A Nonparametric Approach for Multiple Change Point Analysis of Multivariate Data

David S. Matteson and Nicholas A. James
Cornell University*

April 30, 2013

Abstract

Change point analysis has applications in a wide variety of fields. The general problem concerns the inference of a change in distribution for a set of time-ordered observations. Sequential detection is an online version in which new data is continually arriving and is analyzed adaptively. We are concerned with the related, but distinct, offline version, in which retrospective analysis of an entire sequence is performed. For a set of multivariate observations of arbitrary dimension, we consider nonparametric estimation of both the number of change points and the positions at which they occur. We do not make any assumptions regarding the nature of the change in distribution or any distribution assumptions beyond the existence of the αth absolute moment, for some $\alpha \in (0, 2)$. Estimation is based on hierarchical clustering and we propose both divisive and agglomerative algorithms. The divisive method is shown to provide consistent estimates of both the number and location of change points under standard regularity assumptions. We compare the proposed approach with competing methods in a simulation study. Methods from cluster analysis are applied to assess performance and to allow simple comparisons of location estimates, even when the estimated number differs. We conclude with applications in genetics, finance and spatio-temporal analysis.

KEY WORDS: Cluster analysis; Multivariate time series; Permutation tests; Signal processing; U-statistics.

Short title: Nonparametric Change Point Analysis

*Matteson is an Assistant Professor, Department of Statistical Science, Cornell University, 1196 Comstock Hall, Ithaca, NY 14853 (Email: matteson@cornell.edu; Web: `http://www.stat.cornell.edu/~matteson/`). James is a PhD Candidate, School of Operations Research and Information Engineering, Cornell University, 206 Rhodes Hall, Ithaca, NY 14853 (Email: nj89@cornell.edu; Web: `https://courses.cit.cornell.edu/nj89/`).

1 Introduction

Change point analysis is the process of detecting distributional changes within time-ordered observations. This arises in financial modeling (Talih and Hengartner, 2005), where correlated assets are traded and models are based on historical data represented as multivariate time series. It is applied in bioinformatics (Muggeo and Adelfio, 2011) to identify genes that are associated with specific cancers and other diseases. Change point analysis is also used to detect credit card fraud (Bolton and Hand, 2002) and other anomalies (Sequeira and Zaki, 2002; Akoglu and Faloutsos, 2010); and for data classification in data mining (Mampaey and Vreeken, 2011). Applications can also be found in signal processing, where change point analysis can be used to detect significant changes within a stream of images (Kim et al., 2009).

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 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 Carlin et al. (1992) and Lavielle and Teyssière (2006) analysis is performed by maximizing a log-likelihood function, while Page (1954) examines the ratio of log-likelihood functions to estimate change points. Additionally, Davis et al. (2006) 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 (Hariz et al., 2007). Nonparametric approaches often rely heavily on the estimation of density functions (Kawahara and Sugiyama, 2011), though they have also been performed using rank statistics (Lung-Yut-Fong et al., 2011). 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, Johnson et al. (2011) 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 Yao (1987) and Zhang and Siegmund (2007). The proposed approach is motivated by methods from cluster analysis (Székely and Rizzo, 2005).

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, Hawkins (2001) estimates change point locations by maximizing a likelihood function. Lavielle and Teyssière (2006) accomplish the same task by minimizing a loss function. Sequential methods generally estimate change points one at a time (Guralnik and Srivastava, 1999), although some have the ability to estimate two or more at any given stage (Olshen and Venkatraman, 2004). 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 we describe our methodology; its properties are discussed in Section 3. In Sections 4 and 5 we present the results of our procedure when applied to simulated and real data, respectively. In Section 6 we propose an alternative algorithm and illustrate its use on a novel spatio-temporal application. Concluding remarks are in Section 7 and technical details are stated in the Appendix.

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 Rizzo and Székely (2010) 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 $\mathcal{O}(T^k)$ 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 Harchaoui and Cappe (2007), Rigaill (2010) and Lung-Yut-Fong et al. (2011).

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 (Vostrikova, 1981; Cho and Fryzlewicz, 2012) and model selection procedures (Lavielle and Teyssière, 2006; Arlot et al., 2012) 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 (Vostrikova, 1981) with a multivariate divergence measure from Székely and Rizzo (2005). We first discuss measuring differences in multivariate distributions. We then pro-

pose 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.

2.1 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

$$\int_{\mathbb{R}^d} |\phi_x(t) - \phi_y(t)|^2 \, w(t) \, dt, \tag{1}$$

in which $w(t)$ denotes an arbitrary positive weight function, for which the above integral exists. In consideration of Lemma 9 (see Appendix), we use the following weight function

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

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

$$\mathcal{D}(X,Y;\alpha) = \int_{\mathbb{R}^d} |\phi_x(t) - \phi_y(t)|^2 \left(\frac{2\pi^{d/2}\Gamma(1-\alpha/2)}{\alpha 2^{\alpha}\Gamma((d+\alpha)/2)} |t|^{d+\alpha} \right)^{-1} dt. \tag{3}$$

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 Székely and Rizzo (2005) as

$$\mathcal{E}(X,Y;\alpha) = 2E|X-Y|^{\alpha} - E|X-X'|^{\alpha} - E|Y-Y'|^{\alpha}. \tag{4}$$

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 Székely and Rizzo (2005).

The equivalence established in Lemma 1 motivates a remarkably simple empirical divergence measure for multivariate distributions based on U-statistics. Let $\boldsymbol{X}_n = \{X_i : i = 1, \ldots, n\}$ and $\boldsymbol{Y}_m = \{Y_j : j = 1, \ldots, m\}$ 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 (4) may be defined as

$$\widehat{\mathcal{E}}(\boldsymbol{X}_n, \boldsymbol{Y}_m; \alpha) = \frac{2}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} |X_i - Y_j|^\alpha - \binom{n}{2}^{-1} \sum_{1 \le i < k \le n} |X_i - X_k|^\alpha - \binom{m}{2}^{-1} \sum_{1 \le j < k \le m} |Y_j - Y_k|^\alpha. \quad (5)$$

This measure is based on Euclidean distances between sample elements and is $\mathcal{O}(m^2 \vee n^2)$, whereas the sample counterpart of Equation (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 (Hoeffding, 1961) 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 2.4.

2.2 Estimating the Location of a Change Point

Let

$$\widehat{\mathcal{Q}}(\boldsymbol{X}_n, \boldsymbol{Y}_m; \alpha) = \frac{mn}{m+n}\widehat{\mathcal{E}}(\boldsymbol{X}_n, \boldsymbol{Y}_m; \alpha) \quad (6)$$

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 $1 \le \tau < \kappa \le T$ 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

$$(\hat{\tau}, \hat{\kappa}) = \underset{(\tau,\kappa)}{\operatorname{argmax}} \ \widehat{\mathcal{Q}}(\boldsymbol{X}_\tau, \boldsymbol{Y}_\tau(\kappa); \alpha). \quad (7)$$

It is possible to calculate the argmax in Equation (7) in $\mathcal{O}(T^2)$ 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 Venkatraman (1992), 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 Olshen and Venkatraman (2004).

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 (7). Now, let

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

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

$$\hat{q}_k = \hat{\mathcal{Q}}(\boldsymbol{X}_{\hat{\tau}_k}, \boldsymbol{Y}_{\hat{\tau}_k}(\hat{\kappa}_k); \alpha), \tag{8}$$

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 $\mathcal{O}(kT^2)$, in which k is the unknown number of change points.

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 and 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 Gandy (2009).

The permutation test may be performed at each stage in the iterative estimation 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.

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.

3.1 Single Change Point

In Section 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 2. *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 3. *Suppose Assumption 2 holds, then*

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

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 4. *Suppose Assumption 2 holds. Let $\hat{\tau}_T$ denote the estimated change point location for a sample of size T, as defined in Equation (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

$$\widehat{\mathcal{E}}(\boldsymbol{X}_T(\tilde{\gamma}), \boldsymbol{Y}_T(\tilde{\gamma}); \alpha) \overset{a.s.}{\to} \left(\frac{\gamma}{\tilde{\gamma}}\mathbb{1}_{\tilde{\gamma}\geq\gamma} + \frac{1-\gamma}{1-\tilde{\gamma}}\mathbb{1}_{\tilde{\gamma}<\gamma}\right)^2 \mathcal{E}(X, Y; \alpha) = h(\tilde{\gamma}; \gamma)\mathcal{E}(X, Y; \alpha) \tag{9}$$

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

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

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

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

we have

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

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

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

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,

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

□

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.

3.2 Multiple Change Points

The consistency result presented in Vostrikova (1981) 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 5. *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 5, 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 (3) between these two mixture distributions is equal to

$$\int_{\mathbb{R}^d} \left| \frac{\gamma^{(1)}}{\tilde{\gamma}} \phi_x(t) + \left(\frac{\tilde{\gamma} - \gamma^{(1)}}{\tilde{\gamma}} \right) \phi_u(t) - \left(\frac{1 - \gamma^{(2)}}{1 - \tilde{\gamma}} \right) \phi_y(t) - \left(\frac{\gamma^{(2)} - \tilde{\gamma}}{1 - \tilde{\gamma}} \right) \phi_u(t) \right|^2 w(t; \alpha)\, dt \tag{10}$$

Lemma 6. *Suppose that Assumption 5 holds for some $\alpha \in (0, 2)$, then the divergence measure in Equation (10) is maximized when either $\tilde{\gamma} = \gamma^{(1)}$ or $\tilde{\gamma} = \gamma^{(2)}$.*

Proof. Equation (10) can be rewritten as

$$f(\tilde{\gamma}) = \int_{\mathbb{R}^d} \left| \frac{\gamma^{(1)}}{\tilde{\gamma}} [\phi_x(t) - \phi_u(t)] + \frac{1 - \gamma^{(2)}}{1 - \tilde{\gamma}} [\phi_u(t) - \phi_y(t)] \right|^2 w(t; \alpha)\, dt. \tag{11}$$

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

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

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 (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 3 the within distances for $X_T(\tilde{\gamma})$ and $Y_T(\tilde{\gamma})$ converge uniformly to

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

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

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 5, for each $i = 0, 1, \ldots, k$, there exist distributions F_i, G_i, and H_i such that for $\gamma^{(i)} \leq \tilde{\gamma} \leq \gamma^{(i+1)}$, Equation (11) holds; otherwise $f_i(\tilde{\gamma}) = 0$. By Lemmas 6 and 7, $f_i(\tilde{\gamma})$ is maximized when $\tilde{\gamma} = \gamma^{(i)}$ or $\tilde{\gamma} = \gamma^{(i+1)}$ for $i = 1, 2, \ldots, k-1$. By Theorem 4, $f_0(\tilde{\gamma})$ and $f_k(\tilde{\gamma})$ are maximized at $\gamma^{(1)}$ and $\gamma^{(k)}$, respectively.

Theorem 8. *Suppose that Assumption 5 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 (7), and $\mathscr{A}_T = \{y \in [\delta_T, 1 - \delta_T] : f(y) \geq f(\gamma),\ \forall \gamma\}$. 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 7. 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 $\tilde{\gamma}_1, \tilde{\gamma}_2 \in [\gamma^{(i)}, \gamma^{(i+1)}]$,

$$|\tilde{\gamma}_1(1 - \tilde{\gamma}_1)f_i(\tilde{\gamma}_1) - \tilde{\gamma}_2(1 - \tilde{\gamma}_2)f_i(\tilde{\gamma}_2)| > c_i|\tilde{\gamma}_1 - \tilde{\gamma}_2| + o(|\tilde{\gamma}_1 - \tilde{\gamma}_2|). \tag{12}$$

Let $\epsilon > 0$. By Equation (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 $\gamma^* = \text{argmin}_{x \in \mathscr{A}_T} |\hat{\gamma}_T - x|$, 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, $P\Big(\lim_{T\to\infty} d(\hat{\tau}_T/T, \mathscr{A}_T) \leq \eta(\epsilon)\Big) = 1$, and since ϵ was arbitrary, we have established the claim. □

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 (see, Lung-Yut-Fong et al., 2011), which is based on a generalization of a Wilcoxon/Mann-Whitney (marginal) rank based approach, the parametric Pruned Exact Linear Time (PELT) procedure (Killick et al., 2012), and the nonparametric Kernel Change Point (KCP) procedure (Arlot et al., 2012). Each simulation applies these methods to a set of 1,000 independent sequences with two change points, and computes the average Rand index (Fowlkes and Mallows, 1983; Hubert and Arabie, 1985), defined below, and approximate standard errors. All computation was completed using the statistical software `R` (R Development Core Team, 2012), using the `ecp` package (see James and Matteson, 2013).

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.

4.1 Comparing Sets of Change Point Estimates

To measure the performance of a particular method we calculate the Rand index (Rand, 1971) as well as Morey and Agresti's Adjusted Rand index (Morey and Agresti, 1984). 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 Hubert and Arabie (1985), 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 Hubert and Arabie (1985) and Fowlkes and Mallows (1983), 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.

4.2 Univariate Analysis

In this section we compare the simulation performance of the E-Divisive, MultiRank, 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 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 8.

	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 1: 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.

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. 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.

	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: 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.

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 are shown in Table 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 3: 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.

5 Applications

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

5.1 Genetics Data

We first consider the genome data from Bleakley and Vert (2011). 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 Bleakley and Vert (2011) 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 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.

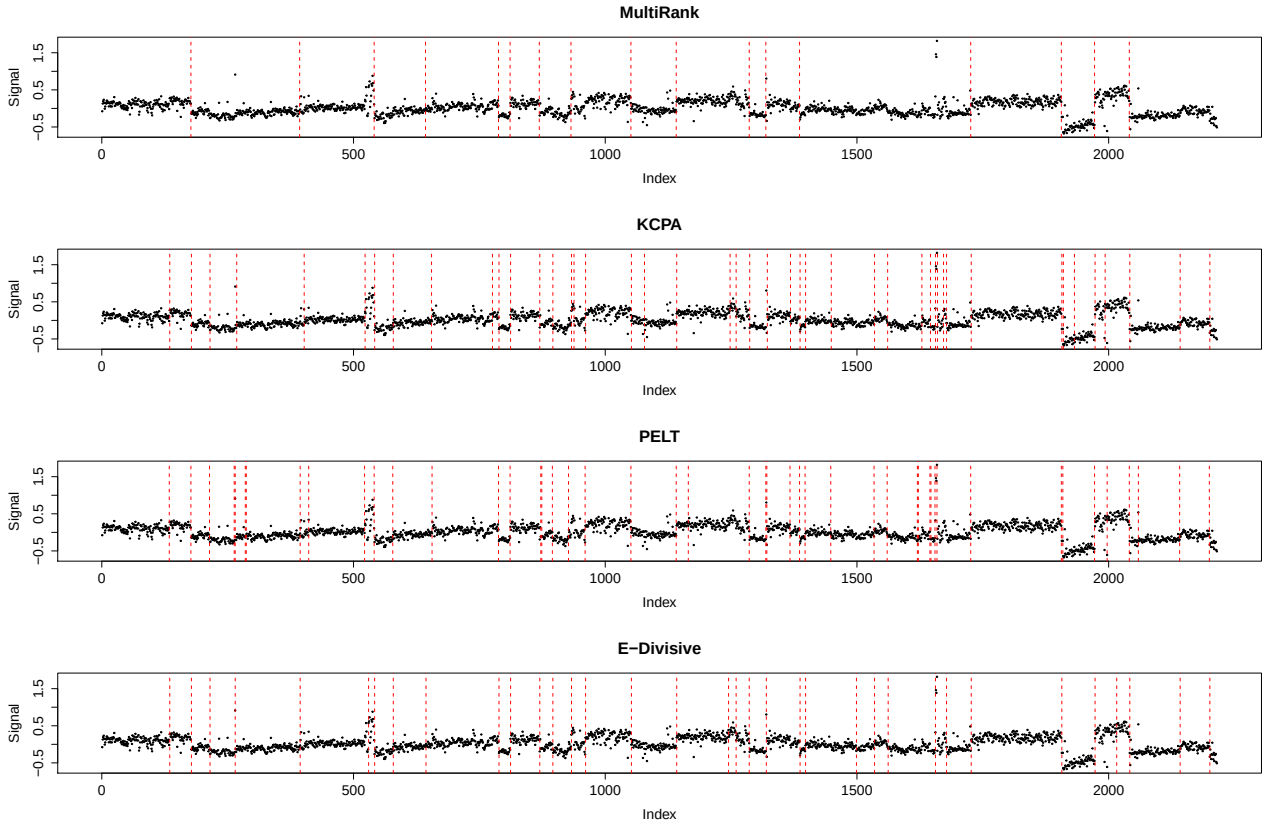

Figure 1: The normalized relative aCGH signal for the tenth individual with a bladder tumor; the estimated change point locations for the MultiRank, KCPA, PELT and E-Divisive methods are indicated by the dashed vertical lines.

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 $\alpha = 1$, 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 with vertical lines to denote the estimated change point locations at April 2000 and October 2002.

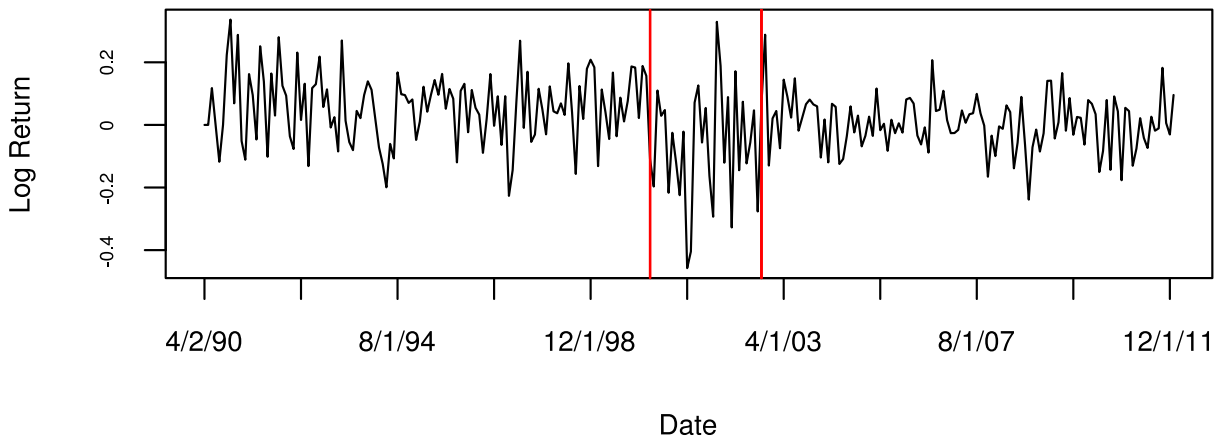

Figure 2: Monthly log returns for Cisco Systems Inc. stock, from April 1990 through January 2012; the E-Divisive procedure estimates significant changes in distribution at the vertical lines April 2000 and October 2002.

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, which became China's largest urban communications network.

Figure 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.

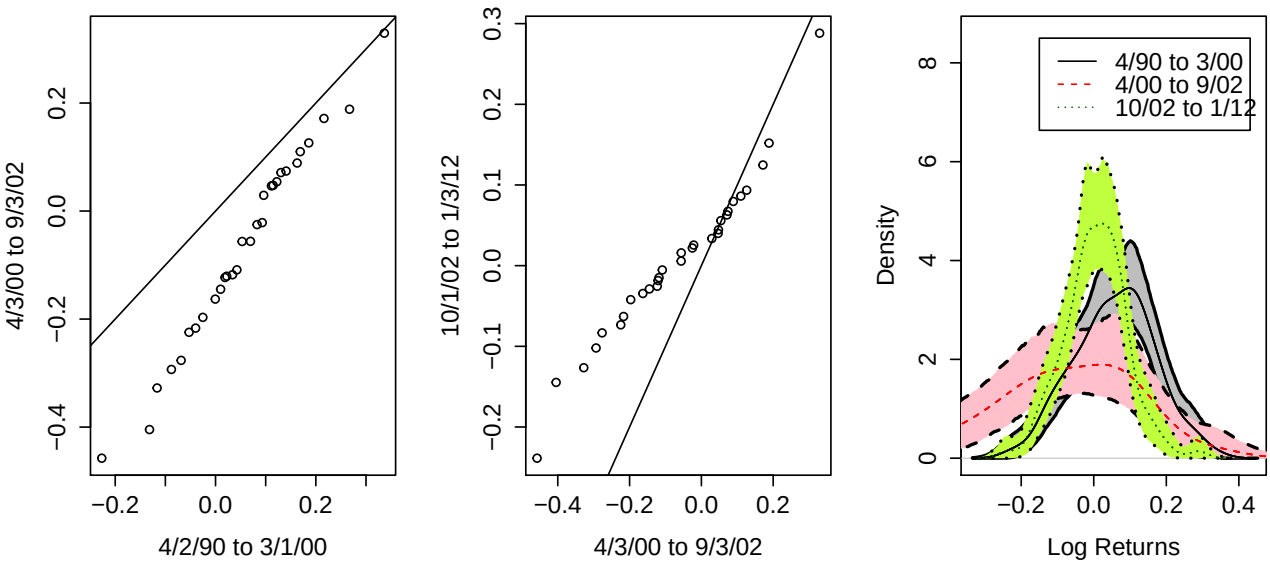

Figure 3: Distributional comparisons between the estimated change points from the E-Divisive procedure: (a,b) quantile-quantile plots between adjacent time periods; and (c) kernel density estimates for each period with 95% confidence bands.

To graphically support the assumption of independent observations within clusters, Figure 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 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.

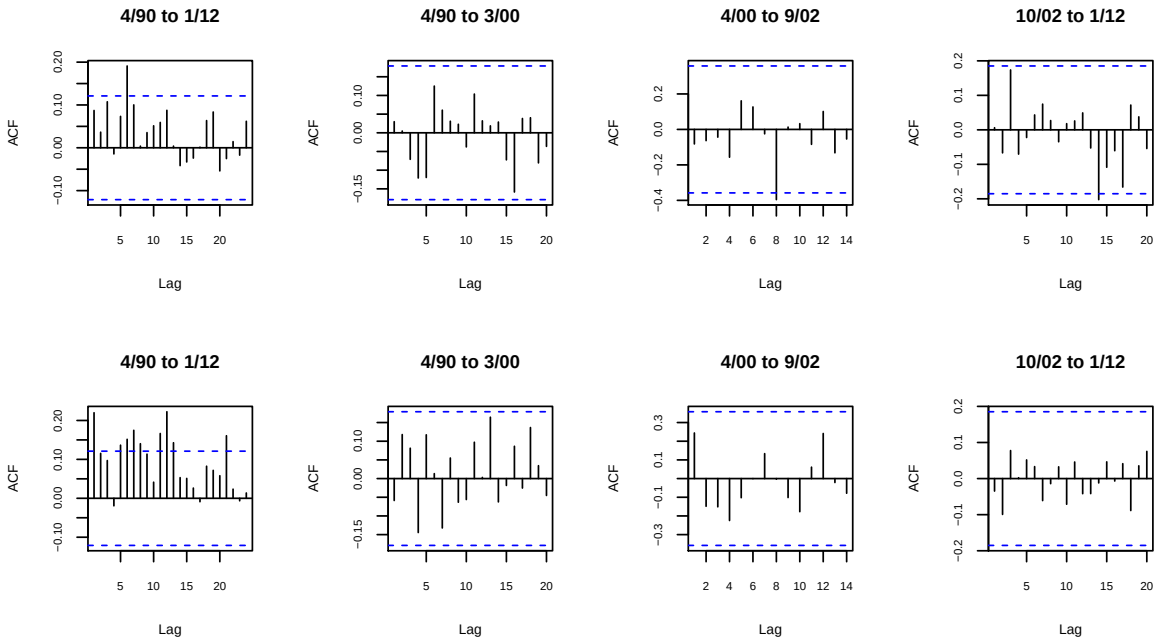

Figure 4: Sample auto-correlation function for the returns (top row) and the squared returns (bottom row), for the entire period (first column) and each estimated sub-period (later columns). The dashed horizontal lines represent approximate 95% confidence intervals about zero.

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.

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.

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

$$\widehat{\mathcal{S}}_n(\mathcal{C}; \alpha) = \sum_{i=1}^{n-1} \widehat{\mathcal{Q}}(C_i, C_{i+1}; \alpha), \tag{13}$$

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 (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 $\mathcal{O}(T^2)$.

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 Matteson et al. (2011); 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 5. The goodness-of-fit measure in Equation (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 5. We note a persistently large density in the downtown region and various shape changes in the outlying regions.

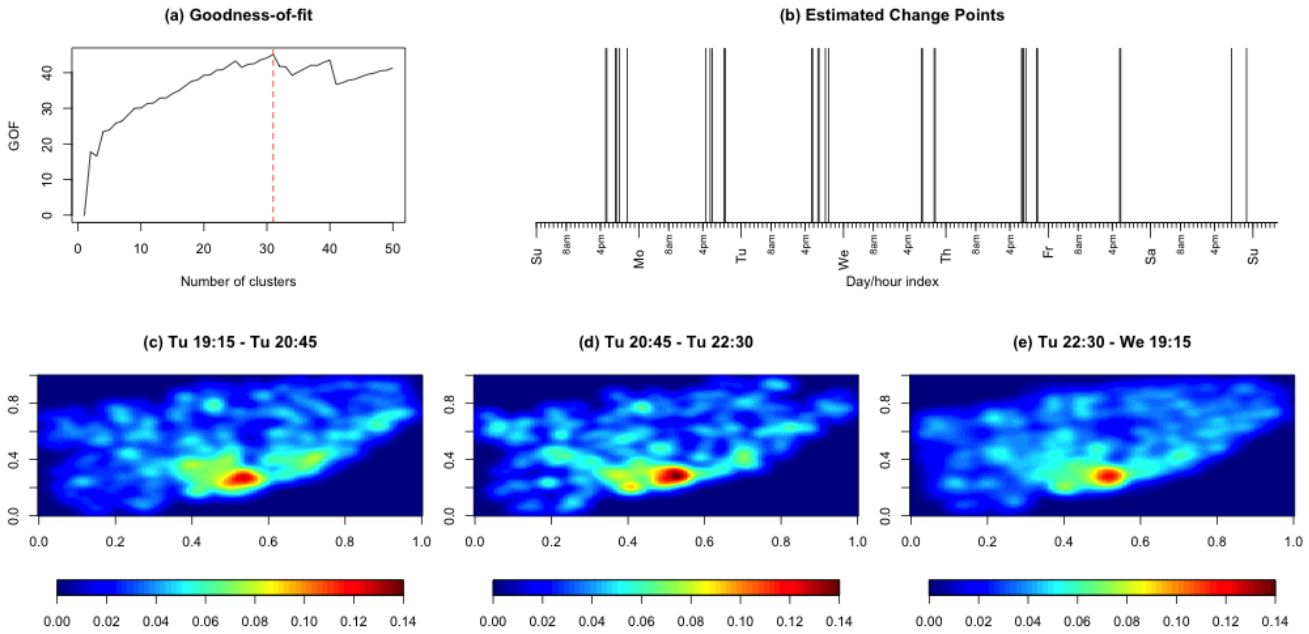

Figure 5: Results from application of the proposed agglomerative algorithm on the Toronto EMS ambulance data: (a) the goodness-of-fit measure of Equation (13); (b) the 31 estimated change point locations; and spatial density estimates, on the square-root scale, from observation in three adjacent cluster periods (c) Tuesday 19:15 - 20:45, (d) Tuesday 20:45 - 22:30, and (e) Tuesday 22:30 - Wednesday 19:15.

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 Hawkins (2001), Lavielle and Teyssière (2006), or Lung-Yut-Fong et al. (2011). Furthermore, this advantage does not come at the expense of additional computational complexity; similar to the previously mentioned methods, the proposed approach is $\mathcal{O}(kT^2)$.

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.

Acknowledgments

We would like to thank the three referees and Associate Editor for their careful review of the manuscript and helpful comments. The authors sincerely thank Toronto EMS for sharing their data. The authors are also grateful to Louis C. Segalini for his research assistance in preparing the financial data analysis. This work was partially supported by National Science Foundation Grant Number CMMI-0926814.

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

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

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

Proof. See page 177 in Székely and Rizzo (2005). □

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 (1) is equal to zero. Take $w(t)$ equal to $w(t;\alpha)$, as defined in Equation (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

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

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 9 it follows that

$$\mathcal{D}(X,Y;\alpha) = \int |\phi_x(t) - \phi_y(t)|^2 \, w(t;\alpha)\, dt$$

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

Finally, $\mathcal{E}(X, Y; \alpha) \geq 0$ since the integrand in Equation (3) is non-negative. □

References

Akoglu, L., and Faloutsos, C. (2010), Event Detection in Time Series of Mobile Communication Graphs,, in *Proc. of Army Science Conference*.

Arlot, S., Celisse, A., and Harchaoui, Z. (2012), "Kernel change-point detection," *arXiv preprint arXiv:1202.3878*, .

Bleakley, K., and Vert, J.-P. (2011), The Group Fused Lasso for Multiple Change-Point Detection,, Technical Report HAL-00602121, Bioinformatics Center (CBIO).

Bolton, R., and Hand, D. (2002), "Statistical Fraud Detection: A Review," *Statistical Science*, 17, 235 – 255.

Carlin, B. P., Gelfand, A. E., and Smith, A. F. (1992), "Hierarchical Bayesian Analysis of Changepoint Problems," *Applied Statistics*, 41(2), 389 – 405.

Cho, H., and Fryzlewicz, P. (2012), "Multiscale and Multilevel Technique for Consistent Segmentation of Nonstationary Time Series," *Statistica Sinica*, 22.

Davis, R., Lee, T., and Rodriguez-Yam, G. (2006), "Structural Break Estimation for Nonstationary Time Series Models," *Journal of the American Statistical Association*, 101(473), 223 – 239.

Fowlkes, E. B., and Mallows, C. L. (1983), "A Method for Comparing Two Hierarchical Clusterings," *Journal of the American Statistical Association*, 78(383), 553 – 569.

Gandy, A. (2009), "Sequential Implementation of Monte Carlo Tests with Uniformly Bounded Resampling Risk," *Journal of the American Statistical Association*, 104(488), 1504–1511.

Guralnik, V., and Srivastava, J. (1999), Event Detection From Time Series Data,, in *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '99, ACM.

Harchaoui, Z., and Cappe, O. (2007), Retrospective Mutiple Change-Point Estimation with Kernels,, in *Statistical Signal Processing, 2007. SSP '07. IEEE/SP 14th Workshop on*, pp. 768 –772.

Hariz, S. B., Wylie, J. J., and Zhang, Q. (2007), "Optimal Rate of Convergence for Nonparametric Change-Point Estimators for Nonstationary Sequences," *The Annals of Statistics*, 35(4), 1802 – 1826.

Hawkins, D. M. (2001), "Fitting Multiple Change-Point Models to Data," *Computational Statistics and Data Analysis*, 37(3), 323 – 341.

Hoeffding, W. (1961), The Strong Law of Large Numbers for U-statistics,, Technical Report 302, North Carolina State University. Dept. of Statistics.

Hubert, L., and Arabie, P. (1985), "Comparing Partitions," *Journal of Classification*, 2(1), 193 – 218.

James, N. A., and Matteson, D. S. (2013), ecp: An R Package for Nonparametric Multiple Change Point Analysis of Multivariate Data,, Technical report, Cornell University.

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

Kawahara, Y., and Sugiyama, M. (2011), "Sequential Change-Point Detection Based on Direct Density-Ratio Estimation," *Statistical Analysis and Data Mining*, 5(2), 114–127.

Killick, R., Fearnhead, P., and Eckley, I. (2012), "Optimal Detection of Changepoints With a Linear Computational Cost," *Journal of the American Statistical Association*, 107(500), 1590 – 1598.

Kim, A., Marzban, C., Percival, D., and Stuetzie, W. (2009), "Using Labeled Data to Evaluate Change Detctors in a Multivariate Streaming Environment," *Signal Processing*, 89(12), 2529 – 2536.

Lavielle, M., and Teyssière, G. (2006), "Detection of Multiple Change-points in Multivariate Time Series," *Lithuanian Mathematical Journal*, 46(3), 287 – 306.

Lung-Yut-Fong, A., Lévy-Leduc, C., and Cappé, O. (2011), "Homogeneity and Change-Point Detection Tests for Multivariate Data Using Rank Statistics," *arXiv:1107.1971*, .

Mampaey, M., and Vreeken, J. (2011), "Summarizing Categorical Data by Clustering Attributes," *Data Mining and Knowledge Discovery*, 24, 1 – 44.

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

Morey, L. C., and Agresti, A. (1984), "The Measurement of Classification Agreement: An Adjustment to the Rand Statistic for Chance Agreement.," *Educational and Psychological Measurement*, 44, 33 – 37.

Muggeo, V. M., and Adelfio, G. (2011), "Efficient Change Point Detection for Genomic Sequences of Continuous Measurements," *Bioinformatics*, 27, 161 – 166.

Olshen, A. B., and Venkatraman, E. (2004), "Circular Binary Segmentation for the Analysis of Array-Based DNA Copy Number Data," *Biostatistics*, 5, 557 – 572.

Page, E. (1954), "Continuous Inspection Schemes," *Biometrika*, 41, 100 – 115.

R Development Core Team (2012), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Rand, W. M. (1971), "Objective Criteria for the Evaluation of Clustering Methods," *Journal of the American Statistical Association*, 66, 846 – 850.

Rigaill, G. (2010), "Pruned Dynamic Programming for Optimal Multiple Change-Point Detection," *arXiv:1004.0887*, .

Rizzo, M. L., and Székely, G. J. (2010), "Disco Analysis: A Nonparametric Extension of Analysis of Variance," *The Annals of Applied Statistics*, 4(2), 1034–1055.

Sequeira, K., and Zaki, M. (2002), 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.

Székely, G. J., and Rizzo, M. L. (2005), "Hierarchical Clustering via Joint Between-Within Distances: Extending Ward's Minimum Variance Method," *Journal of Classification*, 22(2), 151 – 183.

Talih, M., and Hengartner, N. (2005), "Structural Learning With Time-Varying Components: Tracking the Cross-Section of Financial Time Series," *Journal of the Royal Statistical Society*, 67, 321 – 341.

Venkatraman, E. (1992), Consistency Results in Multiple Change-Point Problems, PhD thesis, Stanford University.

Vostrikova, L. (1981), "Detection Disorder in Multidimensional Random Processes," *Soviet Math Dokl.*, 24, 55 – 59.

Yao, Y. C. (1987), "Estimating the Number of Change-Points via Schwarz Criterion," *Statistics & Probability Letters*, 6, 181 – 189.

Zhang, N. R., and Siegmund, D. O. (2007), "A Modified Bayes Information Criterion with Applications to the Analysis of Comparative Genomic Hybridization Data," *Biometrics*, 63, 22 – 32.