A Comparison of Single and Multiple Changepoint Techniques for Time Series Data[⋆,⋆⋆]

Xuesheng Shi[a,b,*,1], Colin Gallagher[a], Robert Lund[b] and Rebecca Killick[c]

[a]School of Mathematics and Statistics, Clemson University, Clemson, SC 29634-0975
[b]Department of Statistics, The University of California - Santa Cruz, Santa Cruz, CA 95064
[c]Department of Mathematics and Statistics, Lancaster University, United Kingdom, LA1 4YF

ARTICLE INFO

Keywords:
AMOC Techniques
ARMA Models
Binary Segmentation
Brownian Bridge
CUSUM Tests
Likelihood Ratio Test
Minimum Description Length
One-step-ahead Prediction Residuals
Penalized Likelihoods
Wild Binary Segmentation.

ABSTRACT

Correlated time series data arise in many applications. This paper describes and compares several prominent single and multiple changepoint techniques for correlated time series. In the single changepoint problem, various cumulative sum (CUSUM) and likelihood ratio statistics, along with boundary cropping scenarios and scaling methods (e.g., scaling to an extreme value or Brownian Bridge limit) are compared. A recently developed test based on summing squared CUSUM statistics over all time indices is shown to have controlled Type I error and superior detection power. In the multiple changepoint setting, penalized likelihoods drive the discourse, with AIC, BIC, mBIC, and MDL penalties being considered. Binary and wild binary segmentation techniques are also compared. A new distance metric is introduced that measures differeces between two multiple changepoint segmentations. Algorithmic and computational concerns are discussed and simulations are given to support all conclusions. In the end, the multiple changepoint setting admits no clear methodological winner, performance depending on the particular scenario. Nonetheless, some practical guidance emerges.

1. Introduction

Changepoints (abrupt shifts) arise in many time series due to changes in recording equipment, observers, etc. In climatology, temperature trends computed from raw data can be misleading if homogeneity adjustments for station relocation moves and gauge changes are not *a priori* made to the record. Lu and Lund (2007) give an example where trend conclusions reverse when changepoint information is neglected. Cases with multiple changepoints are also frequently encountered; for example, in climatology, United States weather stations average about six station moves and/or gauge changes per century of operation (Menne, Williams, and Vose, 2009).

This paper intends to guide the researcher on the best changepoint techniques to use in common time series scenarios. Assumptions are crucial in changepoint analyses and can significantly alter conclusions; here, correlation issues take center stage. It is known that changepoint inferences made from positively correlated series can be spurious if correlation is not taken into account. Even lag one correlations as small as 0.25 can have deleterious consequences on changepoint conclusions (Lund, Wang, Lu, Reeves, Gallagher, and Feng, 2007).

This paper's primary contribution is to extend/modify many of the popular changepoint methods for IID data to correlated settings. Much of our work lies with developing methods that put all techniques, to the best extent possible, on the same footing in time series settings. For example, single changepoint tests will be shown to work best when applied to estimated versions of the series' one-step-ahead prediction residuals, computed under a null hypothesis of no changepoints. Because of this, tests that handle one-step-ahead prediction residuals need to be developed. Two other novel contributions in this article are: (1) developing and proposing a new single changepoint test based on the square of the cumulative sum of one-step-ahead prediction residuals (see Section 2.2), and 2) developing a new distance that compares multiple changepoint segmentations (see Section 5.1). The comparative aspect of the paper is yet another contribution — and there is much to compare. In addition to comparing different statistics via Type I errors and powers, the paper also compares different asymptotic scaling methods.

⋆ Killick gratefully acknowledges funding from EP/R01860X/1, EP/T021020/1 and NE/T006102/1.
⋆⋆ There is supplementary material to this paper available at: https://github.com/xuehens/ChangepointComparison
*Corresponding author
xshi38@ucsc.edu (X. Shi)
ORCID(s): 0000-0003-0185-2670 (X. Shi); 0000-0003-0583-3960 (R. Killick)

Academic changepoint research commenced with the single changepoint case for independent and identically distributed (IID) data in Page (1955). The subject is now vast, with hundreds of papers devoted to the topic. With our objectives, some concessions are necessary. Foremost, this paper examines mean shift changepoints only; that is, while series mean levels are allowed to abruptly shift, the variances and correlations of the series are held constant (stationary) in time. Changepoints can also occur in variances (volatilities) (Chapman, Eckley, and Killick, 2020), in the series' correlation structures (Davis, Lee, and Rodriguez-Yam, 2006; Aue and Horváth, 2013; Picard, 1985), or even in the marginal distribution of the series (Gallagher, Lund, and Robbins, 2012). Secondarily, the simulation results reported here are for Gaussian series only. Robust and non-parametric changepoint methods for non-Gaussian dependent data exist and can be based on the spectrum Picard (1985), empirical characteristic functions Hušková and Meintanis (2006), M-estimators (Hušková and Marušiaková, 2012; Hušková, 2013; Chochola, Hušková, Prášková, and Steinebach, 2013; Prášková and Chochola, 2014), or bootstrapping (Hušková and Kirch, 2008, 2012; Kirch, 2008). Thirdly, we compare the most common types of techniques within the literature, notably excluding those based on energy statistics (Matteson and James, 2014), moving sums (Eichinger and Kirch, 2018), and U statistics (Dehling, Fried, Garcia, and Wendler, 2015).

The rest of this paper proceeds as follows. Section 2 overviews single changepoint detection methods, typically referred to as at most one changepoint (AMOC) tests. Here, a variety of test statistics and their scalings are reviewed and adapted to the time series setting. Akin to the classifications in Aue and Horváth (2013), we specifically discuss two methods for modifying changepoint techniques based on IID data: 1) retain the IID test statistic and modify the limiting distribution for any correlation; and 2) modify the test statistic to account for the correlation; similar discussions appear in Robbins, Gallagher, Lund, and Aue (2011) and Aue and Horváth (2013). Section 3 compares AMOC detectors in a simulation study. Thereafter, we move to the case of multiple changepoints. Here, performance assessment becomes more challenging. For this, a novel changepoint configuration distance specifically designed for our comparisons is developed. Simulations in Section 4 consider a variety of multiple changepoint configurations. We summarize results in Section 6 with recommendations for practitioners.

2. Single Changepoint Techniques

Let $\{X_t\}_{t=1}^N$ be the observed time series and $\gamma(h) = \text{Cov}(X_{t+h}, X_t)$ be the lag h autocovariance of the series. We want to test whether there exists a change in the mean structure while assuming the second order structure is constant over time. An AMOC model with the changepoint occurring at the unknown time k is

$$X_t = \begin{cases} \mu + \epsilon_t, & \text{for } 1 \le t \le k, \\ \mu + \Delta + \epsilon_t, & \text{for } k+1 \le t \le N \end{cases}, \tag{1}$$

where μ is an unknown location parameter, Δ is the magnitude of mean shift at time k, and $\{\epsilon_t\}$ is a stationary time series with zero mean and lag h autocovariance $\gamma(h)$. A hypothesis test for this scenario is:

$$H_0 : \Delta = 0 \quad \text{versus} \quad H_1 : \Delta \neq 0 \quad \text{for some } k \in \{1, \dots, N-1\}. \tag{2}$$

When $\{\epsilon_t\}$ is IID, cumulative sum (CUSUM) and likelihood ratio tests (LRT) are well understood, see Chen and Gupta (2011). When incorporating general stationary autocovariance aspects into a changepoint testing framework, there are two common strategies: 1) keep the IID test statistic and identify any changes in the limiting distribution induced by the correlation; and 2) incorporate the autocovariance within the test statistic. Antoch, Hušková, and Prášková (1997) provide a summary of the first approach for many common changepoint statistics and provide simulations indicating how autocorrelation impacts the performance of the hypothesis tests; Kirch (2007) uses resampling techniques to improve the finite sample performance of these tests. Robbins et al. (2011) shows that estimating and using the autocorrelation (the second approach) is preferable with CUSUM and LRTs.

2.1. CUSUM Tests

The CUSUM method was first introduced by Page (1955) and compares sample means before and after each admissible changepoint time via the statistic

$$\max_{1 \le k < N} \left|\text{CUSUM}_X(k)\right| := \max_{1 \le k < N} \left| \frac{1}{\sqrt{N}} \left[\sum_{t=1}^{k} X_t - \frac{k}{N} \sum_{t=1}^{N} X_t \right] \right|. \tag{3}$$

CUSUM tests have relatively poor detection power when the changepoint occurs near the boundaries (times 1 or N). False detection is more likely to be signaled near the boundaries (i.e., when one of the segment sample means has a comparatively high variance). Because of this, cropped-CUSUM methods, which weight or ignore observations close to the two boundaries, were developed. Specifically, cropping strategies examine weighted statistics of form

$$\max_{1\le k<N} w_k \left|\mathrm{CUSUM}_X(k)\right|,$$

where w_k is some weight (this can be zero for some k). Typically, the weights are smaller for k near the two boundary times of 1 and N. Simulations for cropped settings analogous to those below are presented in the supplementary material; in general, one loses power by cropping. See Csörgo and Horváth (1997) for more on cropping.

In our first scenario, the IID test statistic described in (3) is used. Its asymptotic distribution for correlated data, under the null hypothesis of no changepoints, is known from MacNeill (1974), Csörgo and Horváth (1997), and Theorem 1 in Robbins et al. (2011).

Theorem 1. *Assume that $\{X_t\}$ follows (1), $\{\epsilon_t\}$ admits the causal linear representation $\epsilon_t = \sum_{i=0}^{\infty}\psi_i Z_{t-i}$ where $\sum_{i=0}^{\infty}|\psi_i| < \infty$, and $\hat{\eta}^2$ is a $\sqrt{N}$-based consistent estimator of η^2 under the null hypothesis, the long-run variance parameter*

$$\eta^2 := \lim_{n\to\infty}\frac{1}{n}Var\left(\sum_{t=1}^{n}\epsilon_t\right). \tag{4}$$

Then under H_0,

$$\frac{1}{\hat{\eta}}\max_{1\le k<N}\left|\mathrm{CUSUM}_X(k)\right| \xrightarrow{D} \sup_{t\in[0,1]}|B(t)|. \tag{5}$$

Here, it is assumed that $\{Z_t\}$ is IID with zero mean, variance σ^2, a finite fourth moment, and $\sum_{j=0}^{\infty}|\psi_j| < \infty$. Moreover, $\{B(t), t\in[0,1]\}$ denotes a standard Brownian bridge process obeying $B(t) = W(t) - tW(1)$, where $\{W(t), t\ge 0\}$ is a standard Wiener process.

Theorem 1 requires estimation of η^2, which is often challenging (Stoica and Moses, 2005).

While this result provides an asymptotic test, strong correlation often degrades CUSUM performance (Robbins et al., 2011). That is, convergence to the limit law is faster for independent data than for positively correlated data. As such, it is often beneficial to decorrelate heavily dependent data before applying CUSUM methods. This brings us to our second approach, which incorporates the correlation within the test statistic. For CUSUM methods, this is achieved by replacing the data by one-step-ahead linear prediction residuals.

The autoregressive moving average (ARMA) one-step-ahead linear prediction residuals are defined as:

$$\hat{Z}_t = \dot{X}_t - \hat{\phi}_1\dot{X}_{t-1} - \cdots - \hat{\phi}_p\dot{X}_{t-p} - \hat{\theta}_1\hat{Z}_{t-1} - \cdots - \hat{\theta}_q\hat{Z}_{t-q}, \tag{6}$$

where $\dot{X}_t = X_t - \hat{\mu}_t$, $\hat{\phi}_1, \ldots, \hat{\phi}_p$ are the estimated autoregressive coefficients, and $\hat{\theta}_1, \ldots, \hat{\theta}_q$ are the estimated moving-average coefficients. Here, the edge conditions take $\dot{X}_t = \hat{Z}_t = 0$ for any $t < 0$. Our notation uses σ^2 as the variance of any Z_t. We do not delve into ARMA order selection issues, and take p and q as known. Should this not be the case, one can revert to standard AIC and BIC methods to choose ARMA orders. For the CUSUM and SCUSUM tests described below, parameters are estimated under the changepoint free null hypothesis; in particular, $\hat{\mu}_t \equiv \bar{X} = N^{-1}\sum_{t=1}^{N}X_t$ is used to demean the time series. To evaluate Gaussian likelihoods, the innovations form of the likelihood is used; see Brockwell and Davis (1991). ARMA parameters are estimated in standard ways (for example, Yule Walker methods for autoregressions); again see Brockwell and Davis (1991) for additional detail.

The quantity $\hat{\sigma}^2 = N^{-1}\sum_{t=1}^{N}\hat{Z}_t^2$ is used to estimate the variance of Z_t. The residual CUSUM statistic is

$$\max_{1\le k<N}\left|\mathrm{CUSUM}_Z(k)\right| := \max_{1\le k<N}\left|\frac{1}{\sqrt{N}}\left(\sum_{t=1}^{k}\hat{Z}_t - \frac{k}{N}\sum_{t=1}^{N}\hat{Z}_t\right)\right|, \tag{7}$$

where our notation appends a subscript of Z to indicate use of prediction residuals.

The asymptotic distribution of the CUSUM of the one-step-ahead prediction residuals was established in Theorem 2 of Robbins et al. (2011).

Theorem 2. *Suppose that $\{\epsilon_t\}$ is a causal and invertible ARMA series with IID $\{Z_t\}$ having zero mean, variance σ^2, and with $E[Z_t^4] < \infty$. Let $\{\hat{Z}_t\}$ be the estimated one-step-ahead prediction residuals in (6). Then under the null hypothesis of no changepoints,*

$$\frac{1}{\hat{\sigma}} \max_{1 \le k < N} |\text{CUSUM}_Z(k)| - \frac{1}{\hat{\eta}} \max_{1 \le k < N} |\text{CUSUM}_X(k)| = o_p(1), \tag{8}$$

when all ARMA parameters and η^2 are estimated via some $\sqrt{N}$-consistent manner. It hence follows that

$$\frac{1}{\hat{\sigma}} \max_{1 \le k < N} |\text{CUSUM}_Z(k)| \xrightarrow{\mathcal{D}} \sup_{0 \le t \le 1} |B(t)|. \tag{9}$$

Both of these statistics are compared in Section 3.

2.2. SCUSUM Tests

As an alternative to using partial sums to detect mean shifts, several authors have considered summing the squares of the partial sums. The resulting test statistic converges to the integral of the square of a Brownian Bridge. With SCUSUM denoting the test's acronym, for IID data, the test statistic is

$$\text{SCUSUM} := \frac{1}{N} \sum_{k=1}^{N} \left[\frac{\text{CUSUM}(k)}{\hat{\sigma}} \right]^2. \tag{10}$$

The statistic in (10) also has a Bayesian interpretation with a discrete uniform prior over the changepoint time set $\{1, 2, \cdots, N\}$. The scenario is similar to the average likelihood ratio test considered in Chan and Walther (2013). The squared CUSUM (SCUSUM) test does not by itself yield an estimate of the changepoint location. If the SCUSUM test indicates that a changepoint is preferred, then its location is estimated as the argument(s) that maximizes the absolute CUSUM statistic.

We again consider two approaches for modifying the SCUSUM test for correlation. First, the distribution of the statistic in (10) for autocorrelated data under the null hypothesis can be quantified. The following result follows from Theorem 1 via an application of the continuous mapping theorem.

Theorem 3. *Assume that $\{X_t\}$ follows (1), $\{\epsilon_t\}$ admits the causal linear representation in Theorem 1, and $\hat{\eta}^2$ is a null hypothesis based $\sqrt{N}$-consistent estimator of η^2, the long-run variance in (4). Then under H_0,*

$$\text{SCUSUM}_X = \frac{1}{N} \sum_{k=1}^{N} \left[\frac{\text{CUSUM}_X(k)}{\hat{\eta}} \right]^2 \xrightarrow{\mathcal{D}} \int_0^1 B^2(t) dt. \tag{11}$$

Our second approach for incorporating correlation uses the one-step-ahead prediction residuals in place of the original data. The SCUSUM test statistic for this scheme is

$$\text{SCUSUM}_Z := \frac{1}{N} \sum_{k=1}^{N} \left[\frac{\text{CUSUM}_Z(k)}{\hat{\sigma}} \right]^2. \tag{12}$$

The asymptotic distribution of (12) can be established from Theorem 2 via the continuous mapping theorem.

Theorem 4. *With CUSUM_Z defined as in Theorem 2 and under the same assumptions in Theorem 2, under the null hypothesis of no changepoints,*

$$\text{SCUSUM}_Z = \frac{1}{N} \sum_{k=1}^{N} \left[\frac{\text{CUSUM}_Z(k)}{\hat{\sigma}} \right]^2 \xrightarrow{\mathcal{D}} \int_0^1 B^2(t)\, dt. \tag{13}$$

The distribution of $\int_0^1 B(t)^2 dt$ was investigated in Tolmatz (2002). We note that Bai (1993) proposed using the sum of the square of partial sums of ARMA residuals to detect a single changepoint in autocorrelated data; this test statistic converges to the integral of a squared *Brownian Motion* rather than the integral of the square of a *Brownian Bridge*. To our knowledge, the variant in (12) has not previously been proposed nor studied in the literature.

The differences between CUSUM and CUSUM_Z statistics were investigated in Robbins et al. (2011). Their simulations indicate that the latter statistic is superior to the former in terms of type I error and power. Our simulations confirm this finding. As such, in the remainder of the paper, we do not consider SCUSUM tests (without the subscript Z) further.

2.3. Likelihood Ratio Tests

While CUSUM tests are non-parametric, likelihood ratio tests (LRTs) are inherently parametric. Several error distributions have been considered in LRTs by previous authors, by far the most common being normal — this is the distribution considered here.

The LRT compares the likelihood under the null hypothesis to likelihoods under alternatives with a changepoint. The LRT statistic for a changepoint has the general form

$$\Lambda = \max_{1\le k<N} \Lambda_k, \qquad \Lambda_k = \frac{L_0(\hat{\mu}_0)}{L_k(\hat{\mu}_1, \hat{\mu}_2)}, \tag{14}$$

where L_0 denotes a null hypothesis likelihood and L_k is an alternative likelihood when the changepoint occurs at time k. Elaborating, $\hat{\mu}_0$ is the maximum likelihood estimator (MLE) for $E[X_t]$ under H_0, and $\hat{\mu}_1$ and $\hat{\mu}_2$ are the MLEs for the means of the two segments under the alternative when there is a mean shift at time k. The end statistic is then the maximum over all admissible changepoint locations k. When correlation exists in $\{X_t\}$, the form of the Gaussian likelihood can be found in Brockwell and Davis (1991); this form may contain additional ARMA or other correlation parameters that require estimation.

When the errors follow a causal and invertible Gaussian ARMA process, Jandhyala, Fotopoulos, MacNeill, and Liu (2013); Aue and Horváth (2013) develop asymptotics, scaling to an extreme value limit. While the asymptotics require one to estimate the ARMA parameters in calculation of the Λ_k statistics, the limit distribution does not depend on the ARMA parameters, nor does the scheme require any cropping of the boundary times.

Theorem 5. *Equations (1)-(3) in (Jandhyala et al., 2013). Suppose that $\{\epsilon_t\}$ is a causal and invertible ARMA series with IID $\{Z_t\}$ satisfying the assumptions in Theorem 2. Then the LRT statistic is*

$$U = \max_{1\le k<N} \left(-2\log(\Lambda_k)\right), \quad \Lambda_k = \left(\frac{\hat{\sigma}_k^2}{\hat{\sigma}_{H_0}^2}\right)^{\frac{N}{2}}. \tag{15}$$

Here, $\hat{\sigma}_k^2$ is the MLE estimate of the ARMA white noise process variance when there is a changepoint at time k and $\hat{\sigma}_{H_0}^2$ is an estimate of this same variance under the changepoint free null hypothesis. This statistic can be scaled to a Gumbel extreme value limit:

$$W_U := \sqrt{2U\log\log(N)} - \left[2\log\log(N) + \frac{1}{2}\log\log\log(N) - \frac{1}{2}\log\pi\right].$$

Then under H_0,

$$\lim_{N\to\infty} \mathbb{P}(W_U \le x) = \exp(-2\exp(-x)), \qquad -\infty < x < \infty. \tag{16}$$

Specifically, H_0 is rejected when W_U is too large to be explained by the distribution in (16).

Another way of scaling the Λ_k statistics involves cropping boundary times. Like the CUSUM test, the LRT is volatile at times near the boundaries. In fact, $\Lambda \xrightarrow{D} \infty$ as $N \to \infty$ should the maximum be taken over the entire range $1 \le k < N$ under the null hypothesis of no changepoints. A common cropped LRT simply truncates admissible times near the two boundaries; for example, with $0 < \ell < h < 1$, ℓ being close to zero and h being close to unity, set

$$U_{\text{crop}} = \max_{\ell\le k/N\le h} (-2\log(\Lambda_k)). \tag{17}$$

Robbins et al. (2011) shows that

$$U_{\text{crop}} \xrightarrow{D} \sup_{\ell\le t\le h} \frac{B^2(t)}{t(1-t)}. \tag{18}$$

As the next section shows, LRTs are not competitive in changepoint detection problems. While simulations are presented for the above extreme value test in the next section, simulations for cropped LRTs are delegated to the supplementary material — both methods perform poorly.

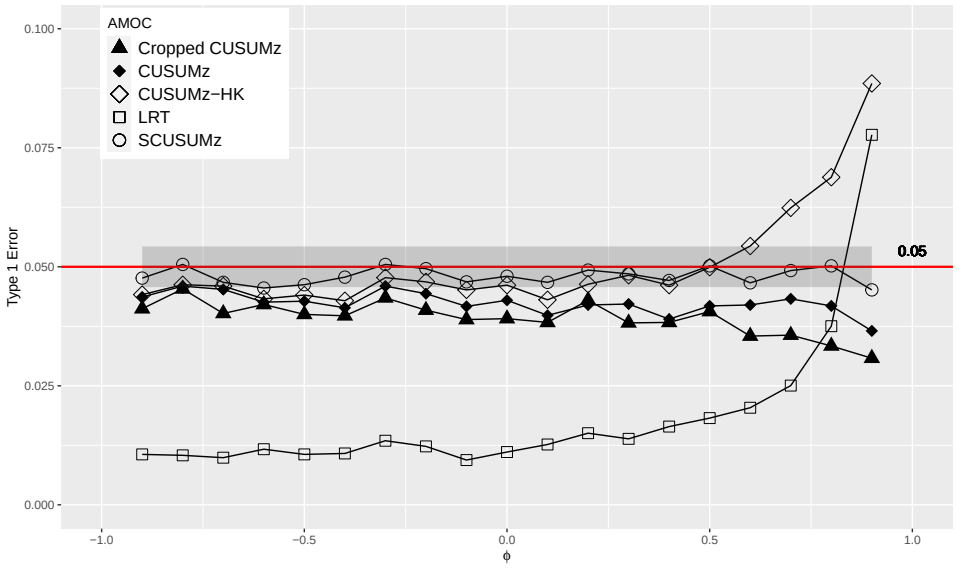

Figure 1: Type I Errors for an AR(1) Series with Different ϕ When $N = 1,000$. The grey band is a 95% pointwise confidence band based on the binomial standard errors $\sqrt{p(1-p)/N}$ assuming $p = 0.05$.

As a final comment here, deriving a LRT test for IID data, and then replacing the data with one-step-ahead prediction residuals, another avenue for dealing with dependence, does not yield a methodologically distinct path. Specifically, if one derives a LRT statistic for independent Gaussian series and then substitutes one-step-ahead prediction residuals in place of the original data, the limit law in (18) again arises. The boundaries again must be cropped to ensure a proper limiting distribution. The discussion around (1.4.22) — (1.4.27) in (Csörgo and Horváth, 1997) provides additional detail on this route; see also Lavielle and Moulines (2000) for more on LRTs for correlated data.

3. AMOC Simulations

This section investigates the finite sample performance of the Section 2 tests (cropped CUSUM_Z, CUSUM_Z, SCUSUM_Z, LRT) through simulation. Results for the cropped test statistics are delegated to the supplementary material; results for the other tests are presented here.

Desirable tests have reasonable (non-inflated) false detection rates when no changepoints exist, and large detection powers when a changepoint is present, regardless of the degree of correlation. For each statistic under consideration, the impact of autocorrelation on the Type I error is first explored. We then examine detection powers of the tests when a changepoint exists. First order Gaussian autoregressions (AR(1)) are considered here with $\sigma^2 = 1$; other structures are examined in the supplementary material.

Figure 1 summarizes results with $N = 1,000$ across varying values of the AR(1) correlation parameter ϕ. Ten thousand independent simulations were run for each considered value of ϕ to produce the figure. Our conclusions do not vary for different N — see the supplementary material. Figure 1 shows that the only method to retain a controlled Type I error across all ϕ is the SCUSUM_Z. The LRT is the worst performing method, being far too conservative except when $\phi = 0.95$, when it becomes highly inflated. The poor performance of the LRT is likely due to the slow convergence to its extreme value limit, which has been previously noted (see page 25 of Csörgo and Horváth (1997)). The CUSUM_Z method is also slightly conservative, becoming more so as ϕ increases. We would expect the 0.05 type I error to be reasonably maintained when $N = 1,000$.

We now consider test detection powers. In general, the detection power of an AMOC test depends on the degree of correlation, the size of the mean shift, and the location of the changepoint time (Robbins, Gallagher, and Lund, 2016). It is reasonable to expect power to be a function of the quantity $|\Delta|/\eta$ — the magnitude of the mean shift scaled to the series standard deviation. Figures 2 ($\Delta = 0.15$) and 3 ($\Delta = 0.3$) show empirical powers based on 10,000 independent Gaussian simulated series of length $N = 1,000$. Sample powers are plotted as a function of ϕ when the mean shift lies in the center of the series (time 501). The figures demonstrate the drastic effects of autocorrelation on the power of

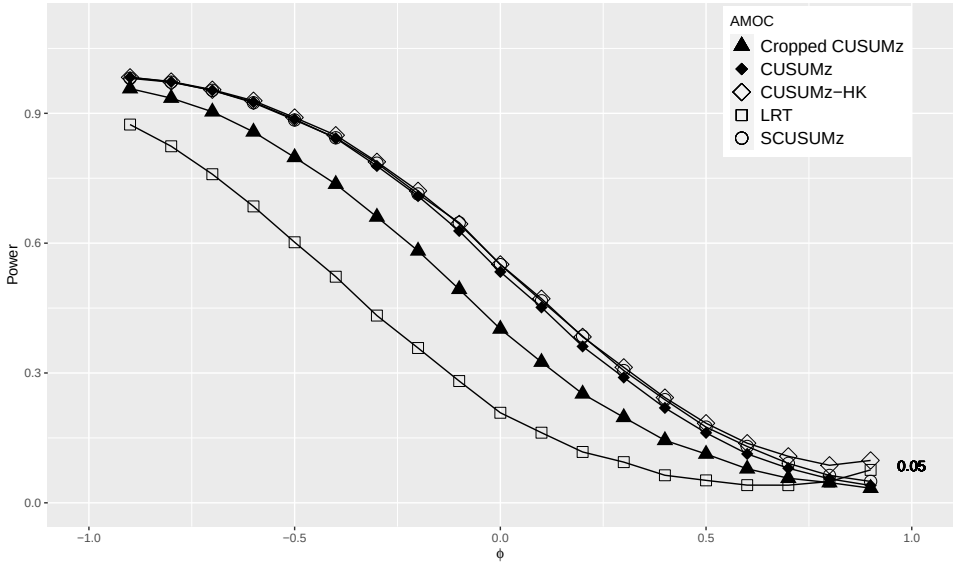

Figure 2: Detection Powers for an AR(1) Series with Different ϕ. Here, $N = 1,000$ and $\Delta = 0.15$.

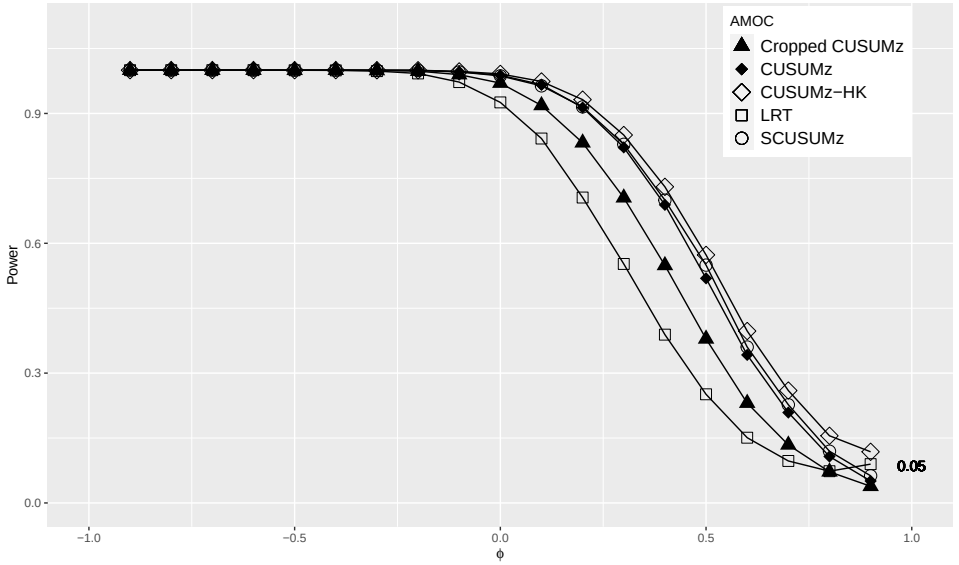

Figure 3: Detection Power for an AR(1) Series with Different ϕ. Here, $N = 1,000$ and $\Delta = 0.3$.

changepoint tests. While the LRT had the highest empirical power when $\phi = 0.95$, the estimated changepoint location of LRT is biased and more variable than that for the CUSUM$_Z$ and SCUSUM$_Z$ tests — see Figures 4 and 5. The LRT test also has a Type I error far exceeding 0.05; as such, its higher power does not suggest better overall performance. Overall, the CUSUM$_Z$ and SCUSUM$_Z$ tests are more powerful than the others. Note also that SCUSUM$_Z$ has higher power than CUSUM$_Z$ for each ϕ considered. Additional simulations (not shown) duplicate this conclusion for other sample sizes. The SCUSUM$_Z$ statistic is clearly the best.

The variance of an AR(1) series is $\sigma^2/(1-\phi^2)$ and changes with ϕ. Analogous simulations to the above where Δ is taken to make $|\Delta|/\eta$ constant for all ϕ were conducted. This makes power comparisons fairer across varying ϕ. The results are shown in the supplementary material. The performance orderings of the methods in the above figures does not change.

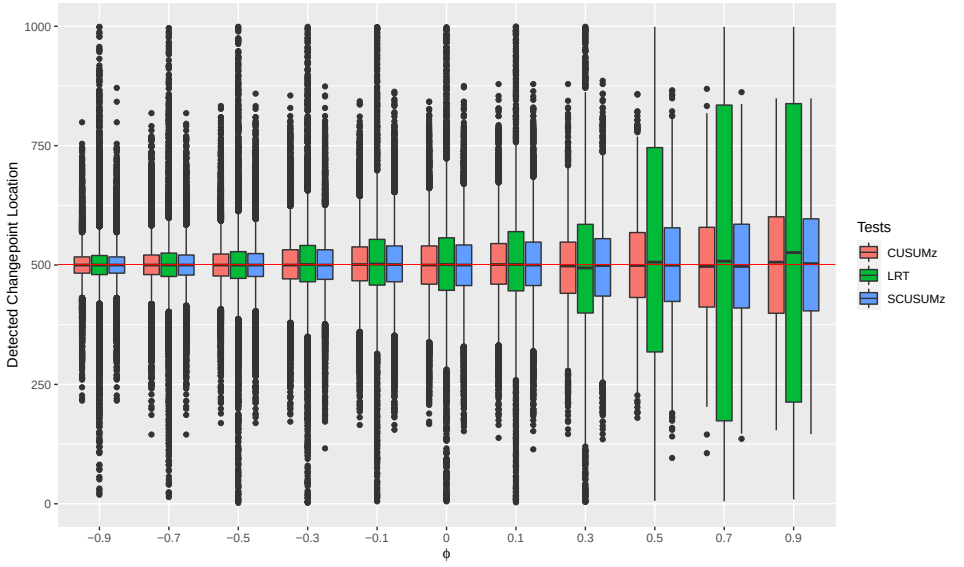

Figure 4: Boxplots of Detected Changepoint Locations for an AR(1) Series with Different ϕ. Here, $N = 1,000$ and $\Delta = 0.15$.

It is possible to increase detection power (2) by estimating the ARMA model parameters under the alternative hypothesis. A procedure inspired by Hušková and Kirch (2008) finds an initial estimate of the changepoint time using the argument that maximizes (3), then demeans the the two segments to estimate any time series nuisance parameters. After this, any AMOC statistic from the previous section can be used. We find that this method indeed increases power, but sometimes also increases Type I error. For the CUSUM_Z statistic, this is seen in Figures 1-3 where the procedure, denoted by CUSUM_Z-HK has increased power, but an inflated type I error for $\phi > 0.5$. Similar results (not shown here) were found for the other statistics. While using the estimator in (19) below could improve results, this was not pursued since our purpose is to make relative comparisons between the tests.

Finally, we examine the effect of the changepoint location. Simulation specifications are as above, but the location of the changepoint is now varied and ϕ is fixed as 0.5. Figure 6 displays empirical powers. The largest detection powers occur when the changepoint is near the center of the record, as expected, with power decreasing as the changepoint time moves towards a boundary. The SCUSUM_Z appears to be the most accurate overall; however, the LRT test is preferable when the changepoint occurs near the beginning of the record.

4. Multiple Changepoint Techniques

Now suppose that $\{X_t\}_{t=1}^N$ has an unknown number of changepoints, denoted by m, occurring at the unknown ordered times $1 < \tau_1 < \tau_2 < \cdots < \tau_m \leq N$. Boundary conditions take $\tau_0 = 1$ and $\tau_{m+1} = N + 1$. These m changepoints partition the series into $m+1$ distinct regimes, the i^{th} regime having its own distinct mean and containing the data points $\{X_{\tau_i+1}, \ldots, X_{\tau_{i+1}}\}$. The model can be written as $X_t = \kappa_t + \epsilon_t$, where $\kappa_t = \mu_{r(t)}$ and $r(t)$ denotes the regime index at time t, which takes values in $\{0, 1, \ldots, m\}$, and $\{\epsilon_t\}$ is a stationary causal and invertible ARMA(p, q) time series that applies to all regimes. Observe that

$$\kappa_t = \begin{cases} \mu_0, & 1 \leq t \leq \tau_1, \\ \mu_1, & \tau_1 + 1 \leq t \leq \tau_2, \\ \quad \vdots & \\ \mu_m, & \tau_m + 1 \leq t \leq N \end{cases}.$$

There are many challenges in the multiple changepoint problem. For us, estimation of the global autocovariance function that applies to all regimes, which is considered further in Section 4.2, is difficult. One also has to estimate an

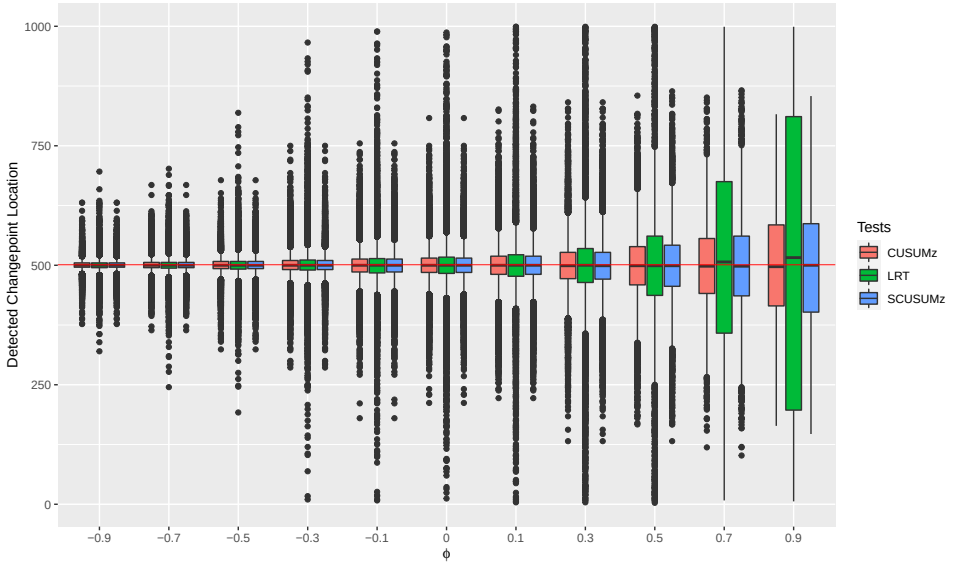

Figure 5: Detected Changepoint Location for an AR(1) Series with Different ϕ. Here, $N = 1000$ and $\Delta = 0.3$.

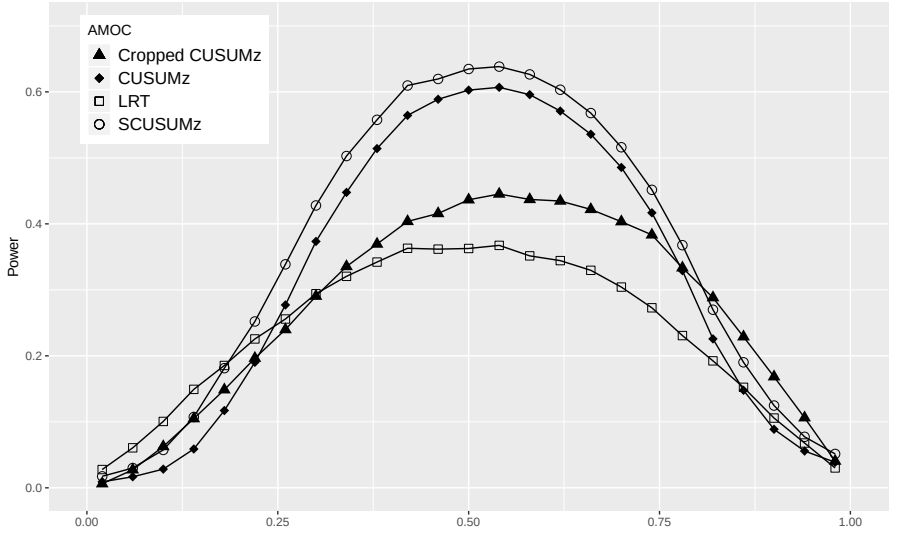

Figure 6: A Graph of τ/N Against Detection Power with $N = 500$ and $\Delta = 0.5$ for an AR(1) Series with $\phi = 0.5$.

unknown number of changepoints, their locations, and all segment parameters in a computationally feasible manner for some of the techniques.

While many authors have considered multiple changepoint issues, most assume IID $\{\epsilon_t\}$. For IID errors, dynamic programming based approaches (Auger and Lawrence, 1989; Killick, Fearnhead, and Eckley, 2012), model selection methods using LASSO (Harchaoui and Lévy-Leduc, 2010; Shen, Gallagher, and Lu, 2014), and moving sum statistics (Eichinger and Kirch, 2018) have all been applied to multiple changepoint problems — this list is not exhaustive. As in the AMOC setting, techniques for independent series may not work well for dependent series (Davis et al., 2006; Li and Lund, 2012; Chakar, Lebarbier, Lévy-Leduc, and Robin, 2017).

The multiple changepoint techniques considered here can be put into two broad categories: 1) recursive segmentation and algorithmic methods using AMOC techniques, and 2) direct approaches that fit all series subsegments jointly.

The two approaches are completely different in their perspective. Elaborating, most recursive techniques employ AMOC single changepoint methods in an iterative manner, identifying at most one additional changepoint in each series subsegment at each recursion level. In contrast, direct techniques model and estimate the multiple changepoint configuration jointly; here, penalization methods typically drive the discourse. No hypothesis testing paradigm underlies any multiple changepoint approach — there is not a clear alternative hypothesis here. Some multiple changepoint techniques apply only to special time series structures. For example, Chakar et al. (2017) is exclusively designed for AR(1) series. Their techniques are not considered here as they cannot be applied to all of our considered scenarios.

4.1. Recursive Segmentation Approaches

Recursive segmentation approaches first focus on finding a single changepoint (usually the most prominent one), thereafter iterating in some manner to identify additional changepoints. The primary tool here is binary segmentation (Scott and Knott, 1974), which estimates a multiple changepoint configuration via any AMOC method. Elaborating, binary segmentation first tests the entire series for a single changepoint. Should a changepoint be found, the series is split about the changepoint time into two subsegments that are further analyzed for additional changepoints using the AMOC strategy. The process is repeated until no subsegment tests positive for a changepoint. Binary segmentation works best when the changepoints are well separated and the segment means are distinct. In our comparisons, the AMOC statistic adopted for binary segmentation is the SCUSUM test applied to one-step-ahead prediction residuals, which won our AMOC comparisons in the previous section.

Extensions of binary segmentation abound and include circular binary segmentation (Olshen, Venkatraman, Lucito, and Wigler, 2004), which seeks to identify a segment of data that has a distinct mean from the rest of the series. Another popular binary segmentation extension is wild binary segmentation (WBS) Fryzlewicz (2014). WBS samples subsegments of the entire series having random lengths and performs an AMOC test on each sampled subsegment. Fryzlewicz (2014) suggests sampling at least $(9N^2)\log(N^2\delta^{-1})/(\delta^2)$ subsegments, where δ is the minimum spacing between changepoints (see Assumption 3.2 of Fryzlewicz (2014)) as this produces a high probability of drawing a favorable subsegment. WBS is a randomized search and hence may return different segmentations on different runs. In our simulations, WBS uses a standard CUSUM test rather than the cropped CUSUM or SCUSUM since its threshold was developed particularly for standard CUSUM methods. In addition, the threshold constant $C = 1.3$ is used as suggested in Fryzlewicz (2014).

Binary segmentation approaches and their variants are simple to implement and are computationally fast. However, they are not guaranteed to achieve the global optimal solution as they essentially are a "greedy algorithm" that sequentially makes decisions based solely on information during the current step. Also inherent in these approaches is the need for the AMOC statistic to behave appropriately when multiple mean shifts are present — this may not happen.

To apply the above segmentation methods in the presence of autocorrelation, we need to develop estimates of the time series autocorrelation parameters that are robust to mean shifts. This autocorrelation needs to be estimated *a priori* to segmentation. The next section elaborates further.

4.2. Global Autocovariance Estimation

For our work, the autocovariance of the series is assumed constant across time and applies to all series subsegments. This autocovariance function will be used to decorrelate the series before applying any binary segmentation search methods to the one-step-ahead prediction residuals. Unfortunately, accurate estimation of the autocovariance function requires knowledge of the underlying mean structure. In the single changepoint case, the long-run covariance parameter in (4) arises in the limit laws; however, this does not extend to multiple changepoint settings, where no theoretical equivalent of (5) exists.

In our setup, the second order (covariance) model parameters are deemed nuisance parameters and are estimated using the entire series. To account for the impact of unknown mean shifts on these estimators, Yule-Walker type moment equations will be used on the first order difference of $\{X_t\}$. The first order difference $X_t - X_{t-1}$ is used because $E[X_t - X_{t-1}] = 0$ unless a changepoint occurs at time t. Define $d_t = X_t - X_{t-1}$ and note that $d_t = \epsilon_t - \epsilon_{t-1}$ except when time t is a changepoint. Let $\gamma_d(h) = \mathrm{Cov}(d_t, d_{t-h})$, $\rho_d(h) = \gamma_d(h)/\gamma_d(0)$ and $\boldsymbol{\rho}_d = (\rho_d(1), \cdots, \rho_d(p))'$. For the AR($p$) case, which is our primary interest, estimators of the AR(p) parameters are based on $\{d_t\}$ and have the form

$$\widehat{\boldsymbol{\phi}} = \widehat{\boldsymbol{M}}^{-1}\widehat{\boldsymbol{\rho}}_d, \tag{19}$$

where $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)'$ and

$$\boldsymbol{M} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & -\left(\frac{1}{2} + \rho_d(1)\right) & \cdots & -\left(\frac{1}{2} + \sum_{j=1}^{p-2} \rho_d(j)\right) \\ \rho_d(1) & \rho_d(0) & \rho_d(1) & \cdots & \rho_d(p-2) \\ \rho_d(2) & \rho_d(1) & \rho_d(0) & \cdots & \rho_d(p-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_d(p-1) & \rho_d(p-2) & \rho_d(p-3) & \cdots & \rho_d(0) \end{bmatrix}.$$

The elements in $\hat{\boldsymbol{M}}$ and $\hat{\boldsymbol{\rho}}_d$ simply replace $\rho_d(h)$ with

$$\hat{\rho}_d(h) = \frac{\hat{\gamma}_d(h)}{\hat{\gamma}_d(0)} = \frac{\sum_{t=2}^{n-h}(X_t - X_{t-1})(X_{t+h} - X_{t+h-1})}{\sum_{t=2}^{n}(X_t - X_{t-1})^2}.$$

While Gallagher, Killick, Lund, and Shi (2021) study these AR(p) estimators in detail, the intuition behind them is that if m is small relative to N, then the mean shifts will have negligible impact on the estimated covariance structure of the differences since d_t has zero mean except at the changepoint times. Gallagher et al. (2021) demonstrate that this estimate of the covariance outperforms alternatives such as direct and windowed estimation. Due to this, the Yule-Walker moment estimators in (19) will be used in our simulations to decorrelate the series for binary segmentation and wild binary segmentation.

4.3. Direct Modelling Approaches

Direct modelling approaches analyze the whole series at once, optimizing an objective function with a penalty term that controls the number of changepoints. The techniques seek a changepoint configuration that minimizes

$$F(m; \tau_1, \ldots, \tau_m) := C(m; \tau_1, \ldots, \tau_m) + P(m; \tau_1, \ldots, \tau_m), \tag{20}$$

where C is the cost of putting m changepoints at the times $\tau_1, \ldots, \tau_m$ and P is a penalty term to prevent over-fitting. There are many ways to define the cost and penalties. A frequently used cost is the negative log-likelihood. Here, we will use

$$C(m; \tau_1, \ldots, \tau_m) = -2\log(L_{\text{opt}}(\boldsymbol{\theta}|m; \tau_1, \ldots \tau_m)).$$

where $L_{\text{opt}}(\boldsymbol{\theta}|m; \tau_1, \ldots, \tau_m)$ is the time series likelihood (Gaussian based) optimized over all parameters $\boldsymbol{\theta}$ given that m changepoints occur at the times $\tau_1, \ldots, \tau_m$. From a given changepoint configuration, finding this optimal likelihood is a simple time series model fitting exercise that can be rapidly computed.

Penalties can be constructed in a variety of ways. Common penalties include minimum description lengths (MDL), modified Bayesian Information Criterion (mBIC), and the classic BIC penalty. AIC is another popular penalty, despite it not providing consistent estimates of the number or locations of the changepoint(s). Of these four penalties, AIC and BIC are simple multiples of the number of changepoints, while the MDL and mBIC further incorporate changepoint time information. The form of these penalties are listed in Table 1. In these tables, Gaussian dynamics implies that $-2\log(L_{\text{opt}}(\boldsymbol{\theta}|m; \tau_1, \ldots \tau_m))$ is $N \ln(\hat{\sigma}^2)$ plus some constant that does not depend on $\boldsymbol{\theta}$. The mBIC and MDL penalties are multiplied by two to be consistent with the AIC and BIC definitions that use twice the negative log-likelihood. Here, $\hat{\sigma}^2$ is the estimated white noise variance of the $\{\epsilon_t\}$ process that drives the ARMA errors.

MDL penalties are based on information theory and are discussed further in Davis et al. (2006) and Li and Lund (2012). The mBIC penalty is developed in Zhang and Siegmund (2007). These two penalties are taken as zero when $m = 0$. The mBIC penalty tends to be larger for the same changepoint configuration than the MDL penalty; as such, MDL tends to select models with more changepoints than mBIC.

With penalized likelihood approaches, a computational bottleneck arises. Since there are $\binom{N-1}{m}$ different admissible changepoint configurations in a series with m changepoints (time N cannot be a changepoint), there are 2^{N-1} different changepoint configurations to consider when analyzing the entire series. This huge count makes an exhaustive model search — one that evaluates all admissible changepoint configurations — virtually impossible to conduct, even when N is a small as 100. Unfortunately, PELT (Killick et al., 2012) and FPOP (Maidstone, Hocking, Rigaill, and Fearnhead, 2017), two rapid dynamic programming based techniques, require the objective function to be additive

Table 1
Penalized Likelihood Objective Functions

Criteria	Objective Function
AIC	$N\ln(\hat{\sigma}^2) + 2(2m+3)$
BIC	$N\ln(\hat{\sigma}^2) + (2m+2)\ln(N)$
mBIC	$N\ln(\hat{\sigma}^2) + 3m\ln(N) + \sum_{i=1}^{m+1}\ln\left(\frac{\tau_i-\tau_{i-1}}{N}\right)$
MDL	$N\ln(\hat{\sigma}^2) + 2\ln(m) + \sum_{i=1}^{m+1}\ln(\tau_i-\tau_{i-1}) + 2\sum_{i=2}^{m}\ln(\tau_i)$

over distinct regimes. The presence of global parameters like the autocovariance function violates this requirement. Regime-additive likelihoods will not arise when $\{\epsilon_t\}$ is ARMA(p,q), although Bai and Perron (1998) argues that any boundary contribution is negligible if the ARMA parameters are allowed to change at each changepoint time (this is not the case here). Unfortunately, the objective function in (20) is not convex, and its optimization is delicate. We will use a genetic algorithm (GA), which have successfully dealt with this and similar changepoint optimization problems (Davis et al., 2006; Li and Lund, 2012).

A GA is an intelligent random walk search that is unlikely to evaluate suboptimal changepoint configurations. Research indicates that genetic algorithms perform well in nonconvex optimization problems (Hajela, 1990). Our GA encodes the changepoint configuration into a binary string and uses the `R GA` package from Scrucca (2013). This GA has proven reliable with our problems.

5. Multiple Changepoint Simulations

In presenting simulation results for different scenarios, the main body of the text will only present graphic(s) that are judged informative. In general, for each simulation case considered, graphics of configuration distances to truth, average number of detected changepoints, and empirical probabilities of estimating the correct number of changepoints were produced. The supplementary material contains any graphics that are not included in the main body. Similarly, we focus on unit shift mean sizes in the main text body unless otherwise noted; results for different mean shift sizes are presented in the supplementary material.

The changepoint configurations that we consider are illustrated in Figure 7, which shows sample time series generated under the various mean shift configurations. These configurations range from scenarios with no or few changepoints to those with a large number of changepoints. All series have length $N = 500$.

Due to AIC's popularity, it is worth mentioning that this penalty performs miserably in all scenarios, always selecting an excessive number of changepoints. This issue was also mentioned in Yao (1988). Since plotting AIC results would degrade our other graphical comparisons, AIC results are not presented so that we may accentuate differences in the remaining methods.

5.1. Comparing Multiple Changepoint Segmentations

Before presenting our simulations, we discuss how to compare an estimated multiple changepoint segmentation to its true value. The estimated multiple changepoint configuration could have a different number of changepoints than the true configuration. For a single changepoint method, such a comparison is easy: examine first whether the method flags a changepoint, and then its distance from the true changepoint time. With multiple changepoint configurations, this comparison is complicated by the fact that different segmentations may have different numbers of changepoints: which changepoint times in one particular configuration correspond to those in the other may be nebulous.

To compare different methods, a distance between the two changepoint configurations $\mathcal{C}_1 = (m;\tau_1,\ldots,\tau_m)$ and $\mathcal{C}_2 = (k;\eta_1,\ldots,\eta_k)$ will now be developed. Several distances have been utilized by the multiple changepoint community. Some, such as the mean squared error (MSE) of the fitted means, V-measure, or Hausdorff distance, are not specific to changepoint problems. Others, such as the number of changepoints or true/false positive detection rates, are more tailored to the changepoint problem. However, each of these statistics quantifies only one aspect of the fit. For example, the MSE could be low, but the number of changepoints could still be overestimated; or the number of

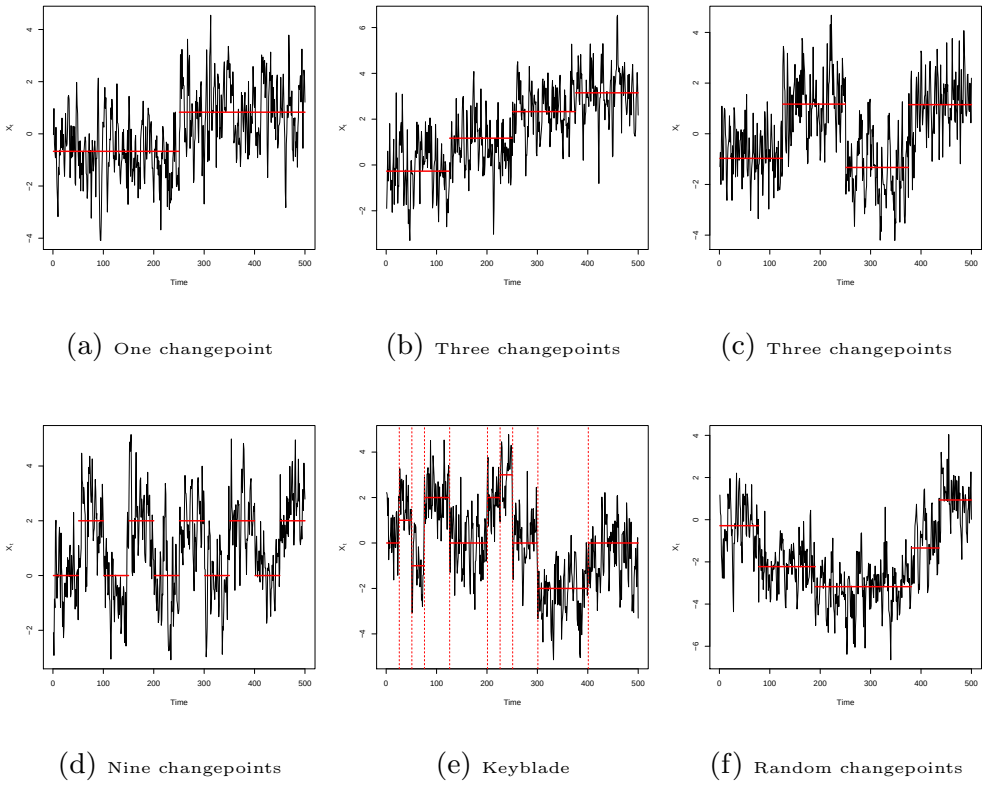

Figure 7: AR(1) Time Series with Different Changepoint Settings

changepoints could be accurate, but their locations inaccurate. As such, we introduce a new changepoint-specific distance balancing the two key components of multiple changepoint analysis: 1) the number of changepoints and 2) their individual locations.

To balance the number and location aspects of changepoint configurations, two components in our distance are needed. The first measures the discrepancy in the numbers of changepoints in the two configurations, for which we use absolute difference. The second component measures discrepancies in the changepoint times. This is trickier to quantify as the number of changepoints may be different in the two configurations and some sort of "matching procedure" is needed. For two changepoint segmentations, $\mathcal{C}_1$ and $\mathcal{C}_2$, the distance used here is

$$d(\mathcal{C}_1, \mathcal{C}_2) = |m - k| + \min\{\mathcal{A}(\mathcal{C}_1, \mathcal{C}_2)\}. \tag{21}$$

The term $|m-k|$ assigns the difference in changepoint numbers for any mismatch in the total number of changepoints. The term $\min \mathcal{A}(\mathcal{C}_1, \mathcal{C}_2)$ reflects the smallest cost that matches changepoint locations in $\mathcal{C}_1$ to those in $\mathcal{C}_2$. This term can be computed via the following linear assignment methods:

$$\mathcal{A}(\mathcal{C}_1, \mathcal{C}_2) = \sum_{i=1}^{k} \sum_{j=1}^{m} c_{i,j} I_{i,j},$$

which is subject to the constraints $\sum_{i=1}^{k} I_{i,j} = 1$, for $j \in \{1, \dots, m\}$ and $\sum_{j=1}^{m} I_{i,j} \leq 1$ for $i \in \{1, \dots, k\}$. Here, the cost of assigning τ_i to η_j is taken simply as $c_{i,j} = |\tau_i - \eta_j|/N$ and $I_{i,j} \in \{0, 1\}$ is the decision variable

$$I_{i,j} = \begin{cases} 1 & \text{if } \tau_i \text{ is assigned to } \eta_j \\ 0 & \text{otherwise} \end{cases}.$$

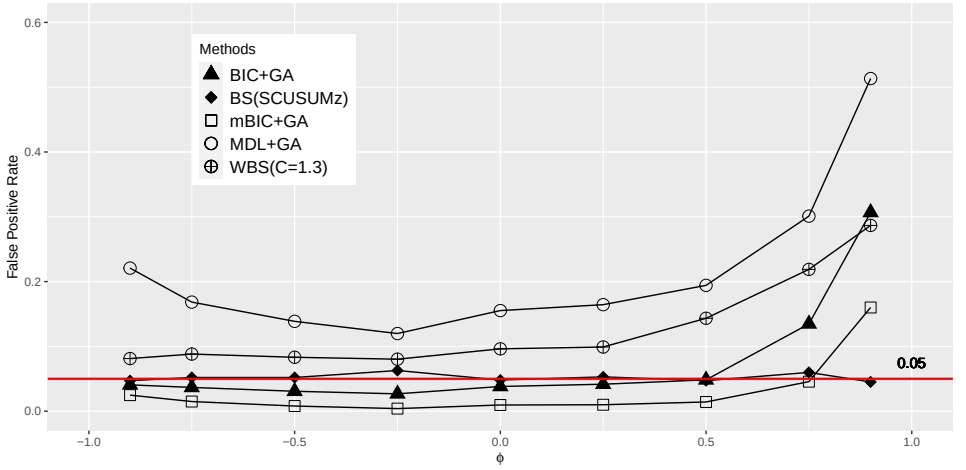

Figure 8: Empirical False Positive Detection Rates for an AR(1) Series with Various ϕ. Truth: No Changepoints.

This linear assignment problem can be efficiently computed via algorithms in Burkard, Dell'Amico, and Martello (2012).

One can verify that (21) defines a legitimate distance satisfying the triangle inequality. The larger the distance is, the worse the two configurations correspond to one another. The term $\min \mathcal{A}(C_1, C_2)$ can be shown to be bounded by unity and measures how closely the two changepoint configurations match up to one another. When both configurations have many changepoints, the distance is dominated by the $|m - k|$ term. In our simulations, estimated multiple changepoint configurations will be compared to the true changepoint configuration with this distance.

5.2. No Changepoints

Many modern multiple changepoint simulation studies increasingly focus on cases with a large number of changepoints, eschewing single and no changepoint scenarios. We include low changepoints scenarios here to help illuminate the differences between the methods.

Our first simulation considers the changepoint free case in an AR(1) Gaussian series having various correlation parameters ϕ and $\sigma^2 = 1$. Figure 8 shows probabilities of falsely declaring one or more changepoints over 1,000 independent simulations. Unlike the single changepoint case, the methods here do not control any false positive rate.

The results show that BIC, mBIC, and binary segmentation perform best, with WBS and MDL performing significantly worse. It is worth noting that WBS has a signicantly higher false positive rate, an issue discussed further in Lund and Shi (2020). Binary segmentation is arguably best here, an expected finding with no changepoints (an AMOC test applied to the series' one-step-ahead prediction residuals should not see a changepoint and stop any recursion at its onset). All methods perform better with negative ϕ than with positive ϕ; performance of all methods degrades as ϕ moves towards unity (as expected).

5.3. A Single Changepoint

We now move to simulations with one changepoint in the same AR(1) setup above. The changepoint is placed in the middle of the series, $t = 251$. Figure 9 shows the average distances between the estimated changepoint configurations and the true configuration. While there are no huge discrepancies between the methods, for heavily correlated series, binary segmentation is the worst and MDL and mBIC the best. Again, all tests degrade as ϕ approaches unity. MDL exhibits the least variability across ϕ. Comparing to the single changepoint results, the multiple changepoint penalties are more conservative than the LRT. Also, since the average distance is less than unity, the correct number of changepoints is often being identified.

5.4. A Three Changepoint Staircase

Our next case moves to a setting with three mean shifts, partitioning the series into four equal-length regimes. The changepoints occur at times 126, 251, and 376, with each changepoint shifting the series upward by one unit (up-up-up). As before, Figure 10 reports average distances. MDL performs the worst for negative ϕ, the other methods

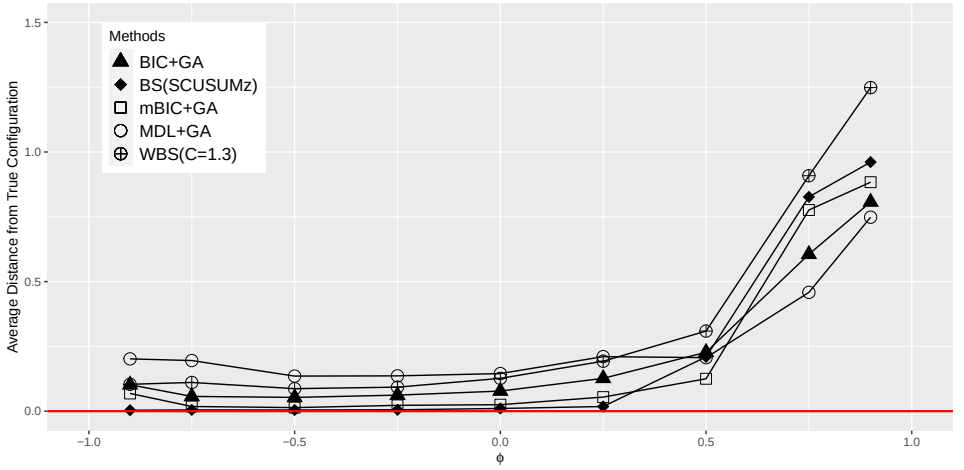

Figure 9: Average Distances for an AR(1) Series with Varying ϕ. Truth: One Changepoint in the Middle Moving the Series Upwards.

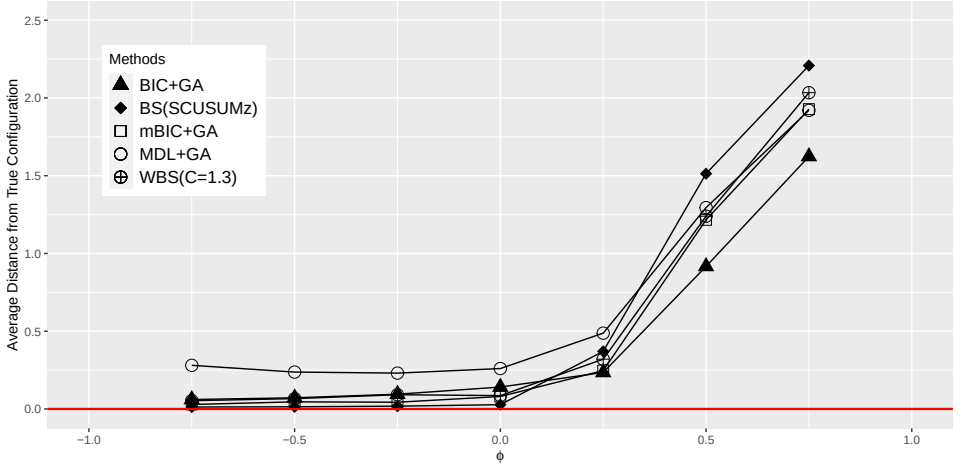

Figure 10: Average Distances for AR(1) Series with Different ϕ. Truth: Three Equally Spaced Changepoints Moving the Series Up-Up-Up.

perform similarly here. Perhaps surprisingly, binary segmentation starts to degrade when ϕ becomes positive, with the other methods also degrading, but to a lesser extent. BIC performs best across all ϕ.

5.5. Three Alternating Changepoints

Next, we consider another three changepoint configuration, the changepoint times again being equally spaced, but this time moving the series up, then down, and then up again (up-down-up). Figure 11 reports the distances. All methods have a harder time here than with the last up-up-up changepoint configuration. In this setting, binary segmentation becomes fooled and estimates too few changepoints; mBIC is inferior to the other methods. MDL and WBS work better, the surprise winner being BIC.

5.6. A Nine Changepoint Staircase

Next, we move to cases with nine changepoints. Our first set of simulations equally spaces all changepoint times in the record, each moving the series higher (All Up). Because the changepoints are more difficult to detect, we have increased the absolute mean shift magnitude to two units — this serves to induce more separation between the methods, allowing for an easier comparison. Figure 12 displays distances for this setting. The winners are BIC and MDL; losers

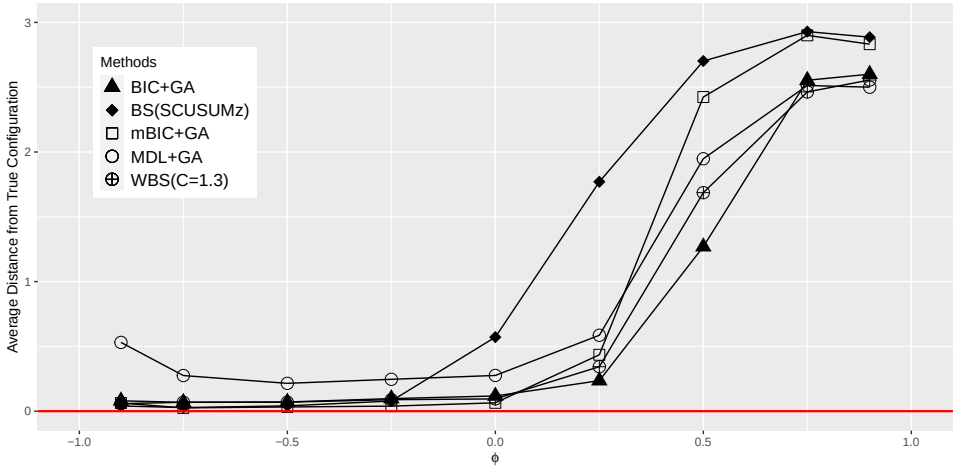

Figure 11: Average Distances for an AR(1) Series with Varying ϕ. Truth: Three Equally Spaced Changepoints Moving the Series Up-Down-Up.

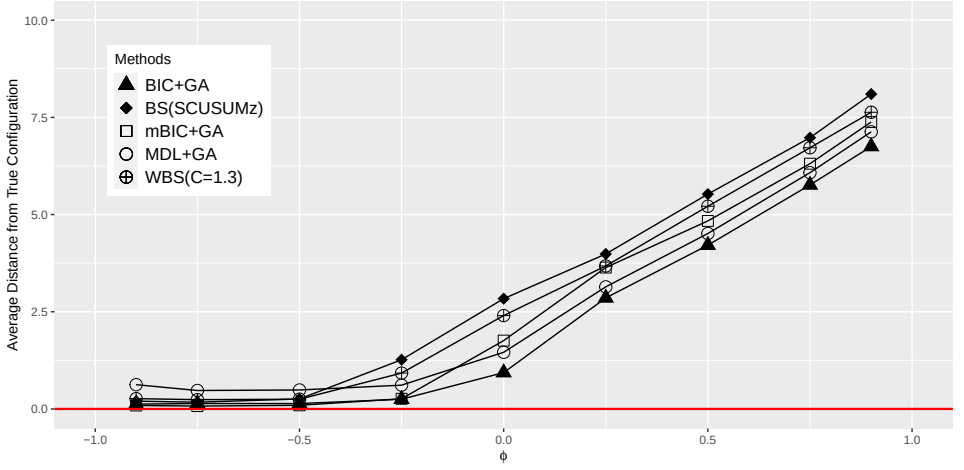

Figure 12: Average Distances for an AR(1) Series with Varying ϕ. Truth: Nine Changepoints, All Up.

are WBS and binary segmentation.

5.7. Nine Alternating Changepoints

Our next set of simulations again considers nine changepoints, but the directions of the equally spaced mean shift sizes of magnitude two are now alternated in an Up-Down-Up-Down-Up-Down-Up-Down-Up fashion (Alternating). Figure 13 displays results. The best method here is BIC again with WBS doing better than in the previous setting; mBIC is a laggard and binary segmentation is again the worst.

5.8. Nine Keyblade Changepoints

As a different type of setup, we next consider the nine changepoint setting where the sizes of the nine mean shifts vary, their shift directions vary, and the changepoint times are not equally spaced. Figure 7(d) shows our chosen pattern for $E[X_t]$, which we call a "keyblade". The distances in Figure 14 reveal BIC and MDL as winners, and WBS and binary segmentation as inferior.

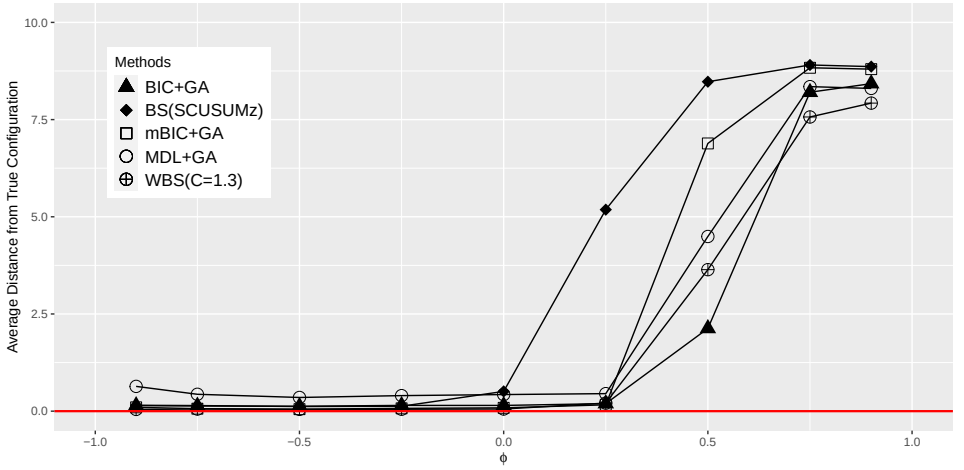

Figure 13: Average Distances for an AR(1) Series with Varying ϕ. Truth: Nine Alternating Changepoints.

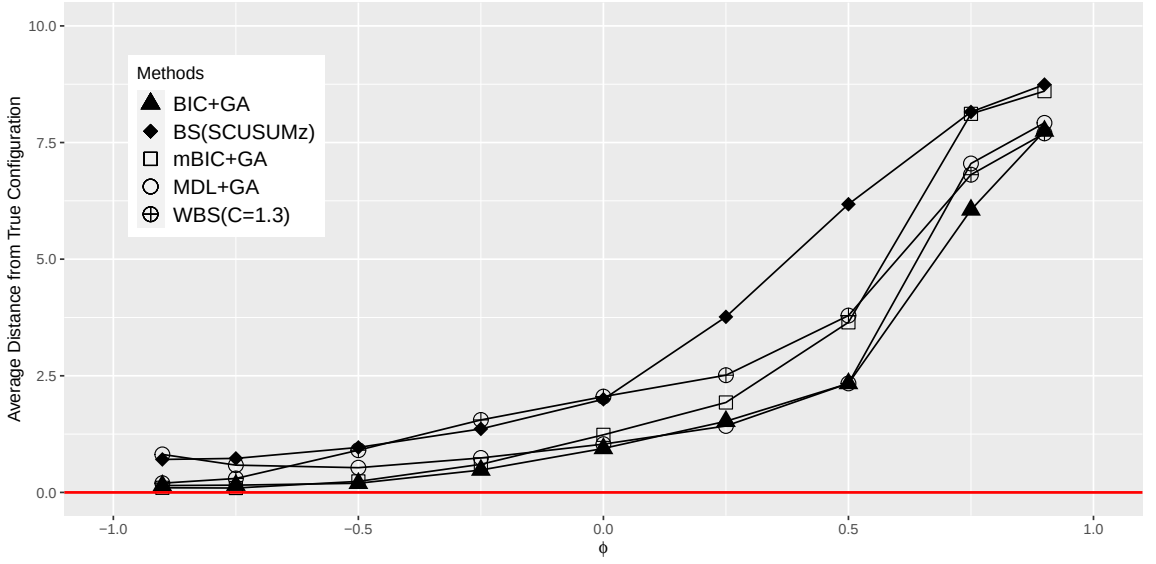

Figure 14: Average Distances for the Keyblade AR(1) Series with Varying ϕ. Truth: Nine Changepoints.

5.9. Random Changepoints

We now consider settings with a random number of changepoints simulated from a Poisson distribution with a mean of five. The locations of any mean shifts are placed uniformly in the set $\{2, \ldots, N\}$ without replacement. The mean of each segment is simulated from a normal distribution with a zero mean and a standard deviation of 1.5. Figures 15 summarizes the results: BIC and MDL are again superior and binary segmentation inferior.

5.10. Varying Series Lengths

The performance of the simple BIC penalty so far was surprising to us — especially since this penalty does not depend on the changepoint times. To investigate this issue further, we fix the AR(1) parameter at $\phi = 0.5$ and compare BIC and mBIC distances as N varies with one and three changepoints. Here, the changepoints induce equal length regimes, all mean shift sizes are of a unit magnitude, and their directions alternate with the first direction being upwards. Table 2 reports average BIC and mBIC distances when $N \in \{500, 1000, 2500\}$. As the sample size increases, the additional penalty the mBIC places on the length of the segments results in fewer changepoints identified than BIC.

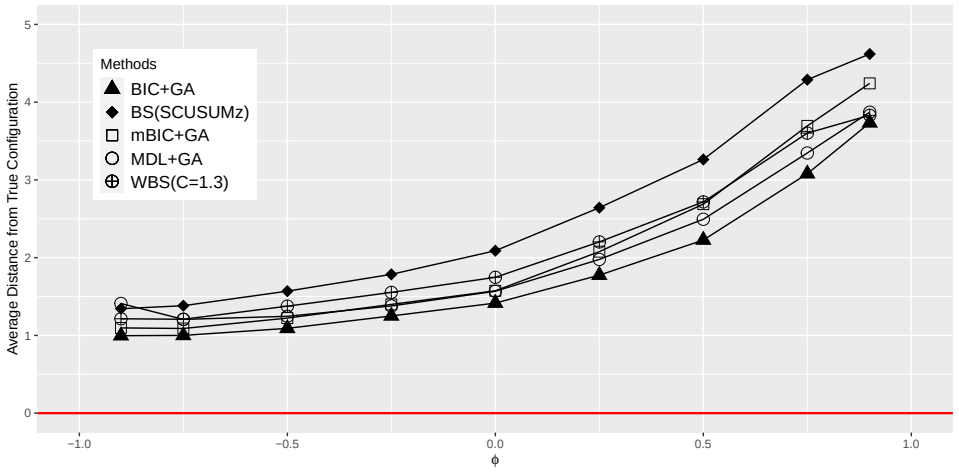

Figure 15: Average Distance between the Estimated and True Changepoint Locations.

Table 2
Comparison of BIC and mBIC. Truth: m changepoints, all of a unit magnitude, placed in alternating directions that equally space the record length for an AR(1) series with varying lengths N. Here, $\sigma^2 = 1$ and $\phi = 0.5$.

Avg. Distance	$m = 1$		$m = 3$	
	BIC	mBIC	BIC	mBIC
$N = 500$	0.227	0.125	1.270	2.420
$N = 1000$	0.126	0.066	0.311	0.921
$N = 2500$	0.121	0.047	0.123	0.066

Table 3
Average Distance for an AR(1) Series with Varying Mean Shift Magnitudes.

Δ	BIC+GA	mBIC+GA	MDL+GA	BS(SCUSUM$_Z$)	WBS(C=1.3)
$\Delta = 1$	1.269	2.424	1.948	2.702	1.686
$\Delta = 2$	0.140	0.051	0.209	0.843	0.149
$\Delta = 3$	0.126	0.042	0.188	0.077	0.079

As N grows, there is a tendency for BIC to add (erroneous) changepoints in some samples. Thus, as the number of changepoints and N grows, mBIC does tend to beat BIC. This leads us to recommend mBIC over BIC for larger N or numbers of changepoints.

Before moving to non-AR(1) settings, we examine method performance as the mean shift magnitudes increase. Here, we fix $N = 500, \phi = 0.5$, and $\sigma^2 = 1$ and consider three alternating changepoints placed at the times 126, 251, and 376. Mean shift magnitudes Δ are varied from 1 to 3. Average distances over 1,000 simulations are reported in Table 3. As the mean shift magnitudes increases, all methods improve. BIC and MDL, frequent winners of past scenarios, perform worst when the mean shift size is largest; moreover, WBS and binary segmentation, two frequent past losers, perform best. mBIC reports the smallest average distance when $\Delta \geq 2$.

Our final simulation task considers other autoregressive error structures. We begin with AR(2) errors and the case of no changepoints. Table 4 shows false positive rates of signaling one or more changepoints when in truth none exist for various AR(2) parameters ϕ_1 and ϕ_2. In this and all four tables below, 1,000 independent simulations are

Table 4
False Positive Rates for an AR(2) Series with Varying $\{\phi_1, \phi_2\}$. Truth: No Changepoints.

$\{\phi_1, \phi_2\}$	BIC+GA	mBIC+GA	MDL+GA	BS(SCUSUM$_Z$)	WBS(C=1.3)
{0.6, 0.35}	21.5%	2.5%	38.8%	22.6%	50.0%
{0.6, 0.3}	17.5%	2.6%	33.2%	10.2%	36.6%
{0.6, -0.1}	5.9%	1.1%	15.6%	0.3%	17.4%
{0.5, -0.2}	4.1%	1.6%	13.6%	0.0%	11.7%
{0.2, -0.5}	3.0%	0.6%	9.4%	0.1%	9.1%

Table 5
Average Distances for an AR(2) Series with Varying $\{\phi_1, \phi_2\}$. Truth: Three Alternating Changepoints of Size $\Delta = 2$.

$\{\phi_1, \phi_2\}$	BIC+GA	mBIC+GA	MDL+GA	BS(SCUSUM$_Z$)	WBS(C=1.3)
{0.6, 0.35}	2.757	2.932	2.759	2.633	2.265
{0.6, 0.30}	2.484	2.895	2.510	2.742	2.337
{0.6, -0.1}	0.167	0.052	0.182	0.818	0.193
{0.5, -0.2}	0.131	0.032	0.163	0.072	0.101
{0.2, -0.5}	0.086	0.023	0.111	0.040	0.068

Table 6
False Positive Rates for an AR(4) Series with Varying $\{\phi_1, \phi_2, \phi_3, \phi_4\}$. Truth: No Changepoints.

$\{\phi_1,\ \phi_2,\ \phi_3,\ \phi_4\}$	BIC+GA	mBIC+GA	MDL+GA	BS(SCUSUM$_Z$)	WBS(C=1.3)
{0.5, 0.25, 0.15, 0.05}	66.5 %	44.8%	76.4%	29.7%	54.4%
{0.6, 0.3, 0.1, -0.3}	16.7 %	8.5%	42.0%	0.6%	21.5%
{0.6, 0.3, -0.3, -0.1}	9.9%	4.9%	32.5%	0.1%	14.8%
{0.6, -0.4, -0.2, -0.1}	5.0%	2.5%	27.0%	0.2%	10.3%
{0.6, -0.4, 0.3, -0.2}	5.3%	1.6%	22.9%	0.2%	17.4%

Table 7
Average Distances for AR(4) Errors with Varying $\{\phi_1, \phi_2, \phi_3, \phi_4\}$. Truth: Three Alternating Changepoints of Size $\Delta = 2$.

$\{\phi_1,\ \phi_2,\ \phi_3,\ \phi_4\}$	BIC+GA	mBIC+GA	MDL+GA	BS(SCUSUM$_Z$)	WBS(C=1.3)
{0.5, 0.25, 0.15, 0.05}	2.723	2.420	3.360	2.516	2.151
{0.6, 0.3, 0.1, -0.3}	0.615	1.582	1.292	2.318	1.256
{0.6, 0.3, -0.3, -0.1}	0.205	0.107	0.251	0.834	0.211
{0.6, -0.4, -0.2, -0.1}	0.127	0.055	0.319	0.031	0.079
{0.6, -0.4, 0.3, -0.2}	0.161	0.066	0.246	0.228	0.101

conducted, $N = 500$, $\sigma^2 = 1$, and all mean shift sizes are two units (this adds additional information to the above unit mean shift simulations). The structure of the four tables below are discussed in tandem after their presentation.

Table 5 reports average distances for the AR(2) scenario of the last table, but now with three changepoints. The three shifts induce four equal length regimes and shift the series mean in an up-down-up manner.

Table 6 shows false positive rates of signaling one or more changepoints when in truth there are none for various parameter choices in an AR(4) series.

Finally, Table 7 reports average distances over 1,000 independent simulations for the same AR(4) scenario above. The mean shift specifications are repeated from Table 5.

In the above tables, when there are no changepoints, binary segmentation appears best and MDL and WBS worst, as was the case for AR(1) errors. In the tables with three changepoints and heavily positively correlated errors, MDL, BIC, and WBS all do comparatively well; when the correlation becomes negative, the situation reverses and mBIC and binary segmentation are best. These aspects also held for AR(1) series, although we did not remark about the negatively correlated results.

To summarize our overall conclusions on multiple changepoints, the following points emerge:

- AIC and binary segmentation are not competitive. Binary segmentation worked well only when no or few

changepoints existed and worsened when multiple mean shifts act in opposite directions. We do not recommend either of these techniques.

- Although its penalty does not depend on the changepoint times, BIC performed surprisingly well across a wide range of scenarios. However, as N gets larger, mBIC becomes superior.
- MDL was often the best performing penalized likelihood technique in heavily correlated scenarios, but does not work as well with negatively correlated series. MDL also tends to lose to mBIC when the changepoint mean shift sizes are large or when changepoints are infrequent.
- MDL and WBS techniques should be used with caution if there is a possibility that no changes are present, as they have high false positive rates.
- BIC and mBIC perform well in the low frequency changepoint settings.

We close with one more comment that is not apparent from the reported results. The MDL penalty works reasonably in a large variety of positively correlated scenarios. However, when it is wrong, it has a tendency to put changepoints times in pairs near each other. This is an attempt by the method to identify an outlier. If one imposes a minimum spacing between changepoint times to combat this, MDL will perform better.

6. Comments and Conclusions

This paper presented a systematic comparison of common single and multiple changepoint techniques in time series settings. Previous work had demonstrated how applying techniques that assume IID to data could lead to erroneous conclusions. Here, we focused on how IID methods could be modified to work in the time series setting, either by correcting the asymptotic distribution, or by modifying the test statistic.

In constructing our comprehensive approach, a summary of the major different techniques available was made in a single manuscript; hence, this paper has utility as a reference. A new multiple changepoint distance was also developed that combines the two important features of changepoint detection, identification of the correct number and location(s) of the changepoints, within a single metric.

In the single changepoint case, it was found that the best techniques apply IID methods to the time series of one-step-ahead prediction residuals. The best performing single changepoint detection method was the sum of CUSUM statistic in Bai (1993). Extreme value based asymptotic tests exhibited poor detection power.

In the multiple changepoint case, conclusions were more nebulous; however, binary segmentation and AIC are not recommended. The penalized likelihoods MDL, mBIC, and BIC all are worthy of additional study. WBS also performed reasonably and deserve additional attention, especially given its relatively recent entrance into the literature. At this point, it is still not clear whether pure algorithmic techniques can beat penalized likelihood methods. It is our view that one should use BIC penalized likelihood methods for the case of large numbers of changepoints and/or small data lengths, with mBIC recommended for smaller numbers of changepoints and/or longer lengths of data.

References

Antoch, J., Hušková, M., Prášková, Z., 1997. Effect of dependence on statistics for determination of change. Journal of Statistical Planning and Inference 60 (2), 291–310.

Aue, A., Horváth, L., 2013. Structural breaks in time series. Journal of Time Series Analysis 34 (1), 1–16.

Auger, I. E., Lawrence, C. E., 1989. Algorithms for the optimal identification of segment neighborhoods. Bulletin of Mathematical Biology 51 (1), 39–54.

Bai, J., 1993. On the partial sums of residuals in autoregressive and moving average models. Journal of Time Series Analysis 14 (3), 247–260.

Bai, J., Perron, P., 1998. Estimating and testing linear models with multiple structural changes. Econometrica 66 (1), 47–78.

Brockwell, P., Davis, R., 1991. Time Series: Theory and Methods, 2nd Edition. Springer-Verlag.

Burkard, R., Dell'Amico, M., Martello, S., 2012. Assignment Problems, Revised Reprint. Vol. 106. SIAM.

Chakar, S., Lebarbier, E., Lévy-Leduc, C., Robin, S., 2017. A robust approach for estimating change-points in the mean of an AR(1) process. Bernoulli 23 (2), 1408–1447.

Chan, H. P., Walther, G., 2013. Detection with the scan and the average likelihood ratio. Statistica Sinica 23 (1), 409–428.

Chapman, J.-L., Eckley, I. A., Killick, R., 2020. A nonparametric approach to detecting changes in variance in locally stationary time series. Environmetrics 31 (1).

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

Chochola, O., Hušková, M., Prášková, Z., Steinebach, J. G., 2013. Robust monitoring of CAPM portfolio betas. Journal of Multivariate Analysis 115, 374–395.
Csörgo, M., Horváth, L., 1997. Limit Theorems in Change-point Analysis. John Wiley & Sons.
Davis, R. A., Lee, T. C. M., Rodriguez-Yam, G. A., 2006. Structural break estimation for nonstationary time series models. Journal of the American Statistical Association 101 (473), 223–239.
Dehling, H., Fried, R., Garcia, I., Wendler, M., 2015. Change-point detection under dependence based on two-sample U-statistics. In: Asymptotic laws and methods in stochastics. Springer, pp. 195–220.
Eichinger, B., Kirch, C., 02 2018. A MOSUM procedure for the estimation of multiple random change points. Bernoulli 24 (1), 526–564.
Fryzlewicz, P., 2014. Wild binary segmentation for multiple change-point detection. The Annals of Statistics 42 (6), 2243–2281.
Gallagher, C., Killick, R., Lund, R., Shi, X., 2021. Autocovariance estimation in the presence of changepoints. arXiv preprint arXiv:2102.10669.
Gallagher, C., Lund, R., Robbins, M., 2012. Changepoint detection in daily precipitation data. Environmetrics 23 (5), 407–419.
Hajela, P., 1990. Genetic search - an approach to the nonconvex optimization problem. AIAA Journal 28 (7), 1205–1210.
Harchaoui, Z., Lévy-Leduc, C., 2010. Multiple change-point estimation with a total variation penalty. Journal of the American Statistical Association 105 (492), 1480–1493.
Hušková, M., 2013. Robust change point analysis. In: Robustness and Complex Data Structures. Springer, pp. 171–190.
Hušková, M., Kirch, C., 2008. Bootstrapping confidence intervals for the change-point of time series. Journal of Time Series Analysis 29 (6), 947–972.
Hušková, M., Kirch, C., 2012. Bootstrapping sequential change-point tests for linear regression. Metrika 75 (5), 673–708.
Hušková, M., Marušiaková, M., 2012. M-procedures for detection of changes for dependent observations. Communications in Statistics - Simulation and Computation 41 (7), 1032–1050.
Hušková, M., Meintanis, S. G., 2006. Change point analysis based on empirical characteristic functions. Metrika 63 (2), 145–168.
Jandhyala, V., Fotopoulos, S., MacNeill, I., Liu, P., 2013. Inference for single and multiple change-points in time series. Journal of Time Series Analysis 34 (4), 423–446.
Killick, R., Fearnhead, P., Eckley, I. A., 2012. Optimal detection of changepoints with a linear computational cost. Journal of the American Statistical Association 107 (500), 1590–1598.
Kirch, C., 2007. Resampling in the frequency domain of time series to determine critical values for change-point tests. Statistics & Risk Modeling 25 (3), 1–25.
Kirch, C., 2008. Bootstrapping sequential change-point tests. Sequential Analysis 27 (3), 330–349.
Lavielle, M., Moulines, E., 2000. Least-squares estimation of an unknown number of shifts in a time series. Journal of Time Series Analysis 21 (1), 33–59.
Li, S., Lund, R., 2012. Multiple changepoint detection via genetic algorithms. Journal of Climate 25 (2), 674–686.
Lu, Q., Lund, R. B., 2007. Simple linear regression with multiple level shifts. Canadian Journal of Statistics 35 (3), 447–458.
Lund, R., Shi, X., 2020. Short communication: Detecting possibly frequent change-points: wild binary segmentation 2 and steepest-drop model selection. Journal of the Korean Statistical Society 49 (4), 1090–1095.
Lund, R., Wang, X. L., Lu, Q. Q., Reeves, J., Gallagher, C., Feng, Y., 2007. Changepoint detection in periodic and autocorrelated time series. Journal of Climate 20 (20), 5178–5190.
MacNeill, I. B., 1974. Tests for change of parameter at unknown times and distributions of some related functionals on Brownian motion. The Annals of Statistics, 950–962.
Maidstone, R., Hocking, T., Rigaill, G., Fearnhead, P., 2017. On optimal multiple changepoint algorithms for large data. Statistics and Computing 27 (2), 519–533.
Matteson, D. S., James, N. A., 2014. A nonparametric approach for multiple change point analysis of multivariate data. Journal of the American Statistical Association 109 (505), 334–345.
Menne, M. J., Williams, Claude N., J., Vose, R. S., 07 2009. The U.S. Historical Climatology Network Monthly Temperature Data, Version 2. Bulletin of the American Meteorological Society 90 (7), 993–1008.
Olshen, A. B., Venkatraman, E., Lucito, R., Wigler, M., 2004. Circular binary segmentation for the analysis of array-based dna copy number data. Biostatistics 5 (4), 557–572.
Page, E., 1955. A test for a change in a parameter occurring at an unknown point. Biometrika 42 (3/4), 523–527.
Picard, D., 1985. Testing and estimating change-points in time series. Advances in Applied Probability 17, 841–867.
Prášková, Z., Chochola, O., 2014. M-procedures for detection of a change under weak dependence. Journal of Statistical Planning and Inference 149, 60–76.
Robbins, M., Gallagher, C., Lund, R., Aue, A., 2011. Mean shift testing in correlated data. Journal of Time Series Analysis 32 (5), 498–511.
Robbins, M. W., Gallagher, C. M., Lund, R. B., 2016. A general regression changepoint test for time series data. Journal of the American Statistical Association 111 (514), 670–683.
Scott, A. J., Knott, M., 1974. A cluster analysis method for grouping means in the analysis of variance. Biometrics, 507–512.
Scrucca, L., 2013. GA: a package for genetic algorithms in R. Journal of Statistical Software 53 (4), 1–37.
Shen, J., Gallagher, C. M., Lu, Q., 2014. Detection of multiple undocumented change-points using adaptive lasso. Journal of Applied Statistics 41 (6), 1161–1173.
Stoica, P., Moses, R. L., 2005. Spectral Analysis of Signals. Pearson Prentice Hall.
Tolmatz, L., 2002. On the distribution of the square integral of the Brownian bridge. The Annals of Probability 30 (1), 253–269.
Yao, Y.-C., 1988. Estimating the number of change-points via schwarz'criterion. Statistics & Probability Letters 6 (3), 181–189.
Zhang, N. R., Siegmund, D. O., 2007. A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. Biometrics 63 (1), 22–32.