{\mathcal{S}}_k + \texttt{penalty}(\vec{\tau}(k))$$

where $\vec{\tau}(k)$ is the set of change points associated with the goodness-of-fit statistic $\widehat{\mathcal{S}}_k$. Examples of penalty terms include

```
penalty1 = function(cp){-length(cp)}
penalty2 = function(cp){mean(diff(sort(cp)))}
```

in which `penalty1` penalizes based upon the number of change points, while `penalty2` penalizes based upon the average distance between change points.

The signature of the method used to perform agglomerative analysis is

```
e.agglo(X, member = 1:nrow(X), alpha = 1, penalty =
function(cp){0})
```

The function's arguments are:

- `X` - A $T \times d$ matrix representation of a length T time series, with d-dimensional observations.
- `member` - A numeric vector that provides the initial cluster membership for each observation.
- `alpha` - The index for the test statistic, as described in Section 3.2.
- `penalty` - A function used to penalize the obtained goodness-of-fit statistics. The input for this function is a vector of change point locations `cp`.

The returned value is a list with the following components:

- `opt` - The locations of the estimated change points for the maximized goodness-of-fit statistic with penalization.
- `fit` - A vector detailing the progression of the penalized goodness-of-fit statistic.
- `cluster` - The estimated cluster membership vector.
- `merged` - A $(T-1) \times 2$ matrix indicating which segments were merged at each step of the agglomerative procedure.
- `progression` - A $T \times (T+1)$ matrix detailing the progression of the set of change point estimates.

The update from $\widehat{\mathcal{S}}_k$ to $\widehat{\mathcal{S}}_{k-1}$ is $O(1)$, so the overall computational complexity is $O(T^2)$. Like the `e.divisive` method, this is quadratic in the number of observations, however its complexity does not depend on the number of estimated change points.

3.4.1 Examples

In this section we present two examples that demonstrate the use of the `e.agglo` method.

Change in normal distributions

In this example, we use the data set created in Section 3.3.1. The number and locations of change points are then estimated by using the agglomerative procedure. Since the `e.agglo` method requires an initial segmentation of the data we have chosen to create 40 equally sized segments.

```
> library(ecp)
> member <- rep(1:40,rep(10,40))
> output <- e.agglo(X = X.3.1.1, member = member, alpha = 1)
> output$opt

[1]   1 101 201 301 401

> tail(output$fit,5)

[1] 100.05695 107.82542 104.30608 102.64330 -17.10722
```

```
> output$progression[1,1:10]

[1]  1 11 21 31 41 51 61 71 81 91

> output$merged[1:4,]

     [,1] [,2]
[1,]  -39  -40
[2,]   -1   -2
[3,]  -38    1
[4,]    2   -3
```

Multivariate change in covariance

This example illustrates the use of the `e.agglo` method with multivariate observations. We apply the `e.agglo` procedure to the trivariate data from Section 3.3.1. The data is initially segmented into 15 equally sized segments.

```
> library(ecp)
> member <- rep(1:15,rep(50,15))
> pen = function(x) -length(x)
> output1 <- e.agglo(X = X.3.1.2, member = member, alpha = 1)
> output2 <- e.agglo(X = X.3.1.2, member = member, alpha = 1,
+                    penalty = pen)
> output1$opt

[1] 1 101 201 301 351 501 601 701 751
```

```
> output2$opt
```

```
[1] 301 501
```

In this case, if we don't penalize the procedure it generates too many change points, as can be seen by the retsult of `output1`. When penalizing based upon the number of change points we obtain a much more accurate result, as shown by `output2`. Here the `e.agglo` method has indicated that observations 1 through 300 and observations 501 through 750 are identically distributed.

3.4.2 Inhomogeneous spatio-temporal point process

We apply the `e.agglo` procedure to a spatio-temporal point process. The examined data set consist of 10,498 observations, each with associated time and spatial coordinates. This data set spans the time interval $[0, 7]$ and has spatial domain $\mathbb{R}^2$. It contains 3 change points, which occur at times $t_1 = 1, t_2 = 3$, and $t_3 = 4.5$. Over each of these subintervals, $[t_i, t_{i+1}]$ the process is an inhomogeneous Poisson point process with intensity function $\lambda(s,t) = f_i(s)$, a 2-d density function, for $i = 1, 2, 3, 4$. This intensity function is chosen to be the density function for a mixture of 3 bivariate normal distributions,

$$\mathcal{N}\left(\begin{pmatrix}-7\\-7\end{pmatrix},\begin{pmatrix}25 & 0\\0 & 25\end{pmatrix}\right),\quad \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}9 & 0\\0 & 1\end{pmatrix}\right),\text{ and }\quad \mathcal{N}\left(\begin{pmatrix}5.5\\0\end{pmatrix},\begin{pmatrix}9 & 0.9\\0.9 & 9\end{pmatrix}\right).$$

For the time periods, $[0, 1], (1, 3], (3, 4.5]$, and $(4.5, 7]$ the respective mixture parameters are

$$\left(\frac{1}{3},\frac{1}{3},\frac{1}{3}\right),\ \left(\frac{1}{5},\frac{1}{2},\frac{3}{10}\right),\ \left(\frac{7}{20},\frac{3}{10},\frac{7}{20}\right),\text{ and }\left(\frac{1}{5},\frac{3}{10},\frac{1}{2}\right).$$

To apply the `e.agglo` procedure we initially segment the observations into 50 segments such that each segment spans an equal amount of time. At its termination, the `e.algo` procedure, with no penalization, identified change points at times 0.998, 3.000, and 4.499. These results can be obtained with the following

```
> library(mvtnorm); library(combinat); library(MASS); library(ecp)
> set.seed(2013)
>
> lambda = 1500  # This is the overall arrival rate per unit time.
> #set of distribution means
> muA = c(-7,-7); muB = c(0,0); muC = c(5.5,0)
> #set of distribution covariance matrices
> covA = 25*diag(2); covB = matrix(c(9,0,0,1),2)
> covC = matrix(c(9,.9,.9,9),2)
> #time intervals
> time.interval = matrix(c(0,1,3,4.5,1,3,4.5,7),4,2)
> #mixing coefficents
> mixing.coef = rbind(c(1/3,1/3,1/3),c(.2,.5,.3),
+                     c(.35,.3,.35), c(.2,.3,.5))
>
> stppData = NULL
> for(i in 1:4){
+  count = rpois(1, lambda* diff(time.interval[i,]))
+  Z = rmultz2(n = count, p = mixing.coef[i,])
+  S = rbind(rmvnorm(Z[1],muA,covA), rmvnorm(Z[2],muB,covB),
+            rmvnorm(Z[3],muC,covC))
+  X = cbind(rep(i,count), runif(n = count, time.interval[i,1],
```

```
+             time.interval[i,2]), S)
+  stppData = rbind(stppData, X[order(X[,2]),])
+ }
>
> member = as.numeric(cut(stppData[,2], breaks = seq(0,7,by=1/12)))
> output = e.agglo(X = stppData[,3:4], member = member, alpha = 1)
```

The `e.agglo` procedure was also run on the above data set using the following penalty function,

- `pen = function(cp){ -length(cp) }`

When using `pen`, change points were also estimated at times 0.998, 3.000, 4.499 The progression of the goodness-of-fit statistic for the different schemes is plotted in Figure 3.6. A comparison of the true densities and the estimated densities obtained from the procedure's results with no penalization are done in Figures 3.7 and 3.8, respectively. As can be see, the estimated results obtained from the `e.agglo` procedure provide a reasonable approximation to the true densities.

3.5 Performance analysis

To compare the performance of different change point methods we used the Rand Index [64] as well as Morey and Agresti's Adjusted Rand Index [55]. These indices provide a measure of similarity between two different segmentations of the same set of observations.

Suppose that the two clusterings of the T observations are given by U =

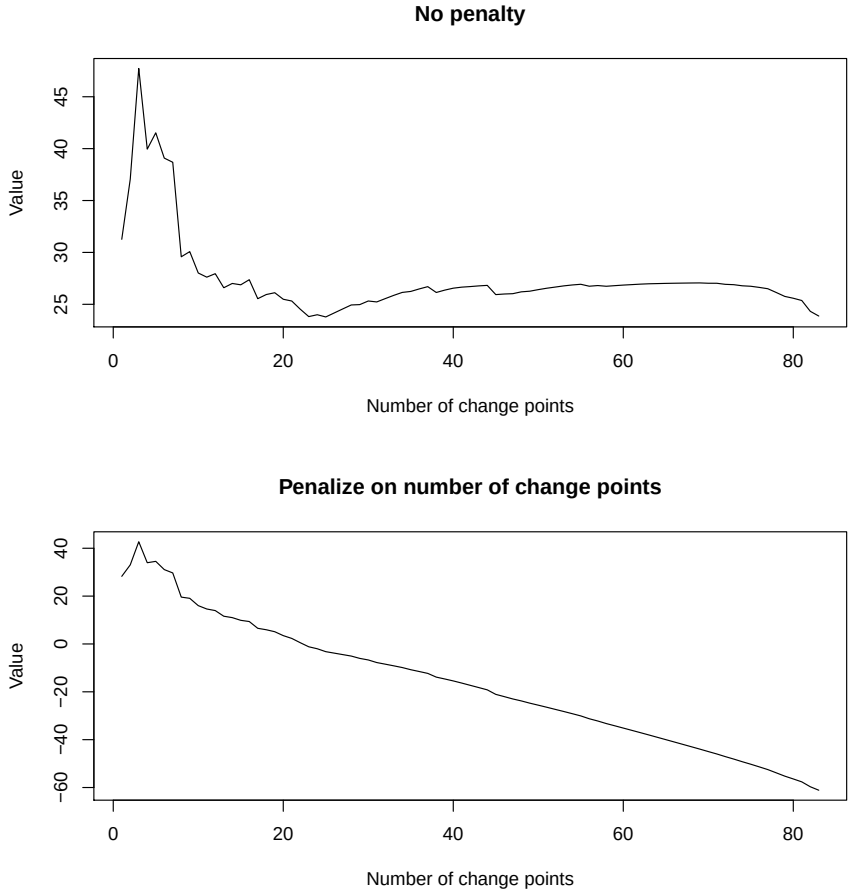

Figure 3.6: E-Agglomerative goodness-of-fit values

The progression of the goodness-of-fit statistic for the various penalization schemes.

$\{U_1, \ldots, U_a\}$ and $V = \{V_1, \ldots, V_b\}$, with a and b clusters, respectively. The Rand Index evaluates similarity by examining the cluster membership of *pairs* of observations. Consider the pairs of observations that belong to the following sets:

{A} Pairs of observations that are in the same cluster under both U & V.

{B} Pairs of observations that are in different clusters under both U & V.

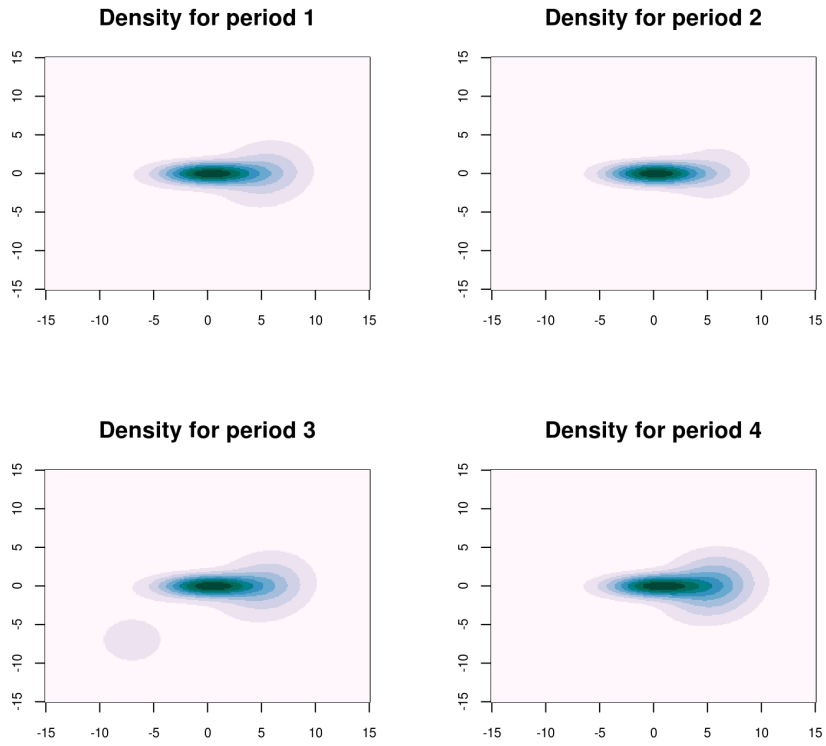

Figure 3.7: True density plots for simulates spatio-temporal point process

True density plots for the different segments of the spatio-temporal point process in Section 3.4.2.

The Rand Index is then defined as

$$\text{Rand} = \frac{\#A + \#B}{\binom{T}{2}}.$$

A shortcoming of the Rand Index is that it does not measure departure from a given baseline model, thus making it difficult to compare two different estimated clusterings. The hypergeometric model is a popular choice for the baseline, and is used by [39] and [21]. This choice of model conditions on the number of clusters as well as their sizes. Under this model we are able to determine the expected value

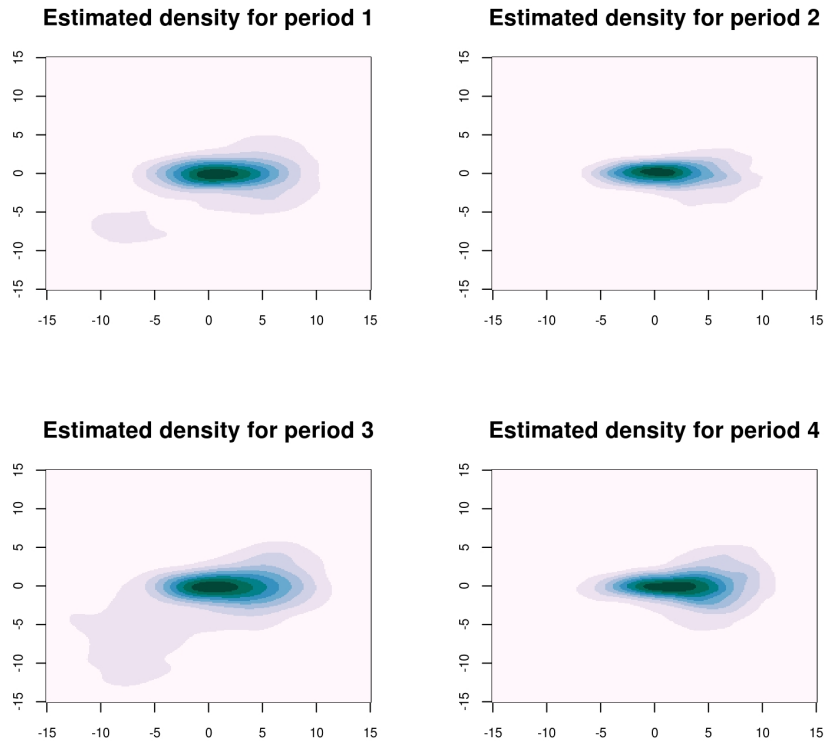

Figure 3.8: Estimated density plots for simulates spatio-temporal point process

Estimated density plots for the estimated segmentation provided by the `e.algo` procedure when applied to the spatio-temporal point process in Section 3.4.2.

of the Rand Index, and using this information the Adjusted Rand Index is

$$\text{Adjusted Rand} = \frac{\text{Rand} - \text{Expected Rand}}{1 - \text{Expected Rand}}.$$

By using the Rand and Adjusted Rand Indices we are able to assess the performance of change point procedures, and compare the performance of different change point procedures. When used to compare different change point procedures, the Rand and Adjusted Rand Indices are sensitive to the two main factors of a change point model; the number of change points, and their locations. In our simulation study the Rand and Adjusted Rand Indices are determined by

comparing the segmentation created by a change point procedure and the true segmentation. We compare the performance of our `e.divisive` procedure against that of our `e.agglo`. The results of the simulations are provided in Tables 3.1 and 3.2. Table 3.1 provides the results for simulations with univariate time series, while Table 3.2 provides the results for the multivariate time series. In these tables, average Rand Index along with standard errors are reported for 1000 simulations. Although not reported, similar results are obtained for the average Adjusted Rand Index.

Both the Rand Index and Adjusted Rand Index can be easily obtained through the use of the `adjustedRand` function in the `clues` package [12]. If `U` and `V` are membership vectors for two different segmentations of the data, then the required index values are obtained as follows,

```
> library(clues)
> RAND <- adjustedRand(U,V)
```

The Rand Index is stored in `RAND[1]`, while `RAND[2]` and `RAND[3]` store various Adjusted Rand indices. These Adjusted Rand indices make different assumptions on the baseline model, and thus arrive at different values for the expected Rand index.

3.6 Conclusion

The `ecp` package is able to perform nonparametric change point analysis of multivariate data. The package provides two primary methods for performing analysis,

	Change in Mean			Change in Variance			Change in Tail		
T	μ	E-Divisive	E-Agglo	σ^2	E-Divisive	E-Agglo	ν	E-Divisive	E-Agglo
150	1	$0.950_{0.001}$	$0.964_{0.004}$	2	$0.907_{0.003}$	$0.914_{0.012}$	16	$0.835_{0.017}$	$0.544_{6.1\times10^{-4}}$
	2	$0.992_{4.6\times10^{-4}}$	$0.991_{0.001}$	5	$0.973_{0.001}$	$0.961_{0.002}$	8	$0.836_{0.020}$	$0.543_{5.9\times10^{-4}}$
	4	$1.000_{3.7\times10^{-5}}$	$1.000_{0.000}$	10	$0.987_{7.1\times10^{-4}}$	$0.978_{0.002}$	2	$0.841_{0.011}$	$0.545_{7.5\times10^{-4}}$
300	1	$0.972_{9.1\times10^{-4}}$	$0.953_{0.002}$	2	$0.929_{0.003}$	$0.948_{0.021}$	16	$0.791_{0.015}$	$0.552_{2.1\times10^{-4}}$
	2	$0.996_{2.2\times10^{-4}}$	$0.994_{6.4\times10^{-4}}$	5	$0.990_{5.1\times10^{-4}}$	$0.976_{0.001}$	8	$0.729_{0.018}$	$0.551_{2.2\times10^{-4}}$
	4	$1.000_{1.0\times10^{-5}}$	$1.000_{0.000}$	10	$0.994_{3.2\times10^{-4}}$	$0.988_{8.9\times10^{-4}}$	2	$0.815_{0.006}$	$0.551_{2.3\times10^{-4}}$
600	1	$0.987_{1.5\times10^{-5}}$	$0.970_{0.001}$	2	$0.968_{0.001}$	$0.551_{2.3\times10^{-4}}$	16	$0.735_{0.019}$	$0.552_{2.1\times10^{-4}}$
	2	$0.998_{3.9\times10^{-6}}$	$0.997_{3.0\times10^{-4}}$	5	$0.995_{2.2\times10^{-4}}$	$0.983_{8.7\times10^{4}}$	8	$0.743_{0.025}$	$0.551_{2.2\times10^{-4}}$
	4	$1.000_{3.1\times10^{-7}}$	$1.000_{0.000}$	10	$0.998_{1.5\times10^{-4}}$	$0.992_{5.5\times10^{-4}}$	2	$0.817_{0.006}$	$0.552_{2.3\times10^{-4}}$

Table 3.1: Results for E-Agglomerative and E-Divisive univariate simulations

Average Rand Index and standard errors from 1,000 simulations for the E-Divisive and E-Agglo methods. Each sample has $T = 150, 300$ or 600 observations, consisting of three equally sized clusters, with distributions $N(0,1), G, N(0,1)$, respectively. For changes in mean $G \equiv N(\mu, 1)$, with $\mu = 1, 2,$ and 4; for changes in variance $G \equiv N(0, \sigma^2)$, with $\sigma^2 = 2, 5,$ and 10; and for changes in tail shape $G \equiv t_\nu(0,1)$, with $\nu = 16, 8,$ and 2.

each of which is able to determine the number of change points without user input. The only necessary user-provided parameter, apart from the data itself, is the choice of α. If α is selected to lie in the interval $(0, 2)$ then the methods provided by this package are able to detect *any* type of distributional change within the observed series, provided that the absolute αth moments exists.

The `e.divisive` method sequentially tests the statistical significance of each change point estimate given the previously estimated change estimates, while the `e.agglo` method proceeds by optimizing a goodness-of-fit statistic. For this reason, we prefer to use the `e.divisive` method, even though its running time is output-sensitive and depends on the number of estimated change points.

	Change in Mean			Change in Correlation		
T	μ	E-Divisive	E-Agglo	ρ	E-Divisive	E-Agglo
300	1	$0.987_{4.7\times10^{-4}}$	$0.978_{0.001}$	0.5	$0.712_{0.018}$	$0.551_{2.5\times10^{-4}}$
	2	$0.992_{8.9\times10^{-5}}$	$0.999_{2.4\times10^{4}}$	0.7	$0.758_{0.021}$	$0.552_{2.4\times10^{-4}}$
	3	$1.000_{1.3\times10^{-5}}$	$1.000_{0.000}$	0.9	$0.769_{0.017}$	$0.550_{3.1\times10^{-4}}$
600	1	$0.994_{2.2\times10^{-4}}$	$0.986_{8.6\times10^{4}}$	0.5	$0.652_{0.022}$	$0.553_{1.4\times10^{-4}}$
	2	$1.000_{4.3\times10^{-5}}$	$0.999_{1.5\times10^{-4}}$	0.7	$0.650_{0.017}$	$0.553_{1.5\times10^{-4}}$
	3	$1.000_{3.3\times10^{-6}}$	$1.000_{0.000}$	0.9	$0.806_{0.019}$	$0.553_{1.8\times10^{-4}}$
900	1	$0.996_{1.6\times10^{-4}}$	$0.991_{6.0\times10^{-4}}$	0.5	$0.658_{0.024}$	$0.554_{9.9\times10^{-5}}$
	2	$1.000_{3.0\times10^{-5}}$	$1.000_{7.3\times10^{-5}}$	0.7	$0.633_{0.022}$	$0.554_{1.1\times10^{-4}}$
	3	$1.000_{5.2\times10^{-6}}$	$1.000_{2.2\times10^{-5}}$	0.9	$0.958_{0.004}$	$0.553_{1.3\times10^{-4}}$

Table 3.2: Results for E-Agglomerative and E-Divisive multivariate simulations

Average Rand Index and standard errors from 1,000 simulations for the E-Divisive and E-Agglo methods, when applied to multivariate time series with $d = 2$. Each sample has $T = 150, 300$ or 600 observations, consisting of three equally sized clusters, with distributions $N_2(0, I), G, N_2(0, I)$, respectively. For changes in mean $G \equiv N_2(\mu, I)$, with $\mu = (1, 1)^\top, (2, 2)^\top$, and $(3, 3)^\top$; for changes in correlation $G \equiv N(0, \Sigma_\rho)$, in which the diagonal elements of Σ_ρ are 1 and the off-diagonal are ρ, with $\rho = 0.5, 0.7$, and 0.9.

Through the provided examples, applications to real data, and simulations [53], we observe that the E-Divisive approach obtains reasonable estimates for the locations of change points. Currently both the `e.divisive` and `e.agglo` methods have running times that are quadratic relative to the size of the time series. Future version of this package will attempt to reduce this to a linear relationship, or provide methods that can be used to quickly provide approximations.

3.7 Appendix

This appendix provides additional details about the implementation of both the `e.divisive` and `e.agglo` methods in the `ecp` package.

3.7.1 Divisive outline

The `e.divisive` method estimates change points with a bisection approach. In Algorithms 1 and 2, segment C_i contains all observations in time interval $[\ell_i, r_i)$. Algorithm 2 demonstrates the procedure used to identify a single change point. The computational time to maximize over (τ, κ) is reduced to $O(T^2)$ by using memoization. Memoization also allows the calculations in the for loop of Algorithm 2 to be performed at most twice. The permutation test is outlined by Algorithm 3. When given the segmentation C, a permutation is only allowed to reorder observations so that they remain within their original segments.

Algorithm 1: Outline of the divisive procedure.

Inputs : Time series Z, significance level p_0, minimum segment size m, the maximum number of permutations for the permutation test R, the uniform resampling error bound eps, epsilon spending rate h, and $\alpha \in (0, 2]$.

Output: A segmentation of the time series.

Create distance matrix $Z_{ij}^{\alpha} = |Z_i - Z_j|^{\alpha}$

while *Have not found a statistically insignificant change point*

  Estimate next most likely change point location

  Test estimated change point for statistical significance

  if *Change point is statistically significant* **then**

    Update the segmentation

  end

endwhile

return Final segmentation

3.7.2 Agglomerative outline

The `e.agglo` method estimates change point by maximizing the goodness-of-fit statistic given by Equation 3.4. The method must be provided an initial segmentation of the series. Segments are then merged in order to maximize the goodness-of-fit statistic. As segments are merged, their between-within distances also need to be updated. The following result due to [73] greatly reduces the computational time necessary to perform these updates.

Lemma 10. *Suppose that C_1, C_2, and C_3 are disjoint segments with respective sizes m_1, m_2, and m_3. Then if C_1 and C_2 are merged to form the segment $C_1 \cup C_2$,*

$$\widehat{\mathcal{E}}(C_1 \cup C_2, C_3; \alpha) = \frac{m_1 + m_3}{m_1 + m_2 + m_3}\widehat{\mathcal{E}}(C_1, C_3; \alpha) + \frac{m_2 + m_3}{m_1 + m_2 + m_3}\widehat{\mathcal{E}}(C_2, C_3; \alpha) - \frac{m_3}{m_1 + m_2 + m_3}\widehat{\mathcal{E}}(C_1, C_2; \alpha).$$

Algorithm 2: Outline of procedure to locate a single change point.

Inputs : Segmentation C, distance matrix D, minimum segment size m.

Output: A triple (x, y, z) containing the following information: a segment identifier, a distance within a segment, a weighed sample divergence.

best $= -\infty$

loc $= 0$

for *Segments* $C_i \in C$

  $A =$ Within distance for $[\ell_i, \ell_i + m)$

  for $\kappa \in \{\ell_i + m + 2, \ldots, r_i + 1\}$

    Calculate and store between and within distances for currenct choice of κ

    Calculate test statistic

    if *Test statistic* $\geq$ *best* **then**

      Update best

      Update loc to m

    end

  endfor

  for $\tau \in \{\ell_i + m + 1, \ldots, r_i - m\}$

    Update within distance for left segment

    for $\kappa \in \{\tau + m + 1, \ldots, r_i + 1\}$

      Update remaining between and within distances for current choice of κ

      Calcualte test statistic

      if *Test statistic* $\geq$ *best* **then**

        Update best

        Update loc to τ

      end

    endfor

  endfor

endfor

return Which segment to divide, loc, and best

Algorithm 3: Outline of the permutation test.

Inputs : Distance matrix D, observed test statistic p, maximum number of permutations R, uniform resampling error bound eps, epsilon spending rate h, segmentation C, minimum segment size m.

Output: An approximate p-value.

over = 1

for $i \in \{1, 2, \ldots, R\}$

Permute rows and columns of D based on the segmentation C to create D'

Obtain test statistic for permuted observations

if *Permuted test statistic* $\geq$ *observed test statistic* **then**

over = over + 1

end

if *An early termination condition is satisfied* **then**

return over/(i+1)

end

endfor

return over/(R+1)

Algorithm 4 is an outline for the agglomerative procedure. In this outline $C_{i+k}(C_{i-k})$ is the segment that is k segments to the right (left) of C_i.

Algorithm 4: Outline of the agglomerative procedure.

Inputs : An initial segmentation C, a time series Z, a penalization function $f(\vec{\tau})$, and $\alpha \in (0, 2]$.

Output: A segmentation of the time series.

Create distance matrix $D_{i,j} = \widehat{\mathcal{E}}(C_i, C_j; \alpha)$

Obtain initial penalized goodness-of-fit (gof) statistic

for $K \in \{N, N+1, \ldots, 2N-3\}$

  Merge best candidate segments

  Update current gof

  if *Current gof* $\geq$ *largest gof so far* **then**

    Update largest gof

  end

endfor

Penalize the sequence of obtained gof statistics

Choose best segmentation based on penalized gof statistics

return Best segmentation

CHAPTER 4

CHANGE POINTS VIA PROBABILISTICALLY PRUNED OBJECTIVES

4.1 Introduction

The analysis of time ordered data, referred to as time series, has become a common practice in both academic and industrial settings. The applications of such analysis span may different fields, each with its own analytical tools. Such fields include network security [7, 71], fraud detection [1, 20], financial modeling [2, 18], climate analysis [77], astronomical observation [22, 78], and many others.

However, when analysis is performed it is generally assumed that the data adheres to some form of homogeneity. This could mean a range of things, depending upon the application area. Some common types of assumed homogeneity include: constant mean, constant variance, and strong or weak stationarity. Depending on the nature of these assumptions it may not be appropriate, or practical, to apply a given analytical procedure to many different types of time series. For instance, an algorithm that assumes weak stationarity would not be suitable for analyzing data that follows a Cauchy distribution, because of its infinite expectation. Furthermore many time series of real data can be seen, even through visual inspection, to violate such homogeneity conditions.

Results obtained under such model misspecification can vary in their degree of inaccuracy [16]. The resulting bias from such misspecification is one of the reasons for the current resurgence of change point analysis. Change point analysis attempts to partition a time series into homogeneous segments. Once again the definition

of homogeneity will depend upon the application area. In this paper we will use a notion of homogeneity that is common in the statistical literature. We will say that a segment is homogeneous if all of its observations are identically distributed. Using this definition of homogeneity, change point analysis can be performed in a variety of ways.

In this paper we consider the following formulation of the offline multiple change point problem. Let $Z_1, Z_2, \ldots, Z_T \in \mathbb{R}^d$ be a length T sequence of independent d-dimensional time ordered observations. The dimension of our observations is arbitrary, but assumed to be fixed. Additionally, let $F_0, F_1, F_2, \ldots,$ be a (possibly infinite) sequence of distributional functions, such that $F_i \neq F_{i+1}$. It is also assumed that in the sequence of observations, there is at least one distributional change. Thus, there exists $k(T) \geq 1$ time indices $0 = \tau_{0,T} < \tau_{1,T} < \cdots < \tau_{k(T),T} < \tau_{k(T)+1,T} = T$, such that $Z_i \overset{iid}{\sim} F_j$, for $\tau_{j,T} < i \leq \tau_{j+1,T}$. From this notation it is clear that the locations of change points $\tau_{j,T}$ depend upon the sample size. However, we will usually suppress this dependence and use the notation τ_j for simplicity. The challenge of multiple change point analysis is to provide a good estimate of both the number of change points, $k(T)$, as well as their respective locations, $\tau_1, \tau_2, \ldots \tau_{k(T)}$. In some cases it is also necessary to provide some information about the distributions $F_0, \ldots, F_{k(T)}$. However, once a segmentation is provided it is usually straightforward to obtain such information.

A popular approach is to fit the observed data to a parametric model. In this setting a change point corresponds to a change in the monitored parameter(s). Earlier work in this area assumes Gaussian observations and proceeds to partition the data through the use of maximum likelihood [50]. More recently, extensions to other members of the Exponential family of distributions and beyond have been

considered [13]. In general, all of these approaches rely on the existence of a likelihood function with an analytic expression. Once the likelihood function is known, analysis is reduced to finding a computationally efficient way to maximize the likelihood over a set of candidate parameter values.

Parametric approaches however, rely heavily upon the assumption that the data behaves according to the predefined model. If this is not the case, then the degree of bias in the obtained results is usually unknown [62]. In practice, it is almost always difficult, if not impossible, to test for adherence to these assumptions. Under such settings, performing nonparametric analysis is a natural way to proceed [10]. Since nonparametric approaches make much weaker assumptions than their parametric counterparts they can be used in a much wider variety of settings; for example, the analysis of internet traffic data, where there is no commonly accepted distributional model. Even though these methods do not directly impose a distributional model for the data, they do make their own types of assumptions [83]. For instance, a common assumption is the existence of a density function, which then allows practitioners to perform maximum likelihood estimation by using estimated densities. However, estimation becomes inaccurate and time consuming when the dimension of the time series increases [33].

Performing multiple change point analysis can easily become computationally intractable. Usually the number of true change points is not known beforehand. However, even if such information were provided, finding the locations is not a simple task. For instance, if it is known that the time series contains k change points then there are $O(T^k)$ possible segmentations. Thus naive approaches to find the best segmentation quickly become impractical. More refined techniques must therefore be employed in order to obtain change point estimates in a reasonable

amount of time.

Most existing procedures for performing retrospective multiple change point analysis can be classified as belonging to one of two groups. The first consists of search procedures which will return what are referred to as *approximate* solutions, while the second consist of those that produce *exact* solutions. As indicated by the name, the approximate procedures tend to produce suboptimal segmentations of the given time series. However, their benefit is that they tend to have provably much lower computational complexity than procedures that return optimal segmentations.

Approximate search algorithms tend to rely heavily on a subroutine for finding a single change point. Estimates for multiple change point locations are produced by iteratively applying this subroutine. Such algorithms are commonly referred to as binary segmentation algorithms. In many cases it can be shown that binary segmentation algorithms have a complexity of $O(T \log T)$. This type of approach to multiple change point analysis was introduced by [76] and has since been adapted by many others. Such adaptations include the Circular Binary Segmentation approach of [59] as well as the E-Divisive approach of [53]. The Wild Binary Segmentation approach of [23] is a variation of binary segmentation that utilizes random intervals in an attempt to further reduce computational time. An extension of this approach to multivariate multiplicative time series called Sparsified Binary Segmentation has been produced by [15]. Each of these procedures have been shown to produce consistent estimates of both the number and locations of change points under a variety of model conditions.

Exact search algorithms return segmentations that are optimal with respect to a prespecified goodness of fit measure, such as a log likelihood function. The naive

approach of searching over all possible segmentations quickly becomes impractical for relatively small time series with a few change points. Therefore, in order to achieve a reasonable computational cost, the utilized goodness of fit measures often satisfy Bellman's Principle of Optimality [6], and can thus be optimized through the use of dynamic programming. However, in most cases the ability to obtain this optimal solution comes with a computational cost. Usually this results in at least $O(T^2)$ computational complexity. Examples of exact algorithms include the Kernel Change Point algorithm, [31] and [4], and the MultiRank algorithm [49]. The complexity of these algorithms also depends upon the number of identified change points. However, a method introduced by [40], as well as the PELT algorithm of [44] can both obtain optimal segmentations with running times that are independent of the number of change points. An additional benefit of the PELT approach is that under certain conditions it is shown to have an expected running time that is linear in the length of the time series.

The second aspect of multiple change point analysis is the determination of the number of change points. The first technique that is commonly used by approximate search algorithms is hypothesis testing. This method continues to identify change points until they are unable to reject the null hypothesis of no change. Such an approach however, is not well suited for many procedures that use exact search algorithms, since many identify all change points at once, instead of sequentially, as is the case with binary segmentation. Many change point algorithms that use an exact search procedure instead turn to penalized optimization. The reasoning behind penalization is that a more complex model, in this case one with more change points, will better fit the observed data. The the penalty thus helps to guard against over segmentation. [79] showed that using the Schwarz Criterion can produce a consistent estimate of the number of change points. It has since

become popular to maximize a penalized goodness of fit measure of the following form,

$$S(k,\beta) = \max_{\tau_1<\tau_2<\cdots<\tau_k} \sum_{i=1}^{k} C(C_i) + \mathcal{P}(k), \tag{4.1}$$

for a penalty function $\mathcal{P}(\cdot)$ and measure of segmentation quality $C(\cdot)$. A common choice for the penalty function is $\mathcal{P}(k) = -\beta k$, for some user defined positive constant β. This type of penalization only takes into consideration the number of change points, and not their location. There are other penalization approaches that not only consider the number of change points, but also the change point locations. See for instance [82] and [30].

An alternative to penalization is to instead generate all optimal segmentations with k change points, up to some prespecified upper limit. This corresponds to evaluating $S(k,0)$ from Equation 4.1 for a range of k values. However, depending on the choice of $C(\cdot)$ it may not be possible to efficiently calculate $S(k,0)$ for a range of k values. And thus the search procedure would have to be run numerous times, which can become rather inefficient. Penalization tends to be faster, but does require the specification of a penalty function or constant. This choice is highly dependent upon the application field and will require some sort of knowledge about the data. Some ways to choose these parameters include cross validation [3], generalized degrees of freedom [70], and slope heuristics [5]. On the other hand, generating all optimal segmentations avoids having to make such a selection.

In the following sections we introduce a change point search procedure which we call cp3o (**C**hange **P**oints via **P**robabilisticly **P**runed **O**bjectives). This is an exact search procedure that can be applied to a larger number of goodness of fit measures in order to reduce the amount of time taken to estimate change point locations. Additionally, the cp3o algorithm allows for the number of change points

to be quickly determined without having to specify a penalty term, while at the same time generating all other optimal segmentations as a byproduct.

As the cp3o procedure can be applied to a general goodness of fit measure we propose one that is based on E-Statistics [73], which we call e-cp3o. The e-cp3o method is a nonparametric algorithm that has the ability to detect *any* type of distributional change. The use of E-Statistics also allows the e-cp3o algorithm to perform analysis on multivariate time series without suffering from the curse of dimensionality.

The results from a variety of simulations show that our method makes a reasonable trade off between speed and accuracy in most cases. In addition to the computational benefits, we show that the cp3o procedure generates consistent estimates for both the number and location of change points when equipped with an appropriate goodness of fit measure. Furthermore, under additional assumptions we also show consistency in the setting where the number of change points is increasing with the length of time series.

The remainder of this paper is organized as follows. In Section 4.2 we discuss the probabilistic pruning procedure used by cp3o, along with conditions necessary to ensure consistency. Section 4.3 is devoted to the development of the e-cp3o algorithm and showing that it satisfies the conditions outlined in Section 4.2. Results for applications to both simulated and real datasets are given in Sections 4.4 and 4.5. Concluding remarks are left for Section 4.6.

4.2 Probabilistic Pruning

When performing change point analysis one must have a quantifiable way of determining whether one segmentation is better than another. When using an exact search procedure this is most commonly accomplished through the use of a goodness of fit measure. Suppose that there are k change points $0 = \tau_0 < \tau_1 < \cdots < \tau_k < \tau_{k+1} = T$. These k locations partition the time series into $k + 1$ segments $C_j = \{Z_i : \tau_{j-1} < i \leq \tau_j\}$. The challenge now is to select the change point locations so that the observations within each segment are identically distributed, and the distribution of observations in adjacent segments are different. Therefore, we will consider sample goodness of fit measures of the following form,

$$\widehat{G}(k) = \max_{\tau_1 < \tau_2 < \cdots < \tau_k} \sum_{j=1}^{k} \widehat{R}(C_j, C_{j+1}), \tag{4.2}$$

in which $\widehat{R}(A, B)$ is a measure of the sample divergence between observation sets A and B. The divergence measure $\widehat{R}$ is such that larger values indicate that the distributions of the observations in the two sets are more distinct. Since each term of the sum in Equation 4.2 depends only upon contiguous observations it is possible to obtain the value of $\widehat{G}(k)$ through dynamic programming.

Using traditional dynamic programming approaches greatly reduces the computational time required to perform the optimization in Equation 4.2. However, the running time of such methods is still quadratic in the length of the time series, thus limiting their applicability. Many of the calculations performed during the dynamic programs do not result in the identification of a new change point. These calculations can be viewed as excessive since they do not provide any additional information about the series' segmentation, and quickly compound to slow down the algorithm. Thus a practical step towards reducing running time, and even pos-

sibly the theoretical computational complexity, is to quickly identify such excessive calculations and have them removed. One way to do this is by continually pruning the set of potential change point locations. [66] proposes a pruning method that can be used when the goodness of fit measure is convex, and can also be adapted for online change point detection. Since the sample divergence measure $\widehat{R}$ is not necessarily convex this pruning approach may not always be applicable. The cp3o procedure therefore performs pruning that is more in line with the approach taken by [44] in developing the PELT method.

Let $0 < v < t < s < u \leq T$ and $Z_a^b = \{Z_a, Z_{a+1}, \ldots, Z_b\}$ for $a \leq b$. Furthermore, suppose that there exists a constant Γ such that

$$\widehat{R}(Z_{v+1}^t, Z_{t+1}^u) - \widehat{R}(Z_{v+1}^t, Z_{t+1}^s) - \widehat{R}(Z_{t+1}^s, Z_{s+1}^u) < \Gamma$$

holds for all $v < t < s < u$. The value of Γ will depend not only on the distribution of our observations, but also the nature of the divergence measure $\widehat{R}$. Therefore, in many settings it may be difficult, if not impossible, to find such a Γ. Instead we consider the following probabilistic formulation. Let $\epsilon > 0$, we then wish to find Γ_ϵ such that for all $v < t < s < u$

$$\mathbb{P}\Big(\widehat{R}(Z_{v+1}^t, Z_{t+1}^u) - \widehat{R}(Z_{v+1}^t, Z_{t+1}^s) - \widehat{R}(Z_{t+1}^s, Z_{s+1}^u) \geq \Gamma_\epsilon\Big) \leq \epsilon.$$

Let $\zeta_k(t)$ denote the value of $\widehat{G}(k)$ when segmenting $Z_1, Z_2, \ldots, Z_t$ with k change points. Using this notation we can express our probabilistic pruning rule as follows.

Lemma 11. *Let v be the optimal change point location preceding t, with $t < s < u \leq T$. If*

$$\zeta_k(t) + \widehat{R}(Z_{v+1}^t, Z_{t+1}^s) + \Gamma_\epsilon < \zeta_k(s),$$

then with probability at least $1 - \epsilon$, t is not the optimal change point location preceding u.

Proof. If

$$\zeta_k(t) + \widehat{R}(Z^t_{v+1}, Z^s_{t+1}) + \widehat{R}(Z^s_{t+1}, Z^u_{s+1}) + \Gamma_\epsilon \quad < \quad \zeta_k(s) + \widehat{R}(Z^s_{t+1}, Z^u_{s+1}),$$

then from the definition of Γ_ϵ we have that

$$\mathbb{P}\Big(\zeta_k(t) + \widehat{R}(Z^t_{v+1}, Z^u_{t+1}) < \zeta_k(s) + \widehat{R}(Z^s_{t+1}, Z^u_{s+1})\Big) \geq 1 - \epsilon.$$

Since the optimal value attained from segmenting $Z_1, Z_2, \ldots, Z_u$ with $k+1$ change points is an upper bound for $\zeta_k(t) + \widehat{R}(Z^s_{t+1}, Z^u_{s+1})$,

$$\mathbb{P}\Big(\zeta_k(t) + \widehat{R}(Z^t_{v+1}, Z^u_{t+1}) < \zeta_{k+1}(u)\Big) \geq 1 - \epsilon.$$

From this we can see that with probability at least $1-\epsilon$, it would be better to have s as the change point prior to u. □

4.2.1 Consistency

As has been mentioned before, when performing multiple change point analysis it is of utmost importance to obtain an accurate estimate of the number of change points, as well as their locations. Therefore, in this section we will show that under a certain asymptotic setting, the estimates generated by maximizing Equation 4.2 generate consistent location estimates.

When showing consistency many authors consider the case in which the number of change points is held constant, while the number of observations tends toward infinity. This seems rather unrealistic, as one would expect to observe additional change points as more data is collected. For this reason we will allow the number of change points, $k(T)$, to possibly tend towards infinity as the length of the time series increases. The asymptotic setting we will consider is similar in nature to that taken by [75] and [83].

In order to establish consistency of the proposed estimators we make the following assumptions.

Assumption 3. *Let $\mathscr{F} = \{F_0, F_1, \dots\}$ be a collection of distribution functions, and $\{R_{j\ell}\}$ a collection of doubly indexed positive finite constants. Suppose that A and B are disjoint sets of observations, such that the observations in A have distribution F_a and those of B hare distribution F_b, for $F_a, F_b \in \mathscr{F}$. The constants $R_{j\ell}$ are such that $\hat{R}_{ab} = \widehat{R}(A, B) \to R_{ab}$ almost surely as $\min(\#A, \#B) \to \infty$. Furthermore let f be a function such that $|\hat{R}_{ab} - R_{ab}| = o(f(\#A \wedge \#B))$ almost surely for all pairs a and b.*

Assumption 4. *Let $\lambda_T = \min_{1 \le j \le k(T)} (\tau_j - \tau_{j-1})$, and suppose $\lambda_T \to \infty$ as $T \to \infty$.*

Assumption 4 states that the number of observations between change points tends towards infinity. This later allows us to apply the law of large numbers.

Assumption 5. *The number of change points $k(T)$ and its upper bound $K(T)$ are such that $k(T) = o\left(\frac{1}{f(T)}\right)$ and $k(T) = o(K(T))$.*

The above assumption controls the rate at which the number of change points can increase. This is directly related to the rate at which our sample estimates converge almost surely to their population counterparts.

Assumption 6. *Let $\mathscr{F}$ be the collection of distribution functions from Assumption 3. From this collection we define a set of random variables $\{X(a, b)\}_{a,b=0}^{\infty}$. For each pair of values $0 \le a \le b$ the random variable $X(a, b)$ has a mixture distribution created with mixture components $F_a, F_{a+1}, \dots, F_b$.*

Then for $r \le q$, and integers $0 = s_0 < s_1 < s_2 < \cdots < s_r < s_{r+1} = q + 1$, define,

$$G^q(r) = \max_{s_1, s_2, \dots, s_r} \sum_{i=1}^{r} R(X(s_{i-1}, s_i - 1), X(s_i, s_{i+1} - 1));$$

and for $r > q$ *we define* $G^q(r)$ *to be equal to* $G^q(q)$. *Assume that* $d_T = \max_{1 \le i,j \le k(T)} G^{k(T)}(i) - G^{k(T)}(j)$, *is such that* $[d_T k(T)]/[K(T)] \to 0$ *as* $T \to \infty$.

Assumption 6 concerns the rate at which additional change points increase the objective function of interest. We will show that a higher upper bound on the number of change points is necessary when each additional change point has the potential to greatly change the value of $G^q(r)$.

Assumption 7. *Let* $0 \le \pi < \gamma < \rho \le 1$ *and* $i_1 < i_2 < i_3 < i_4$ *be positive integers. Suppose that the time series* $Z_1, Z_2, \ldots, Z_T$ *is such that* $Z_{\lfloor \pi T \rfloor + 1}, \ldots, Z_{\lfloor \gamma T \rfloor} \sim X(i_1, i_2)$ *and* $Z_{\lfloor \gamma T \rfloor + 1}, \ldots, Z_{\lfloor \rho T \rfloor} \sim X(i_3, i_4)$ *for every sample of size* T. *For* $\tilde{\gamma} \in (\pi, \rho)$ *we define the following sets* $A(\tilde{\gamma}) = \{Z_{\lfloor \pi T \rfloor}, \ldots, Z_{\lfloor \tilde{\gamma} T \rfloor}\}$ *and* $B(\tilde{\gamma}) = \{Z_{\lfloor \tilde{\gamma} T \rfloor + 1}, \ldots, Z_{\lfloor \rho T \rfloor}\}$. *Assume that there exist a class of functions indexed by* π, γ, *and* ρ; $\Theta_\pi^\rho(\tilde{\gamma}|\gamma) : (\pi, \rho) \mapsto \mathbb{R}$, *such that* $\widehat{R}(A(\tilde{\gamma}), B(\tilde{\gamma})) \to \Theta_\pi^\rho(\tilde{\gamma}|\gamma) R(X(i_1, i_2), X(i_3, i_4))$ *almost surely as* $T \to \infty$. *Finally we assume that* $\Theta_\pi^\rho(\tilde{\gamma}|\gamma)$ *has a unique maximizer at* $\tilde{\gamma} = \gamma$.

Assumption 7 describes the behavior of our goodness of fit measure when it is used to identify a single change point. Essentially this assumption states that the measure will attain its maximum value when the estimated change point location, $\tilde{\gamma}$, and true change point location, γ, coincide.

Change Point Locations

We begin by showing that under Assumptions 3-7, the cp3o procedure will produce consistent estimates for the change point locations.

Lemma 12. *Let* $\mathcal{G}(k(T)) = \{\hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_{k(T)}\}$ *and*

$$\begin{aligned}\mathcal{B}_T(\epsilon) &= \mathcal{B}_T(\epsilon, \{\tau_i\}) \\ &= \left\{(\eta_1, \eta_2, \ldots, \eta_{k(T)}) \in \mathbb{R}^{k(T)} : \frac{|\eta_i - \tau_i|}{T} \leq \epsilon \text{ for } i = 1, 2, \ldots, k(T)\right\}.\end{aligned}$$

Then for all $\epsilon > 0$,

$$\mathbb{P}\left(\mathcal{G}(k(T)) \in \mathcal{B}_T(\epsilon)\right) \to 1$$

as $T \to \infty$.

Proof. Suppose that $\mathcal{G}(k(T)) \notin \mathcal{B}_T(\epsilon)$, then there exists i such that $\frac{|\hat{\tau}_i - \tau_i|}{T} > \epsilon$. Select the largest such i and define the following random variables. Let $M_1 \sim U_1$ where U_1 is the distribution (possibly a mixture) created by the observations between $\hat{\tau}_{i-2}$ and $\hat{\tau}_{i-1}$, $M_2 \sim U_2$ for U_2 having distribution created by the observations between $\hat{\tau}_{i-1}$ and τ_i. Similarly define M_3 for the observations between τ_i and $\hat{\tau}_{i+1}$, and M_4 for the observations between $\hat{\tau}_{i+1}$ and $\hat{\tau}_{i+2}$.

Then the value of the sample goodness of fit measure generated by the estimates of $\mathcal{G}(k(T))$ is

$$\hat{R}(M_1, M_2) + \hat{R}(M_2, M_3) + \hat{R}(M_3, M_4) + A,$$

which due to Assumptions 3 and 5 is equal to

$$\Theta_0^1(\beta_1|\gamma_1)R(U_1, U_2) + \Theta_0^1(\beta_2|\gamma_2)R(U_2, U_3) + \Theta_0^1(\beta_3|\gamma_3)R(U_3, U_4) + B + k(T)o(f(T)).$$

In the above expressions A and B are collections of terms that are not affected by the choice of $\hat{\tau}_i$. The β_i and γ_i terms are as listed below.

$$\beta_1 = \frac{\hat{\tau}_{i-1} - \hat{\tau}_{i-2}}{\hat{\tau}_i - \hat{\tau}_{i-2}} \quad \beta_2 = \frac{\hat{\tau}_i - \hat{\tau}_{i-1}}{\hat{\tau}_{i+1} - \hat{\tau}_{i-1}} \quad \beta_3 = \frac{\hat{\tau}_{i+1} - \hat{\tau}_i}{\hat{\tau}_{i+2} - \hat{\tau}_i}$$

$$\gamma_1 = \frac{\hat{\tau}_{i-1} - \hat{\tau}_{i-2}}{\tau_i - \hat{\tau}_{i-2}} \quad \gamma_2 = \frac{\tau_i - \hat{\tau}_{i-1}}{\hat{\tau}_{i+1} - \hat{\tau}_{i-1}} \quad \gamma_3 = \frac{\hat{\tau}_{i+1} - \tau_i}{\hat{\tau}_{i+2} - \tau_i}$$

Each of the terms in the sum is maximized when $\beta_i = \gamma_i$, which corresponds to $\frac{|\hat{\tau}_i - \tau_i|}{T} \to 0$. By our assumptions, we have that the remainder term $k(T)o(f(T)) = o(1)$. Therefore, if $\frac{|\hat{\tau}_i - \tau_i|}{T}$ is strictly bounded away from 0 then the statistic will be strictly less than the optimal value as $T \to \infty$. However, this contradicts the manner in which $\hat{\tau}_i$ is selected. □

Number of Change Points

Once we have shown that the procedure generates consistent estimate for the change point locations it is simple to show that it will also produce a consistent estimate for the number of change points. We have chosen to implement the procedure outlined below in Assumption 8 to determine the number of change points. However, other approaches could be used and still have the same consistency result.

Assumption 8. *Define* $\nabla\widehat{G}(k) = \widehat{G}(k+1) - \widehat{G}(k)$, $\widehat{\mu}(\nabla\widehat{G}) = \frac{\widehat{G}(K(T)) - \widehat{G}(1)}{K(T)-1}$ *and* $\widehat{\sigma}^2(\nabla\widehat{G}) = \widehat{Var}\{\nabla\widehat{G}(k) : k = 1, 2, \ldots, K(T) - 1\}$. *Then suppose our estimated number of change points is given by*

$$\hat{k}(T) = 1 + \max\left\{\ell : \nabla\widehat{G}(1), \ldots, \nabla\widehat{G}(\ell) > \widehat{\mu}(\nabla\widehat{G}) + \frac{1}{2}\sqrt{\widehat{\sigma}^2(\nabla\widehat{G})}\right\}. \tag{4.3}$$

The selection procedure in Equation 4.3 has similar intuition to the one presented by [47]. Both procedures work on the principle that a true change point will cause a significant change in the goodness of fit. While spurious change point estimates will only cause a minuscule increases/decrease in value. We thus say that a change is significant if it is more than half a standard deviation above the average change. As previously stated, other methods could be used, this just happens to be the one that we chose to implement.

Before proving that we can obtain a consistent estimate for the number of

change points one final assumption is made to ensure that the detection of an additional true change point causes a strictly positive increase in our goodness of fit measure. A similar property is also needed for the finite sample approximation.

Assumption 9. *For every fixed q, suppose the values $G^q(1), G^q(2), \ldots, G^q(q)$ form an increasing sequence. Similarly let $\hat{G}^q(r)$ be the finite sample estimates of $G^q(r)$. Additionally, suppose there exists T_0 such that for $T > T_0$ the values $\hat{G}^{k(T)}(1), \hat{G}^{k(T)}(2), \ldots, \hat{G}^{k(T)}(k(T))$ also form an increasing sequence.*

Lemma 13. *Let $\hat{k}(T)$ be the number of estimated change points for a sample of size T, and that the conditions of Assumptions 3-9 hold. Then*

$$\lim_{T\to\infty} \mathbb{P}\left(\hat{k}(T) = k(T)\right) = 1.$$

Proof. If $k(T)$ is bounded then the proof for the constant $k(T)$ version applies. Suppose that $\hat{k}(T) > k(T)$, then $\nabla\hat{G}^{k(T)}(k(T)) > \mu(\nabla\hat{G}) + \frac{1}{2}\sqrt{\sigma^2(\nabla\hat{G})}$. Letting $\nabla_{ij}^T = \frac{1}{2}[G^{k(T)}(i) - G^{k(T)}(j)]^2$, we note the following inequalities:

$$\begin{aligned}
\mu(\nabla\hat{G}) &= \frac{\hat{G}^{k(T)}(K(T)) - \hat{G}^{k(T)}(1)}{K(T) - 1} \\
&\leq \frac{d_T}{K(T) - 1} + o(1) \\
&\to 0.
\end{aligned}$$

$$\begin{aligned}
\sigma^2(\nabla\hat{G}) &= \frac{2}{(K(T)-1)(K(T)-2)} \sum_{1\leq i<j\leq K(T)-1} \frac{1}{2}\left[G^{k(T)}(j) - G^{k(T)}(i)\right]^2 + o(1) \\
&= \binom{K(T)-1}{2}^{-1} \left[\sum_{1\leq i<j\leq k(T)} \nabla_{ij}^T \quad + \sum_{1\leq i\leq k(T)<j\leq K(T)-1} \nabla_{ij}^T \quad + \sum_{k(T)<i<j\leq K(T)-1} \nabla_{ij}^T \right] + o(1) \\
&\leq \binom{K(T)}{2}^{-1} \left[\binom{k(T)}{2} d_T + k(T)(K(T) - k(T)) d_T + \binom{K(T) - k(T)}{2} o(1) \right] + o(1) \\
&\to 0,
\end{aligned}$$

since $G^{k(T)}(i) = G^{k(T)}(k(T))$ for $i > k(T)$. Thus $\mathbb{P}(\hat{k}(T) > k(T)) \to 0$ as $T \to \infty$.

Next suppose that $\hat{k}(T) < k(T)$. This implies that $\nabla\hat{G}^{k(T)}(\hat{k}(T)) \leq \mu(\nabla\hat{G}) + \frac{1}{2}\sqrt{\sigma^2(\nabla\hat{G})}$. However, since $\nabla G^{k(T)}(i) > 0$ for $i < k(T)$, $\mathbb{P}(\hat{k}(T) < k(T)) \to 0$ as $T \to \infty$. □

Theorem 14. *For all $\epsilon > 0$, as $T \to \infty$,*

$$\mathbb{P}\left(\hat{k}(T) = k(T), \mathcal{G}(k(T)) \in \mathcal{B}_T(\epsilon)\right) \to 1.$$

Proof. Using the results of Lemma 12 and Lemma 13 we have the following;

$$\begin{aligned}\mathbb{P}\left(\hat{k}(T) = k(T), \mathcal{G}(k(T)) \in \mathcal{B}_T(\epsilon)\right) &\geq 1 - \mathbb{P}\left(\hat{k}(T) \neq k(T)\right) \\ &\quad -\mathbb{P}\left(\mathcal{G}(k(T)) \notin \mathcal{B}_T(\epsilon)\right) \\ &\to 1.\end{aligned}$$

□

4.3 Pruning and Energy Statistics

The cp3o procedure introduced in Section 4.2 can be applied with almost any goodness of fit measure $\widehat{R}(\cdot,\cdot)$. However, in order to ensure consistency for both the estimated change point locations, as well as the estimated number of change points, some restrictions must be enforced as outlined in Section 4.2.1.

In this section we make use of a particular class of goodness of fit measures that allows for nonparametric change point analysis. These measures are indexed by $\alpha \in (0, 2)$[1] and allow for the detection of *any* type of distributional change. When

[1]The choice of $\alpha = 2$ is allowed, however, in this case the goodness of fit measure would only be able to detect changes in mean.

a value of α is selected, the only distributional assumptions that are made are that observations are independent and that they all have finite absolute αth absolute moments. This class of measures are based upon the energy statistic of [73], and we thus call the resulting procedure e-cp3o.

The e-cp3o procedure is a nonparametric procedure that makes use of an approximate test statistic and an exact search algorithm in order to locate change points. Computationally the e-cp3o procedure is comparable to other parametric/nonparametric change point methodologies that use approximate search algorithms. In the remainder of this section we give a brief review of E-Statistics, followed by their incorporation into the cp3o framework. Finally, we show that the resulting goodness of fit measure satisfies the conditions necessary for consistency.

4.3.1 The Energy Statistic

As change point analysis is directly related to the detection of differences in distribution we consider the U-statistic introduced in [73]. This statistic provides a simple way to determine whether the independent observations in two sets are identically distributed.

Suppose that we are given samples $\boldsymbol{X}_n = \{X_i : i = 1, \ldots, n\}$ and $\boldsymbol{Y}_m = \{Y_j : j = 1, \ldots, m\}$, that are independent iid samples from distributions F_X and F_Y respectively. Our goal is to determine if $F_X = F_Y$. We then define the following metric on the space of characteristic functions,

$$\mathcal{D}(X, Y|\alpha) = \int_{\mathbb{R}^d} |\phi_x(t) - \phi_y(t)|^2 \omega(t|\alpha)\, dt,$$

i which ϕ_x and ϕ_y are the characteristic functions associated with distributions F_X and F_Y respectively. Also $\omega(t|\alpha)$ is a positive weight function chosen such that the

integral is finite. By the uniqueness of characteristic functions, it is obvious that $\mathcal{D}(X, Y|\alpha) = 0$ if and only if $F_X = F_Y$.

Another metric that can be considered is based on Euclidean distances. Let (X', Y') be an iid copy of (X, Y), then for $\alpha \in (0, 2)$ define

$$\mathcal{E}(X, Y|\alpha) = 2E|X - Y|^\alpha - E|X - X'|^\alpha - E|Y - Y'|^\alpha. \tag{4.4}$$

For an appropriately chosen weight function,

$$\omega(t|\alpha) = \left(\frac{2\pi^{d/2}\Gamma(1 - \alpha/2)}{\alpha 2^\alpha \Gamma((d + \alpha)/2)}|t|^{d+\alpha}\right)^{-1},$$

we have the following lemma.

Lemma 15. *For any pair of independent random variables X and Y, and $\alpha \in (0, 2)$ is such that $E(|X|^\alpha + |Y|^\alpha) < \infty$, then $\mathcal{E}(X, Y|\alpha) = \mathcal{D}(X, Y|\alpha)$ and $\mathcal{E}(X, Y|\alpha) \in [0, \infty)$. Moreover, $\mathcal{E}(X, Y|\alpha) = 0$ if and only if X and Y are identically distributed.*

Proof. See the appendix of [73]. □

Theorem 15 allows for an intuitively simple empirical divergence measure. Let $\boldsymbol{X}_n$ and $\boldsymbol{Y}_m$ be as above, then we can define the empirical counterpart to Equation 4.4

$$\begin{aligned}\widehat{\mathcal{E}}(\boldsymbol{X}_n, \boldsymbol{Y}_m; \alpha) =& \frac{2}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}|X_i - Y_j|^\alpha \\ &- \binom{n}{2}^{-1}\sum_{1\le i<j\le n}|X_i - X_j|^\alpha - \binom{m}{2}^{-1}\sum_{1\le i<j\le m}|Y_i - Y_j|^\alpha.\end{aligned} \tag{4.5}$$

4.3.2 Incomplete Energy Statistic

The computation of the U-statistics presented in Equation 4.5 require $O(n^2 \vee m^2)$ calculations, which makes it impractical for large n or m. We propose working with

an approximate statistic that is obtained by using incomplete U-statistics. In the following formulation of the incomplete U-statistic let $\delta \in \{2, 3, \ldots, \lfloor \sqrt{T} \rfloor\}$.

Suppose that we divide a segment of our time series into two adjacent subseries, $\boldsymbol{X}_n = \{Z_a, Z_{a+1}, \ldots, Z_{a+n-1}\}$ and $\boldsymbol{Y}_m = \{Z_{a+n}, Z_{a+n+1}, \ldots, Z_{a+n+m-1}\}$, and define the following sets

$$
\begin{aligned}
W_X^\delta &= \{(i,j) : a+n-\delta \le i < j < a+n\} \cup \bigcup_{i=0}^{n-\delta-1} \{(a+i, a+i+1)\} \\
W_Y^\delta &= \{(i,j) : a+n \le i < j < a+n+\delta\} \cup \bigcup_{i=\delta-1}^{m-2} \{(a+n+i, a+n+i+1)\} \\
B^\delta &= (\{a+n-1, \ldots, a+n-\delta\} \times \{a+n, \ldots, a+n+\delta-1\}) \\
&\quad \cup \left(\bigcup_{i=\delta+1}^{m \wedge n} \{(a+n-i, a+n+i-1)\} \right)
\end{aligned}
$$

The set B^δ aims at reducing the number of samples needed to compute the between sample distances. While the sets W_X^δ and W_Y^δ reduce the number of terms used for the within sample distances. When making this reduction the sets W_X^δ, W_Y^δ, and B^δ consider all unique pairs within a δ window around the split that creates $\boldsymbol{X}_n$ and $\boldsymbol{Y}_m$. This point corresponds to a potential change point location and thus we use as much information about points close by to determine the empirical divergence.

We then define the incomplete U-statistic $\widetilde{\mathcal{E}}$ as

$$
\begin{aligned}
\widetilde{\mathcal{E}}(\boldsymbol{X}_n, \boldsymbol{Y}_m | \alpha, \delta) = & \frac{2}{\# B^\delta} \sum_{(i,j) \in B^\delta} |X_i - Y_j|^\alpha \\
& - \frac{1}{\# W_X^\delta} \sum_{(i,j) \in W_X^\delta} |X_i - X_j|^\alpha - \frac{1}{\# W_Y^\delta} \sum_{(i,j) \in W_Y^\delta} |Y_i - Y_j|^\alpha .
\end{aligned} \tag{4.6}
$$

Using this approximation greatly reduces our computational complexity from $O(n^2 \vee m^2)$ to $O(n \vee m)$. [58] shows that a strong law of large numbers result

holds for incomplete U-Statistics, and and thus $\widehat{\mathcal{E}}$ and $\widetilde{\mathcal{E}}$ have the same almost sure limit as $n \wedge m \to \infty$.

4.3.3 The e-cp3o Algorithm

We now present the goodness of fit measure that is used by the e-cp3o change point procedure. In addition, we show that the prescribed measure satisfies the necessary consistency requirements from Section 4.2.1.

The e-cp3o algorithm uses an approximate test statistics combined with an exact search algorithm in order to identify change points. Its goodness of fit measure is given by the following weighted U-Statistic,

$$\widehat{\mathcal{R}}(\boldsymbol{X}_n, \boldsymbol{Y}_m|\alpha) = \frac{mn}{(m+n)^2}\widehat{\mathcal{E}}(\boldsymbol{X}_n, \boldsymbol{Y}_m|\alpha). \tag{4.7}$$

Or an approximation can be obtained by using its incomplete counterpart

$$\widetilde{\mathcal{R}}(\boldsymbol{X}_n, \boldsymbol{Y}_m|\alpha, \delta) = \frac{mn}{(m+n)^2}\widetilde{\mathcal{E}}(\boldsymbol{X}_n, \boldsymbol{Y}_m|\alpha, \delta).$$

By using Slutsky's theorem and a result of [67] we have that if $F_X = F_Y$ then $\widehat{\mathcal{R}}(\boldsymbol{X}_n, \boldsymbol{Y}_m|\alpha) \xrightarrow{p} 0$, and otherwise $\widehat{\mathcal{R}}(\boldsymbol{X}_n, \boldsymbol{Y}_m|\alpha)$ tends almost surely to a finite positive constant, provided that $m = \Theta(n)$ (this means that $n = O(m)$ and $m = O(n)$). In fact if $F_X = F_Y$ we have that $\widehat{\mathcal{R}}(\boldsymbol{X}_n, \boldsymbol{Y}_m|\alpha, \delta) \to 0$ almost surely.

In the case of $\widetilde{\mathcal{R}}(\boldsymbol{X}_n, \boldsymbol{Y}_m|\alpha)$, the result of [60, Theorem 4.1] combined and Slutsky's theorem show that under equal distributions, $F_X = F_Y$, $\widetilde{\mathcal{R}}(\boldsymbol{X}_n, \boldsymbol{Y}_m|\alpha, \delta) \xrightarrow{p} 0$. Similarly, $\widehat{\mathcal{R}}(\boldsymbol{X}_n, \boldsymbol{Y}_m|\alpha, \delta)$ also tends towards a positive finite constant provided $m = \Theta(n)$. These properties lead to a very intuitive goodness of fit measure,

$$\widehat{G}(k|\alpha) = \max_{\tau_1 < \tau_2 < \cdots < \tau_k} \sum_{j=1}^{k} \widehat{\mathcal{R}}(C_j, C_{j+1}|\alpha).$$

By using the dynamic programming approach presented by [48], the values $\widehat{G}(\ell|\alpha)$ for $\ell \leq k$ can be computed in $O(kT^3)$ instead of $O(T^{k+2})$ operations. However, the T^3 term makes this an inadequate approach, so the procedure (e-cp3o) is implemented with the similarly defined goodness of fit measure $\widetilde{G}(k|\alpha, \delta)$, which allows for only $O(kT^2)$ operations.

Consistency of e-cp3o

We now show that the goodness of fit measure, $\widetilde{R}(\cdot, \cdot|\alpha, \delta)$, used by e-cp3o satisfies the conditions for a cp3o based procedure to generate consistent estimates. It is assumed that α has been chosen so that all of the αth moments are finite. In the results below we will consider the goodness of fit measure $\widehat{R}$ based upon the complete U-Statistic, even though the e-cp3o procedure is based on its incomplete version $\widetilde{R}$. The reason for this is that $\widehat{R}$ and $\widetilde{R}$ have the same almost sure limits, and we are working in an asymptotic setting.

Proposition 16. *Assumption 7 is satisfied by the e-cp3o goodness of fit measure.*

Proof. Using the result of [53, Theorem 1] we have that

$$\widehat{R}(A(\tilde{\gamma}), B(\tilde{\gamma})|\alpha) \to \tilde{\gamma}(1-\tilde{\gamma})h(\tilde{\gamma}; \gamma)\mathcal{E}(U_1, U_2|\alpha).$$

Such that $U_1 \sim X(i_1, i_2)$ $U_2 \sim X(i_3, i_4)$, and $h(x; y) = \left(\frac{y}{x}\mathbb{1}_{x \geq y} + \frac{1-y}{1-x}\mathbb{1}_{x<y}\right)^2$. Therefore, $R(X(i_1, i_2), X(i_3, i_4)) = \gamma(1-\gamma)\mathcal{E}(U_1, U_2|\alpha)$ and $\Theta_0^1(\tilde{\gamma}|\gamma) = \frac{\tilde{\gamma}(1-\tilde{\gamma})}{\gamma(1-\gamma)}h(\tilde{\gamma}; \gamma)$, which can be shown to have a unique maximizer at $\tilde{\gamma} = \gamma$. □

Proposition 17. *The portion of Assumption 9 about* $\{G^m(r)\}_{r=1}^m$ *holds for the e-cp3o goodness of fit measure.*

Proof. We begin by showing that $G^m(1) < G^m(2)$. Suppose the first change point partitions the time series into two segments, one where observations are distributed according to F and another where they are distributed according to J. Now suppose that J is a created by a linear mixture of the distributions G and H (which may themselves be mixture distributions). Suppose that the second change point is positioned so as to separate these distributions, G and H. Let random variables X, Y, and Z be such that $X \sim F, Y \sim G$, and $Z \sim H$. Then we have that

$$G^m(1) = \int_{\mathbb{R}^d} |\phi_x(t) - \phi_y(\beta t)\phi_z((1-\beta)t)|^2 \omega(t|\alpha)\, dt$$

where β is the mixture coefficient used to create the distribution J. It is clear that this will be maximized either when $\beta = 0$ or $\beta = 1$, in either case we will show that the obtained value is bounded above by

$$G^m(2) = \int_{\mathbb{R}^d} |\phi_x(t) - \phi_y(t)|^2 \omega(t|\alpha)\, dt + \int_{\mathbb{R}^d} |\phi_y(t) - \phi_z(t)|^2 \omega(t|\alpha)\, dt.$$

Case $\beta = 1$: In this setting the value of $G^m(1)$ is equal to the first term in the definition of $G^m(2)$, and since the distributions G and H are distinct, the second term is strictly positive. Thus $G^m(1) < G^m(2)$.

Case $\beta = 0$: In this case $G^m(1) = \int |\phi_x(t) - \phi_z(t)|^2 \omega(t|\alpha) dt$. However, since we have a metric, the triangle inequality immediately shows that $G^m(1) < G^m(2)$.

In the above setting the location of the first change point was held fixed when the second was identified. This need not be the optimal way to partition the time series into three segments. Thus since this potentially suboptimal segmentation results in an upper bound for $G^m(1)$ it follows that the optimal segmentation will also bound $G^m(1)$.

The argument to show that $G^m(r) < G^m(r+1)$ for $r = 2, \ldots, m-1$ is identical. □

Proposition 18. *The portion of Assumption 9 about* $\left\{\hat{G}^{k(T)}(r)\right\}_{r=1}^{k(T)}$ *holds for the e-cp3o goodness of fit measure.*

Proof. In the paper [73], the empirical measure used for the statistic $\widehat{\mathcal{E}}$ is based upon V-statistics, while we instead use U-statistics. The use of V-statistics ensures that the statistic will always have a nonnegative value. This isn't the case when using U-statistics, but the difference in their value can be bounded by a constant multiple of $\frac{1}{T}$. Combining this with the fact that $0 < G^{k(T)}(r) < G^{k(T)}(r+1)$, and $\frac{d_T}{T} \to 0$, we conclude that for T large enough the version of the statistics based on U-statistics will also produce nonnegative values. Therefore for T large enough $\hat{G}^{k(T)}(r) < \hat{G}^{k(T)}(r+1)$. □

4.4 Simulation Study

We now show the effectiveness of our methodology by considering a number of simulation studies. The goal of these studies is to demonstrate that the e-cp3o procedure is able to perform reasonably well in a variety of settings. In these studies we examine both the number of estimated change points as well as their estimated locations.

To assess the performance of the segmentation obtained from the e-cp3o procedure we use Fowlkes and Mallows' adjusted Rand index [21]. This value is calculated by comparing a segmentation based upon estimated change point locations to the known true segmentation. The index takes into account both the number of change points as well as their locations, and lies in the interval $[0, 1]$, where it is equal to 1 if and only if the two segmentations are identical.

For each simulation study we apply various methods to 100 randomly generated time series. We then report the average running time in seconds, the average adjusted Rand value, and the average number of estimated change points.

As the simulations in the following sections will demonstrate, the e-cp3o procedure does not always generate the best running time or average Rand values. However, in every setting it generates results that are either better or comparable to almost all other competitors, when accuracy and speed are viewed together. For this reason we would advocate the use of the e-cp3o procedure as a general purpose change point algorithm, especially for small to moderate length time series.

To perform the probabilistic pruning introduced in Section 4.2 the value of Γ_ϵ must be specified. In our implementation we obtain an estimate of Γ_ϵ in the following way. We uniformly draw $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ random samples from the set

$$\{(v,t,s,u) : v < t < s < u \text{ and } \min\{t-v, s-t, u-s\} \geq \delta\}.$$

For each sample we calculate

$$\widetilde{R}(Z_{v+1}^{t}, Z_{t+1}^{u}; \alpha) - \widetilde{R}(Z_{v+1}^{t}, Z_{t+1}^{s}; \alpha) - \widetilde{R}(Z_{t+1}^{s}, Z_{s+1}^{u}; \alpha),$$

and then set Γ_ϵ equal to the $1-\epsilon$ quantile of these quantities. Any other sampling approach could be used to obtain a value for Γ_ϵ as long it satisfies the probabilistic criterion.

4.4.1 Univariate Simulations

We begin our simulation study by comparing the e-cp3o procedure to the E-Divisive and PELT procedures. These two procedures are implemented in the `ecp` [41] and

	PELT	E-Div	e-cp3o
	T=400, k(T)=3, K(T)=9		
Rand	$0.884_{7\times10^{-3}}$	$0.987_{2\times10^{-3}}$	$0.937_{10^{-2}}$
# of cps	$9.150_{0.4}$	$3.070_{3\times10^{-2}}$	$2.660_{8\times10^{-2}}$
Time(s)	$0.003_{5\times10^{-5}}$	$9.199_{5\times10^{-2}}$	$0.150_{7\times10^{-4}}$
	T=1650, k(T)=10, K(T)=50		
Rand	$0.953_{3\times10^{-3}}$	$0.992_{7\times10^{-4}}$	$0.940_{5\times10^{-3}}$
# of cps	$16.690_{0.4}$	$10.050_{2\times10^{-2}}$	$9.390_{7\times10^{-2}}$
Time(s)	$0.009_{7\times10^{-5}}$	$239.263_{0.7}$	$3.542_{3\times10^{-2}}$

Table 4.1: e-cp3o growing sample size simulation results

Results of the first univariate simulation from Section 4.4.1 with different time series lengths. The true number of change points is given by k(T) and K(T) the upper limit used by e-cp3o. The table contains average values over 100 replicates, with standard error as subscripts.

`changepoint` [45] R packages respectively. This set of simulations consist of independent Gaussian observations which undergo changes in their mean and variance. The distribution parameters were chosen so that $\mu_j \overset{iid}{\sim} U(-10, 10)$ and $\sigma_j^2 \overset{iid}{\sim} U(0, 5)$. For each analyzed time series all of the different change point procedures were run with their default parameter values. For E-Divisive and e-cp3o this corresponds to $\alpha = 1$. And for e-cp3o the minimum segment size was set to 30 observations (corresponding to $\delta = 29$), and a value of $\epsilon = 0.01$ is used for the probabilistic pruning. Since in this simulation study the number of change points increased with the time series length, the value of $K(T)$ would also change. The results of these simulations are in Table 4.1, which also includes additional information about the time series and upper limit $K(T)$. As can be seen from Table 4.1, better results are obtained by combining an exact test statistic with an approximate search algorithm. But these gains in segmentation quality are rather small. Thus, because of the increase in speed and small loss in segmentation quality, we would argue that the e-cp3o procedure should be preferred over the E-Divisive. The PELT procedure

was much faster, but the e-cp3o procedure was able to generate segmentations that were similar in quality as measured by the adjusted Rand index.

The next set of simulations also compares to a nonparametric procedure from the `npcp` R package. This procedure, like the e-cp3o, is designed to detect changes in the joint distribution of multivariate time series. More information about this procedure, which we will denote by NPCP-F, is given in Section 4.4.2. Time series in this simulation study contain two changes in mean followed by a change in tail index. The changes in mean correspond to the data transitioning from a standard normal distribution to a $N(3, 1)$ and then back to standard normal. The tail index change is caused by a transition to a t-distribution with 2.01 degrees of freedom. We expect that all three methods will be able to easily detect the mean changes and will have a more difficult time detecting the change in tail index. As with the previous set of simulations, all procedures are run with their default parameter values. Results for this set of simulations can be found in Table 4.2. Surprisingly, in this set of simulations the e-cp3o procedure was not only significantly faster than the E-Divisive and NPCP-F, but also managed to generate slightly better segmentations on average.

These two simulation studies on univariate time series show that the e-cp3o procedure performs well when compared to other parametric and nonparametric change point algorithms. The first set of simulations showed that it generated segmentations whose quality is comparable to that of an efficient parametric procedure when its parametric assumptions were satisfied. While the second set of simulations showed that it is able to handle more subtle distributional changes, such as a change in tail behavior. The flexibility of the e-cp3o method allows for it to be used when parametric assumptions are met, as well as in settings where

	NPCP-F	E-Div	e-cp3o
	T=400, k(T)=3, K(T)=9		
Rand	$0.820_{5\times10^{-3}}$	$0.828_{5\times10^{-3}}$	$0.874_{7\times10^{-3}}$
# of cps	$2.280_{6\times10^{-2}}$	$2.200_{5\times10^{-2}}$	$2.430_{5\times10^{-2}}$
Time	$4.790_{2\times10^{-2}}$	$6.726_{5\times10^{-2}}$	$0.176_{10^{-3}}$
	T=1600, k(T)=3, K(T)=9		
Rand	$0.839_{6\times10^{-3}}$	$0.864_{8\times10^{-3}}$	$0.917_{5\times10^{-3}}$
# of cps	$2.370_{6\times10^{-2}}$	$2.480_{7\times10^{-2}}$	$2.920_{3\times10^{-2}}$
Time	$71.772_{0.3}$	$143.207_{1.1}$	$2.392_{4\times10^{-2}}$

Table 4.2: e-cp3o univariate simulation results

Simulation results for time series with mean and tail index changes. The subscripts indicate the standard errors for each value.

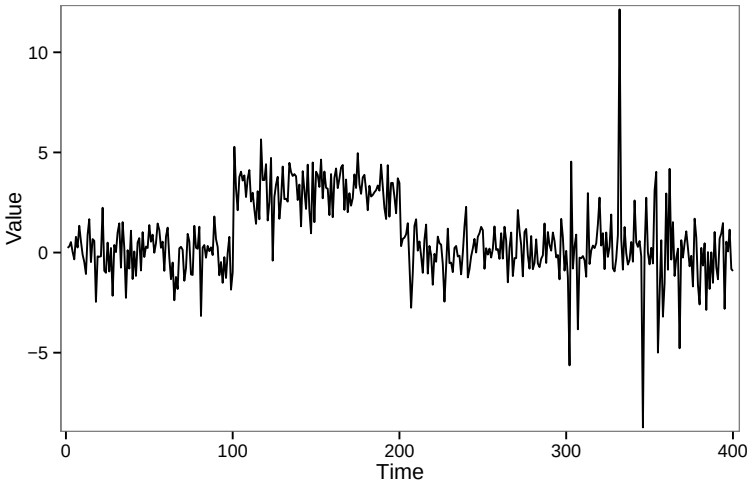

Figure 4.1: Change in mean and tail example

Example of time series with changes in mean and tail index. Mean changes occur at times 100 and 200, while the tail index change is at time 300.

they aren't sure to be satisfied.

4.4.2 Multivariate Simulations

We now examine the performance of the e-cp3o procedure when applied to a multivariate time series. Since a change in mean can be seen as a change in a marginal distribution we could just apply any univariate method to each dimension of the dataset. For this reason we will examine a more complex type of distributional change. In this simulation the distributional change will be due to a change in the copula function [72], while the marginal distributions remain unchanged. Since the PELT procedure as implemented in the `changepoint` package only performs marginal analysis it is not suited for this setting, and will thus not be part of our comparison. We instead consider a method proposed by [27] and implemented in the R package `npcp` by [36]. This package provides two methods that can be used in this setting. One that looks for any change in the joint distribution (NPCP-F) and one designed to detect changes in the copula function (NPCP-C).

For a given set of marginal distributions, the copula function is used to model their dependence. Thus a change in the copula function reflects a change in the dependence structure. This is of particular interest in finance where portfolios of dependent securities are typical [28].

In this simulation we consider a two dimensional process where both marginal distributions are standard normal. While the marginal distributions remain static, the copula function evolves over time. For this simulation the copula undergoes two changes. Initially it is a Clayton copula and then changes to the independence copula and finally becomes a Gumbel copula. The density function for each of the used copulas is provided in Table 4.3 and simulation results in Table 4.4.

As was expected, in Table 4.4 it is clear that the NPCP-C method obtained

Copula	Density $c(u, v)$
Clayton	$\left(\max\{u^{-2.8} + v^{-2.8} - 1,\ 0\}\right)^{-5/14}$
Independence	uv
Gumbel	$\exp\left\{-\left[(-\log(u))^{2.8} + (-\log(v))^{2.8}\right]^{5/14}\right\}$

Table 4.3: Copula densities

The densities for the copula functions used in the multivariate simulations.

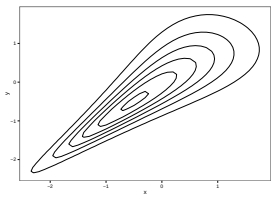

Figure 4.2: Clayton contour plot

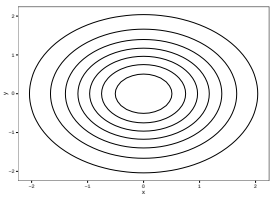

Figure 4.3: Independence contour plot

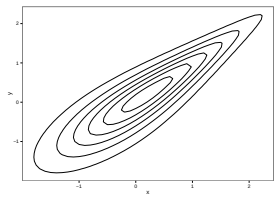

Figure 4.4: Gumbel contour plot

Figure 4.5: Copula contours

	NPCP-C	NPCP-F	e-cp3o
	T=300, k(T)=2, K(T)=9		
Rand	$0.958_{4\times10^{-3}}$	$0.616_{9\times10^{-3}}$	$0.685_{8\times10^{-3}}$
# of cps	$2.130_{4\times10^{-2}}$	$0.320_{6\times10^{-2}}$	$4.000_{0.1}$
Time	$73.163_{0.1}$	$1.871_{6\times10^{-2}}$	$0.125_{4\times10^{-3}}$
	T=1200, k(T)=2, K(T)=9		
Rand	$0.979_{3\times10^{-3}}$	$0.865_{10^{-2}}$	$0.766_{10^{-2}}$
# of cps	$2.150_{4\times10^{-2}}$	$1.790_{9\times10^{-2}}$	$1.570_{7\times10^{-2}}$
Time	$10,580.831_{4.0}$	$41.270_{0.7}$	$1.901_{5\times10^{-3}}$

Table 4.4: e-cp3o multivariate simulation results

Results of the multivariate simulation with different time series lengths. The subscripts indicate the standard errors for each value.

the best average Rand value in all situations. But this comes at a much increased average running time. This becomes very problematic when analysis of a single longer time series can take almost three hours. For shorter time series the e-cp3o provides the best combination between running time, estimated number of change points, and Rand value. For longer time series the NPCP-F procedure is the clear winner.

4.5 Real Data

In this section we apply the e-cp3o procedure to two real data sets. For our first application we make use of a dataset of monthly temperature anomalies. The second consists of monthly foreign exchange (FX) rates between the United States, Russia, Brazil, and Switzerland.

4.5.1 Temperature Anomalies

For the fist application of the e-cp3o procedure we examine the HadCRUT4 dataset of [56]. This dataset consists of monthly global temperature anomalies from 1850 to 2014. Since the dataset consists of anomalies, it does not indicate actual average monthly temperatures, but instead measured deviations from some predefined baseline. The time period used to create the baseline in this case spans 1960 to 1990.

The HadCRUT4 dataset contains two major components; one for land air temperature anomalies and another for sea-surface temperature anomalies. The analysis performed in this section will only consider the land air temperature anomaly

component from the tropical region (30° South to 30° North). This region was chosen because it was the most likely of all the presented regions to have a small difference between the minimum and maximum anomaly value, and be affected by changing seasons. More information about the dataset and the averaging process used can be found in the paper by [56].

From looking at the plot of the tropical land air anomaly time series it is suspected that there is some dependence between observations. This assumption is quickly confirmed by looking at the auto-correlation plot. As a result, we apply the e-cp3o procedure to the differenced data which visually appears to be piecewise stationary. The auto-correlation plot for the differenced data shows that much of the linear dependence has been removed, however, the same plot for the differences squared still indicates some dependence. As with the exchange rate data, we believe that this indicated dependence can be attributed to changes in distribution.

The e-cp3o procedure was applied with a minimum segment length of one year, corresponding to $\delta = 11$; a maximum of $K(T) = 20$ change points were fit, we chose $\alpha = 1$, and $\epsilon = 0.01$. Upon completion we identified change points at the following dates: July 1860, February 1878, January 1918, and February 1973, which are shown in Figure 4.6. With these change points we notice that the auto-correlation plots, for both the differenced and squared differenced data, show almost no statistically significant correlations. This is in line with our original hypothesis that the previously observed correlation was due to the distributional changes within the data.

Furthermore, the February 1973 change point occurs around the same time as the United Nations Conference on the Human Environment. This conference, which was held in June 1972, focused on human interactions with the environ-

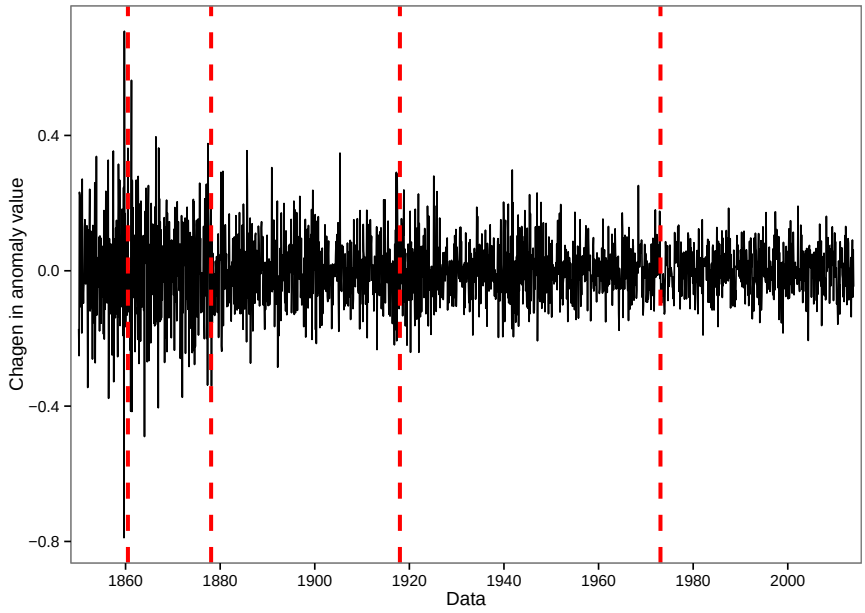

Figure 4.6: Temperature anomalies time series

Change in land air temperature anomalies for the Tropical climate zone from February 1850 to December 2013. The cp3o estimated change point locations are indicated by dashed vertical lines.

ment. From this meeting came a few noteworthy agreed upon principles that have potential to impact land air temperatures:

1. Pollution must not exceed the environment's ability to clean itself
2. Governments would plan their own appropriate pollution policies
3. International organizations should help to improve the environment

These measures, undoubtedly played a role in the decreased average anomaly size, as well as an almost 66% decrease in the variance.

4.5.2 Exchange Rates

We next apply the e-cp3o procedure to a set of spot FX rates obtained through the `R` package `Quandl` [54]. For our analysis we consider the three dimensional time series consisting of monthly FX rates for Brazil (BRL), Russia (RUB), and Switzerland (CHF). All of the rates are against the United States (USD). The time horizon spanned by this time series is September 30, 1996 to February 28, 2014, which results in a total of 210 observations. Looking at the marginal series it is obvious that each of the individual FX rates does not generate a stationary process. Thus, instead of looking at the actual rate, we look at the change in the log process. This transformation results in marginal processes that appear to at least be piecewise stationary.

Our procedure is only guaranteed to work with independent observations, so we must hope that our data either satisfies this condition or is very close to it. The papers by [37, 38] provide evidence that changes in the daily exchange rate are not independent, and that there is a reasonable amount of nonlinear dependence. However, they are not able to conclude whether this observed dependence is due to distributional changes or some other phenomena. For this reason we are instead interested in the change in the monthly exchange rate, which is more likely to either be weakly dependent or show no dependence. To check this we examine the auto/cross-correlation plots for both the difference and difference squared data. This preliminary analysis shows that there is no significant auto or cross-correlation within the differenced data, while for the squared differences there is only significant auto-correlation for Switzerland at a lag of one month.

The e-cp3o procedure is applied with a minimum segment length of six observations (half a year), which corresponds to a value of $\delta = 5$. Furthermore, we have

chosen to fit at most $K(T) = 15$ change points, and values of $\alpha = 1$ and $\epsilon = 0.01$ were used. This specific choice of values resulted in change points being identified at May 31, 1998 and March 31, 2000. These results are depicted in Figure 4.10.

It can be argued that changes in Russia's economic standing leading up to the 1998 ruble crisis are the causes of the May 31, 1998 change point. During the Asian financial crisis many investors were losing faith in the Russian ruble. At one point, the yield on government bonds was as high as 47%. This paired with a 10% inflation rate would normally have been an investor's dream come true. However, people were skeptical of the government's ability to repay these bonds. Furthermore, at this time Russia was using a floating pegged rate for its currency, which resulted in the Central Bank's mass expenditure of USD's which further weakened the ruble's position.

The change point identified at March 31, 2000 also coincides with an economic shift in one of the examined countries. The country most likely to be the cause of this change is Brazil. In 1994 the Brazilian government pegged their currency to the USD. This helped to stabilize the county's inflation rate; however, because of the Asian financial crisis and the ruble crisis many investors were averse to investing in Brazil. In January 1999 the Brazilian Central Bank announced that they would be changing to a free float exchange regime, thus their currency was no longer pegged to the USD. This change devalued the currency and helped to slow the current economic downturn. The change in exchange regime and other factors led to a 48% debt to GDP ratio, besting the IMF target and thus increasing investor faith in Brazil.

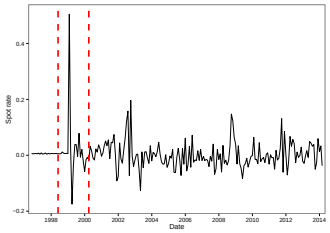

Figure 4.7: Brazil

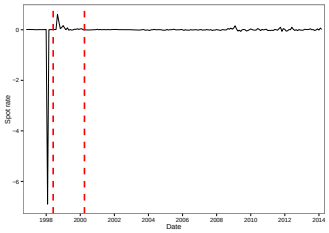

Figure 4.8: Switzerland

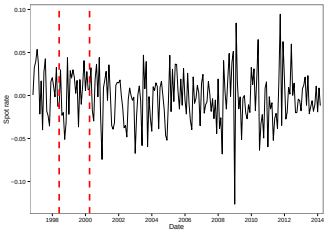

Figure 4.9: Russia

Figure 4.10: Component series for spot rates

Time series for FX spot rates for each of the three countries. Estimated change point locations indicated by vertical lines.

4.6 Conclusion

We have presented an exact search algorithm that incorporates probabilistic pruning in order to reduce the amount of unnecessary calculations. This search method can be used with almost any goodness of fit measure in order to identify change points in multivariate time series. Asymptotic theory has also been provided showing that the cp3o algorithm can generate consistent estimates for both the number of change points as well as the change point locations as the time series increases, provided that a suitable goodness of fit measure is provided. Furthermore, the decoupling of the search procedure and the determination of the number of estimated change points allows for the cp3o algorithm to efficiently generate a collection of optimal segmentations, with differing numbers of change points. This is all accomplished without the user having to specify any sort of penalty constant or function.

By combining the cp3o search algorithm with E-Statistics we developed e-cp3o, a method to perform nonparametric multiple change point analysis that can detect *any* type of distributional change. This method combines an approximate statistic

with an exact search algorithm. The slight loss in accurately estimating change point locations on finite time series is greatly outweighed by the dramatic increase in speed, when compared to similar methods that combine an exact statistic with an approximate search algorithm.

BIBLIOGRAPHY

[1] L. Akoglu and C. Faloutsos. Event detection in time series of mobile communication graphs. In *Proc. of Army Science Conference*, 2010.

[2] Elena Andreou and Eric Ghysels. Detecting multiple breaks in financial market volatility dynamics. *Journal of Applied Econometrics*, 17(5):579–600, 2002.

[3] Sylvain Arlot and Alain Celisse. Segmentation of the mean of heteroscedastic data via cross-validation. *Statistics and Computing*, 21(4):613–632, 2011.

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

[5] Sylvain Arlot and Pascal Massart. Data-driven calibration of penalties for least-squares regression. *The Journal of Machine Learning Research*, 10:245–279, 2009.

[6] Richard Bellman. On the theory of dynamic programming. *Proceedings of the National Academy of Sciences of the United States of America*, 38(8):716, 1952.

[7] Rudolf B Blazek, Hongjoong Kim, Boris Rozovskii, and Alexander Tartakovsky. A novel approach to detection of denial-of-service attacks via adaptive sequential and batch-sequential change-point detection methods. In *Proceedings of IEEE systems, man and cybernetics information assurance workshop*, pages 220–226. Citeseer, 2001.

[8] Kevin Bleakley and Jean-Philippe Vert. The group fused lasso for multiple change-point detection. Technical Report HAL-00602121, Bioinformatics Center (CBIO), June 2011.

[9] Richard Bolton and David Hand. Statistical fraud detection: A review. *Statistical Science*, 17:235 – 255, 2002.

[10] E Brodsky and Boris S Darkhovsky. *Nonparametric Methods in Change Point Problems*. Number 243. Springer, 1993.

[11] Bradley P. Carlin, Alan E. Gelfand, and Adrian F.M. Smith. Hierarchical bayesian analysis of changepoint problems. *Applied Statistics*, 41(2):389 – 405, 1992.

[12] Fang Chang, Weiliang Qiu, Ruben H. Zamar, Ross Lazarus, and Xiaogang Wang. `clues`: An **R** package for nonparametric clustering based on local shrinking. *Journal of Statistical Software*, 33(4):1–16, 2010.

[13] Jie Chen and Arjun K Gupta. *Parametric Statistical Change Point Analysis: With Applications to Genetics, Medicine, and Finance*. Springer, 2011.

[14] H. Cho and P. Fryzlewicz. Multiscale and multilevel technique for consistent segmentation of nonstationary time series. *Statistica Sinica*, 22, 2012.

[15] Haeran Cho and Piotr Fryzlewicz. Multiple change-point detection for high-dimensional time series via sparsified binary segmentation. *Preprint*, 2012.

[16] Tai-leung Chong et al. Partial parameter consistency in a misspecified structural change model. *Economics Letters*, 49(4):351–357, 1995.

[17] R. Davis, T. Lee, and G. Rodriguez-Yam. Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association*, 101(473):223 – 239, 2006.

[18] Alexandra Dias and Paul Embrechts. Change-point analysis for dependence

structures in finance and insurance. In Giorgio Szegö, editor, *Risk Measures for the 21st Century*. Wiley, 2004.

[19] Chandra Erdman and John W. Emerson. bcp: An R package for performing a bayesian analysis of change point problems. *Journal of Statistical Software*, 23(3):1–13, 2007.

[20] Tom Fawcett and Foster Provost. Adaptive fraud detection. *Data Mining and Knowledge Discovery*, 1(3):291–316, 1997.

[21] E. B. Fowlkes and C. L. Mallows. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, 78(383):553 – 569, 1983.

[22] Peter Friedman. A change point detection method for elimination of industrial interference in radio astronomy receivers. In *Statistical Signal and Array Processing, 1996. Proceedings., 8th IEEE Signal Processing Workshop on (Cat. No. 96TB10004*, pages 264–266. IEEE, 1996.

[23] Piotr Fryzlewicz et al. Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42(6):2243–2281, 2014.

[24] Axel Gandy. Sequential implementation of monte carlo tests with uniformly bounded resampling risk. *Journal of the American Statistical Association*, 104(488):1504–1511, 2009.

[25] Axel Gandy. Sequential implementation of monte carlo tests with uniformly bounded resampling risk. *Journal of the American Statistical Association*, 104(488):1504–1511, 2009.

[26] Alan Genz, Frank Bretz, Tetsuhisa Miwa, Xuefei Mi, Friedrich Leisch, Fabian

Scheipl, and Torsten Hothorn. `mvtnorm`*: Multivariate Normal and T Distributions*, 2012.

[27] Edit Gombay and Lajos Horvath. Change-points and bootstrap. *Environmetrics*, 10(6):725–736, 1999.

[28] Dominique Guegan and Jing Zhang. Change analysis of a dynamic copula for measuring dependence in multivariate financial data. *Quantitative Finance*, 10(4):421–430, 2010.

[29] Valery Guralnik and Jaideep Srivastava. Event detection from time series data. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '99. ACM, 1999.

[30] Alexis Hannart and Philippe Naveau. An improved bayesian information criterion for multiple change-point models. *Technometrics*, 54(3):256–268, 2012.

[31] Zaid Harchaoui and Oliver Cappe. Retrospective multiple change-point estimation with kernels. In *Statistical Signal Processing, 2007. SSP '07. IEEE/SP 14th Workshop on*, pages 768 –772, 2007.

[32] Samir B. Hariz, Jonathan J. Wylie, and Qiang Zhang. Optimal rate of convergence for nonparametric change-point estimators for nonstationary sequences. *The Annals of Statistics*, 35(4):1802 – 1826, 2007.

[33] Trevor Hastie, Robert Tibshirani, Jerome Friedman, T Hastie, J Friedman, and R Tibshirani. *The Elements of Statistical Learning*, volume 2. Springer, 2009.

[34] Douglas M. Hawkins. Fitting multiple change-point models to data. *Computational Statistics and Data Analysis*, 37(3):323 – 341, 2001.

[35] Wassily Hoeffding. The strong law of large numbers for U-statistics. Technical Report 302, North Carolina State University. Dept. of Statistics, 1961.

[36] Mark Holmes, Ivan Kojadinovic, and Jean-François Quessy. Nonparametric tests for change-point detection à la gombay and horváth. *Journal of Multivariate Analysis*, 115:16–32, 2013.

[37] David A Hsieh. The statistical properties of daily foreign exchange rates: 1974–1983. *Journal of International Economics*, 24(1):129–145, 1988.

[38] David A Hsieh. Testing for nonlinear dependence in daily foreign exchange rates. *Journal of Business*, 62(3):339–368, 1989.

[39] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193 – 218, 1985.

[40] Brad Jackson, Jeffrey D Scargle, David Barnes, Sundararajan Arabhi, Alina Alt, Peter Gioumousis, Elyus Gwin, Paungkaew Sangtrakulcharoen, Linda Tan, and Tun Tao Tsai. An algorithm for optimal partitioning of data on an interval. *Signal Processing Letters, IEEE*, 12(2):105–108, 2005.

[41] Nicholas A. James and David S. Matteson. ecp: An R package for nonparametric multiple change point analysis of multivariate data. *Journal of Statistical Software*, 62(7):1–25, 2014.

[42] O. Johnson, D. Sejdinovic, J. Cruise, A. Ganesh, and R. Piechocki. Non-parametric change-point detection using string matching algorithms. *arXiv:1106.5714*, June 2011.

[43] Y. Kawahara and M. Sugiyama. Sequential change-point detection based on direct density-ratio estimation. *Statistical Analysis and Data Mining*, 5(2):114–127, 2011.

[44] R. Killick, P. Fearnhead, and IA Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

[45] Rebecca Killick and Idris A. Eckley. changepoint: An R package for changepoint analysis. *Journal of Statistical Software*, 58(3):1–19, 2014.

[46] A.Y. Kim, C. Marzban, D.B. Percival, and W. Stuetzie. Using labeled data to evaluate change detctors in a multivariate streaming environment. *Signal Processing*, 89(12):2529 – 2536, 2009.

[47] Marc Lavielle. Using penalized contrasts for the change-point problem. *Signal processing*, 85(8):1501–1510, 2005.

[48] Marc Lavielle and Gilles Teyssière. Detection of multiple change-points in multivariate time series. *Lithuanian Mathematical Journal*, 46(3):287 – 306, 2006.

[49] A. Lung-Yut-Fong, C. Lévy-Leduc, and O. Cappé. Homogeneity and change-point detection tests for multivariate data using rank statistics. *arXiv:1107.1971*, 2011.

[50] Edgard M Maboudou-Tchao and Douglas M Hawkins. Detection of multiple change-points in multivariate data. *Journal of Applied Statistics*, 40(9):1979–1995, 2013.

[51] Michael Mampaey and Jilles Vreeken. Summarizing categorical data by clustering attributes. *Data Mining and Knowledge Discovery*, 24:1 – 44, 2011.

[52] D. S. Matteson, M. W. McLean, D. B. Woodard, and S. G. Henderson. Forecasting Emergency Medical Service Call Arrival Rates. *The Annals of Applied Statistics*, 5(2B):1379–1406, 2011.

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

[54] Raymond McTaggart and Gergely Daroczi. *Quandl: Quandl Data Connection*, 2013. R package version 2.1.2.

[55] Leslie C. Morey and Alan Agresti. The measurement of classification agreement: An adjustment to the rand statistic for chance agreement. *Educational and Psychological Measurement*, 44:33 – 37, 1984.

[56] Colin P Morice, John J Kennedy, Nick A Rayner, and Phil D Jones. Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The hadcrut4 data set. *Journal of Geophysical Research: Atmospheres (1984–2012)*, 117(D8), 2012.

[57] Vito M.R. Muggeo and Giada Adelfio. Efficient change point detection for genomic sequences of continuous measurements. *Bioinformatics*, 27:161 – 166, 2011.

[58] Masoud M. Nasari. Strong law of large numbers for weighted u-statistics: Application to incomplete u-statistics. *Statistics & Probability Letters*, 82(6):1208 – 1217, 2012.

[59] Adam B. Olshen and E.S. Venkatraman. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5:557 – 572, 2004.

[60] Kevin A O'Neil and Richard A Redner. Asymptotic distributions of weighted u-statistics of degree 2. *The Annals of Probability*, pages 1159–1169, 1993.

[61] E.S. Page. Continuous inspection schemes. *Biometrika*, 41:100 – 115, 1954.

[62] Jean-Yves Pitarakis. Least squares estimation and tests of breaks in mean and variance under misspecification. *Econometrics Journal*, 7(1):32–54, 2004.

[63] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012.

[64] William M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66:846 – 850, 1971.

[65] Guillem Rigaill. Pruned dynamic programming for optimal multiple change-point detection. *arXiv:1004.0887*, 2010.

[66] Guillem Rigaill. Pruned dynamic programming for optimal multiple change-point detection. *arXiv preprint arXiv:1004.0887*, 2010.

[67] Maria L. Rizzo and Gábor J. Székely. Disco analysis: A nonparametric extension of analysis of variance. *The Annals of Applied Statistics*, 4(2):1034–1055, 2010.

[68] Gordon J. Ross. `cpm`*: Sequential Parametric and Nonparametric Change Detection*, 2012. **R** package version 1.0.

[69] Karlton Sequeira and Mohammed Zaki. Admit: Anomaly-based data mining for intrusions. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02. ACM, 2002.

[70] Xiaotong Shen and Jianming Ye. Adaptive model selection. *Journal of the American Statistical Association*, 97(457):210–221, 2002.

[71] Vasilios A Siris and Fotini Papagalou. Application of anomaly detection algorithms for detecting syn flooding attacks. *Computer communications*, 29(9):1433–1442, 2006.

[72] M Sklar. *Fonctions de Répartition à n Dimensions et Leurs Marges*. Université Paris 8, 1959.

[73] Gábor J. Székely and Maria L. Rizzo. Hierarchical clustering via joint between-within distances: Extending ward's minimum variance method. *Journal of Classification*, 22(2):151 – 183, 2005.

[74] Makram Talih and Nicolas Hengartner. Structural learning with time-varying components: Tracking the cross-section of financial time series. *Journal of the Royal Statistical Society*, 67:321 – 341, 2005.

[75] E.S. Venkatraman. *Consistency Results in Multiple Change-Point Problems*. PhD thesis, Stanford University, 1992.

[76] L.J. Vostrikova. Detection disorder in multidimensional random processes. *Soviet Math Dokl.*, 24:55 – 59, 1981.

[77] Hui Wang, Rebecca Killick, and Xiang Fu. Distributional change of monthly precipitation due to climate change: Comprehensive examination of dataset in southeastern united states. *Hydrological Processes*, 2013.

[78] Yao Xie, Jiaji Huang, and Rebecca Willett. Change-point detection for high-dimensional time series with missing data. *Selected Topics in Signal Processing, IEEE Journal of*, 7(1):12–27, 2013.

[79] Yi Ching Yao. Estimating the number of change-points via schwarz criterion. *Statistics & Probability Letters*, 6:181 – 189, 1987.

[80] Achim Zeileis, Christian Kleiber, Walter Krämer, and Kurt Hornik. Testing and dating of structural changes in practice. *Computational Statistics & Data Analysis*, 44:109–123, 2003.

[81] Achim Zeileis, Friedrich Leisch, Kurt Hornik, and Christian Kleiber. strucchange: An r package for testing for structural change in linear regression models. *Journal of Statistical Software*, 7(2):1–38, 2002.

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

[83] Changliang Zou, Guosheng Yin, Long Feng, Zhaojun Wang, et al. Nonparametric maximum likelihood approach to multiple change-point problems. *The Annals of Statistics*, 42(3):970–1002, 2014.