Autocovariance Estimation in the Presence of Changepoints

Colin Gallagher[1†], Rebecca Killick[2†], Robert Lund[3†]
and Xueheng Shi[3*†]

[1]School of Mathematics and Statistics, Clemson University, 105
Sikes Hall, Clemson, 29634, SC, USA.
[2]Department of Mathematics and Statistics, Lancaster
University, Bailrigg, Lancaster, LA1 4YF, England, UK.
[3*]Department of Statistics, University of California, Santa Cruz,
1156 High St, Santa Cruz, 95064, CA, USA.

*Corresponding author(s). E-mail(s): xhshi@ucdavis.edu;
Contributing authors: cgallag@clemson.edu ;
r.killick@lancaster.ac.uk; rolund@ucsc.edu;
[†]These authors contributed equally to this work.

Abstract

This article studies estimation of a stationary autocovariance structure in
the presence of an unknown number of mean shifts. Here, a Yule-Walker
moment estimator for the autoregressive parameters in a dependent time
series contaminated by mean shift changepoints is proposed and stud-
ied. The estimator is based on first order differences of the series and
is proven consistent and asymptotically normal when the number of
changepoints m and the series length N satisfy $m/N \to 0$ as $N \to \infty$.

Keywords: Autoregression, Differencing, Robustness, Rolling Windows,
Segmentation, Yule-Walker Estimates

1 Introduction

Time series dynamics often change due to external events or internal sys-
tematic fluctuations. One common structural change is the mean shift, and

changepoint analyses allow the researcher to identify whether and when abrupt changes in the mean of the series take place. Evolving from the original treatise for a single location parameter shift in [1], the majority (but not all) of changepoint analyses check for shifts in the mean of the series. Since [1], considerable changepoint work has been conducted, including recursive segmentation algorithms such as binary segmentation and wild binary segmentation [2], dynamic programming based approaches such as [3] and [4], moving sum (MOSUM) procedures [5, 6], and simultaneous multi-scale changepoint estimators (SMUCE) [7]. Additional changepoint work includes applications in climatatology [8], economics [9], and disease modelling [10].

Many changepoint techniques assume independent and identically distributed (IID) model errors; however, time series data are typically correlated (e.g., daily temperatures, stock prices, and DNA sequences [11]). Changepoint techniques tend to overestimate the number of changepoints should positive autocorrelation be ignored [12]. In addition, some multiple changepoint models for time series allow all model parameters, including those governing the correlation structure of the series, to change at each changepoint time. These scenarios are easier to handle computationally as dynamic programming techniques can quickly optimize penalized likelihood objective functions; see [4] and [13]. In these cases, the objective function optimized is additive in its segments (regimes). A more parsimonious model allows series means to shift with each changepoint time, but keeps error autocovariances constant across all regimes. These models do not lead to objective function additivity and fast dynamic programming techniques cannot be directly applied (See [12]).

Remedies typically seek to incorporate the autocorrelation structure in the changepoint analysis or to pre-whiten the series prior to any changepoint analysis. In either case, one needs to quantify the autocovariance structure and/or long-run variance of the series. With a good estimate of the series' autocovariance structure, one-step-ahead prediction residuals can be computed — and these residuals are always uncorrelated (independent up to estimation error for Gaussian series). Indeed, a principle of [12, 14] is that good multiple changepoint detection routines can be devised by applying IID methods to the series' one-step-ahead prediction residuals (also called pre-whitening). Perhaps owing to this, considerable recent research has sought to find changepoints in dependent time series. Among these, [15] estimate the long-run variance of the error process via a difference-type variance estimator calculated from local means from different blocks; this estimate is then used to modify SMUCE for dependent data. The authors [6] propose a robust covariance estimation procedure from $M-$estimation to modify a moving sum procedure. Other proposed long-run variance (or time-average variance) estimators for mean shift problems based on robust methods include [16], [17], and [18].

This paper studies autocovariance and long-run variance estimation in the presence of mean shifts in more detail. We devise a method based on first order differencing that outperforms robust and rolling window methods. The

scenario is asymptotically quantified when the model errors obey a causal autoregressive (AR) process.

The rest of this paper proceeds as follows. The next section narrates our setup and discusses approaches to the problem. Section 3 then develops an estimation technique based on lag one differences of the series. Section 4 proves consistency and asymptotic normality of these estimators and Section 5 assesses their performance in simulations. Section 6 applies the results to an annual precipitation series and Section 7 concludes with brief comments.

2 Model and Estimation Approaches

Suppose that $\{X_t\}_{t=1}^N$ is a time series having an unknown number of mean shift changepoints, denoted by m, occurring at the unknown ordered times $1 < \tau_1 < \tau_2 < \cdots < \tau_m \leq N$. These m changepoints partition the series into $m+1$ distinct segments, each segment having its own mean. The model is written as

$$X_t = \kappa_{s(t)} + \epsilon_t. \tag{1}$$

Here, $s(t)$ denotes the series' regime number at time t, which takes values in $\{0, 1, \ldots, m\}$. Then $\kappa_{s(t)} = \mu_i$ is constant for all times in the ith regime:

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

We assume that $\{\epsilon_t\}$ is a stationary causal AR(p) time series that applies to all regimes. The AR order p is assumed known for the moment; BIC penalties will be examined later to select the order of the autoregression should it be unknown. While more general ARMA(p, q) $\{\epsilon_t\}$ could be considered, we work with AR(p) errors because this model class is dense in all stationary short-memory series [19], and estimation, prediction, and forecasting are easily conducted. Adding a moving-average component $q \geq 1$ induces considerably more work and is less commonly found in changepoint applications. The AR(p) $\{\epsilon_t\}$ obeys

$$\epsilon_t = \phi_1 \epsilon_{t-1} + \cdots + \phi_p \epsilon_{t-p} + Z_t, \quad t \in \mathbb{Z}, \tag{2}$$

where $\{Z_t\}$ is IID white noise with a zero mean, variance σ^2, and a finite fourth moment (this enables consistent estimation of the autoregressive parameters $\phi_1, \ldots, \phi_p$).

The next section develops a difference based moment estimation procedure for the mean shift setting. Under this scenario, first-order differences of the series will have a non-zero mean only at the changepoint times. At this point, it might seem prudent to apply ARMA estimation methods that are

robust to outliers to the differenced series. Indeed, many previous authors have considered outlier-robust estimators for ARMA models. For examples, the M-estimators of [20] are shown to be consistent and tractable and the bounded influence propagation (BIP) τ-estimators in [21] merit mention. However, these estimators require the ARMA series to be causal and invertible. In our application, the differenced series has a unit root in its MA component and is hence not invertible. Perhaps worse, future simulations demonstrate that BIP τ-estimators do not perform well in our setting.

3 Moment Estimates based on Differencing

This section derives a system of linear equations that relate the autocorrelations of the differenced series to the AR(p) coefficients. First-order differencing a series eliminates any piecewise constant mean except at times where shifts occur. Authors have previously used differencing to estimate global parameters in the changepoint literature. For example, [22] discuss a class of difference-based estimators for autocovariances in nonparametric changepoint segment regression when the errors are from a stationary m-dependent process. The paper [2] uses differencing to get an estimate of $\mathrm{Var}(X_t)$, although IID errors are assumed in this work. The estimator in (16) below comes from [18] and is also based on differencing. This said, there seems to be no previous literature using differencing to estimate AR(p) parameters in a setting corrupted by mean shifts. As an aside, differencing also detrends a time series; the estimators below perform well if a time series has both changepoints and a linear trend.

Let $\{X_t\}$ be a stationary series satisfying the causal AR(p) difference equation

$$X_t = \mu + \sum_{j=1}^{p} \phi_j(X_{t-j} - \mu) + Z_t, \tag{3}$$

with $\{Z_t\}$ a zero mean IID sequence with a finite fourth moment. Since X_t may be causally expressed in terms of $Z_t, Z_{t-1}, \ldots$, the autoregressive coefficients are uniquely determined by the pth order recursion

$$\gamma_X(h) = \phi_1\gamma_X(h-1) + \cdots + \phi_p\gamma_X(h-p), \quad h = 1, 2, \ldots \tag{4}$$

and its boundary conditions [19]. Here, $\gamma_X(h) = \mathrm{Cov}(X_t, X_{t-h})$ and we use the analogous notation $\rho_X(h) = \mathrm{Corr}(X_t, X_{t-h})$. Consider the sequence of first differences defined by $d_t = X_t - X_{t-1}$. Then $\{d_t\}$ is stationary with

$$\gamma_d(h) := \mathrm{Cov}(d_t, d_{t+h}) = 2\gamma_X(h) - \gamma_X(h-1) - \gamma_X(h+1), \tag{5}$$

and

$$\rho_d(h) := \mathrm{Corr}(d_t, d_{t+h}) = \frac{2\rho_X(h) - \rho_X(h-1) - \rho_X(h+1)}{2[1 - \rho_X(1)]}. \tag{6}$$

One can also show that $\{d_t\}$ satisfies an ARMA$(p,1)$ difference equation with a first-order moving average parameter of -1. We now show that $\phi_1, \ldots, \phi_p$ can be recovered from the autocorrelation function of the differences.

Let $P(A\|B)$ denote the best linear predictor (BLP) of a random variable A from linear combinations of elements in the set B. We assume that B includes a constant term to allow for cases with a nonzero mean. It is well known that for a stationary causal ARMA process, the linear representation of the best linear prediction of future series values from past series values is unique [19]. Equations determining the autoregressive coefficients can be derived by equating two different expressions for the BLP.

Executing on the above, (3) gives

$$P(d_{p+1}\|1;\ X_j, 1 \le j \le p) = P(X_{p+1}\|1;\ X_j, 1 \le j \le p) - X_p$$
$$= \kappa\mu + (\phi_1 - 1)X_p + \sum_{j=2}^{p} \phi_j X_{p+1-j},$$

where $\kappa = 1 - \sum_{j=1}^{p} \phi_j$ ($\kappa \ne 0$ by causality). Substituting $X_{p-j} = X_p - \sum_{j=0}^{p-2} d_{p-j}$ for $j = 1, \ldots, p-1$ into the last line above yields

$$P(d_{p+1}\|1;\ X_j, 1 \le j \le p) = \kappa(\mu - X_p) - \sum_{j=0}^{p-2} d_{p-j} \sum_{i=2+j}^{p} \phi_i.$$

To express the BLP in terms of $\mathbf{d} = (d_p, \ldots, d_1)^T$ only, use the prediction equations to obtain

$$P(X_p\|1;\ \mathbf{d}) = \mu + \mathrm{Cov}(X_p, \mathbf{d})\boldsymbol{\Gamma}_d^{-1}\mathbf{d},$$

where $\boldsymbol{\Gamma}_d$ is the $p \times p$ covariance matrix of $\mathbf{d}$, which is known to be invertible for a causal stationary ARMA $\{d_t\}$ (see Proposition 5.1 in [19]). Combining the above gives

$$P(d_{p+1}\|1;\ \mathbf{d}) = -\kappa v_p d_1 + \sum_{j=1}^{p-1} \left(d_{p+1-j} - \kappa v_j - \sum_{i=j+1}^{p} \phi_i \right), \qquad (7)$$

with $(v_1, v_2, \ldots, v_p) = \mathrm{Cov}(X_p, \mathbf{d})\boldsymbol{\Gamma}_d^{-1}$.

The coefficients $\mathbf{v}^T = (v_1, \ldots, v_p)$ can be written in terms of the correlation function of the differences in (6):

$$\mathbf{v}^T = \gamma_d(0)^{-1}\mathrm{Cov}(X_p, \mathbf{d})(\boldsymbol{\Gamma}_d/\gamma_d(0))^{-1}$$
$$= \mathbf{cR}_d^{-1},$$

where $\mathbf{R}_d$ is the $p \times p$ autocorrelation matrix of $\mathbf{d}$ and

$$\mathbf{c}^T = \gamma_d(0)^{-1}\mathrm{Cov}(X_p, \mathbf{d}) = \left(1/2, 1/2 + \rho_d(1), \ldots, 1/2 + \sum_{k=1}^{p-1}\rho_d(k)\right) \quad (8)$$

can be extracted from (5) and the relation

$$\gamma_X(h) - \gamma_X(h+1) = \frac{\gamma_d(0)}{2} + \sum_{k=1}^{h-1}\gamma_d(k).$$

A second representation of the BLP is given by the prediction equations:

$$P(d_{p+1}\|1; \mathbf{d}) = \sum_{j=1}^{p} u_j d_{p+1-j}, \quad (9)$$

where the predicting coefficients are $(u_1, u_2, \ldots, u_p) = \mathrm{Corr}(d_{p+1}, \mathbf{d})\mathbf{R}_d^{-1}$. Here, u_p is the lag p partial autocorrelation of $\{d_t\}$. Equating the coefficient of d_1 in (7) and (9) yields $-\kappa v_p = u_p$. If $v_p \neq 0$, which we tacitly assume to avoid trifling work, we can set $\kappa = -u_p/v_p$. Equating the coefficients on the right hand sides of (7) and (9), and solving for ϕ produces an expression of the autoregressive coefficients in terms of the autocorrelations of $\{d_t\}$:

$$\phi_k = (u_k - u_{k-1}) - \frac{u_p}{v_p}(v_k - v_{k-1}), \quad k = 1, \ldots, p, \quad (10)$$

where $v_0 = 1$ and $u_0 = -1$. If $\{X_t\}$ satisfies (3), then $\phi_1, \ldots, \phi_p$ satisfy (10). Now let $\{d_t\}$ be a stationary sequence with $v_p \neq 0$ and suppose that $\phi^T = (\phi_1, \ldots, \phi_p)$ satisfies (10):

$$\phi = \mathbf{A}\mathbf{R}_d^{-1}\left(\mathrm{Corr}(d_{p+1}, \mathbf{d}) - \frac{u_p}{v_p}\mathbf{c}\right) + \left(1 + \frac{u_p}{v_p}, 0, \ldots, 0\right)^T,$$

with

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -1 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix}.$$

Since u_p is the partial correlation of $\{d_t\}$ at lag p,

$$u_p = (0, 0, \ldots, 1)\mathbf{R}_d^{-1}\boldsymbol{\rho}_d,$$

with $\boldsymbol{\rho}_d^T = (\rho_d(1), \ldots, \rho_d(h))$. Substituting this into the above linear equation for $\boldsymbol{\phi}$ and simplifying gives

$$\boldsymbol{\phi} = \boldsymbol{M}\boldsymbol{\rho}_d + (1, 0, \ldots, 0)^T, \tag{11}$$

where

$$\mathbf{M} = \mathbf{AR}_d^{-1}\left(\mathbf{I} - \mathbf{c}^*(0, 0, \ldots, 1)\mathbf{R}_d^{-1}\right),$$

with $\mathbf{c}^* = \left(-1/2, 1/2 + \rho_d(1), \ldots, 1/2 + \sum_{k=1}^{p-1}\rho_d(k)\right)^T$. Note that each element in $\boldsymbol{M}$ is a function of $\rho_d(1), \ldots, \rho_d(p)$.

The $p = 1$ case will shed light on the above calculations. Here, (10) and (6) give

$$\phi_1 = 1 + 2\rho_d(1) = \frac{\rho_X(1) - \rho_X(2)}{1 - \rho_X(1)},$$

which exceeds unity whenever $\rho_X(2) < 2\rho_X(1) - 1$, which can happen for some $AR(p)$ models. However, if $\{X_t\}$ follows

$$X_t = \phi X_{t-1} + Z_t, \quad t = 0, \pm 1, \pm 2, \ldots,$$

and $|\phi| < 1$,

$$\frac{\rho_X(1) - \rho_X(2)}{1 - \rho_X(1)} = \frac{\phi - \phi^2}{1 - \phi} = \phi.$$

In general, if $\boldsymbol{\phi}$ is from a causal $AR(p)$ model satisfying (4), then (11) provides a one-to-one transformation between $\rho_d(1), \ldots, \rho_d(p)$ and $\boldsymbol{\phi}$. However, if $\{X_t\}$ does not follow a causal $AR(p)$ recursion, there is no guarantee that $\boldsymbol{\phi}$ satisfying (10) corresponds to a causal $AR(p)$ model. In practice, this presents no issue since it is easy to check to see if a fitted $AR(p)$ model is causal. If our fitted model is not causal, this is an indication that $\{X_t\}$ is inadequately described by an $AR(p)$ series. In this case, we simply change p and refit until causality is achieved.

Given observations $X_1, \ldots, X_N$, we estimate the lag h sample autocorrelation of the differences from

$$\hat{\rho}_d(h) = \frac{\hat{\gamma}_d(h)}{\hat{\gamma}_d(0)} = \frac{\sum_{t=1}^{N-h}(d_t - \bar{d})(d_{t+h} - \bar{d})}{\sum_{t=1}^{N}(d_t - \bar{d})^2}, \quad h \geq 0.$$

Here, $\bar{d} = (N - 1)^{-1}\sum_{t=1}^{N-1}d_t$ is the sample mean. The $AR(p)$ model will be fit using (10) with $\gamma_d(h)$ replaced by the sample version $\hat{\gamma}_d(h)$:

$$\hat{\mathbf{u}} = \hat{\mathbf{R}}_d^{-1}\hat{\boldsymbol{\rho}}_d \quad \text{and} \quad \hat{\mathbf{v}} = \hat{\mathbf{R}}_d^{-1}\hat{\mathbf{c}}, \tag{12}$$

where $\hat{\mathbf{c}}$ is obtained from (8) by replacing all elements with their estimates.

To ensure that the estimated v_p is not zero, one simply checks this in practice. It is also recommended to check to see if the fitted $AR(p)$ model is causal.

Summarizing, our proposed algorithm for fitting an AR(p) model using differences is

1. Compute $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$. If $\hat{v}_p = 0$, reduce the AR order to $p-1$ and refit.
2. Use (10) with $\mathbf{u} = \hat{\mathbf{u}}$ and $\mathbf{v} = \hat{\mathbf{v}}$ to find $\hat{\boldsymbol{\phi}}$, and check to see that the estimates correspond to a causal model. If the solution is non-causal, change p and refit.

The above algorithm produces a $\hat{\boldsymbol{\phi}}$ for a causal AR(p) process satisfying (11):

$$\hat{\boldsymbol{\phi}} = \hat{\mathbf{M}}\hat{\boldsymbol{\rho}}_d + (1, 0, \ldots, 0)^T, \tag{13}$$

where each element in $\hat{\mathbf{M}}$ corresponds to an element of $\mathbf{M}$ with $\rho_d(h)$ replaced by $\hat{\rho}_d(h)$ for each h. For any stationary sequence of first differences $\{d_t\}$, each element of $\hat{\mathbf{M}}$ converges almost surely to its theoretical value. In particular, as $N \to \infty$, $\hat{\mathbf{M}} \to \mathbf{M}$ in the almost sure sense.

We end this section by estimating σ^2. There are several moment equations that can be used to estimate σ^2. For example, multiplying both sides of the ARMA($p, 1$) difference equation,

$$d_t = \phi_1 d_{t-1} + \cdots + \phi_p d_{t-p} + z_t - z_{t-1},$$

by d_t, taking expectations, and solving for σ^2 yields,

$$\sigma^2 = \gamma_d(0) \left(\frac{1 - \sum_{k=1}^{p} \phi_k \rho_d(k)}{2 - \phi_1} \right).$$

A moment based estimator of the variance is hence

$$\hat{\sigma}^2 = \hat{\gamma}_d(0) \left(\frac{1 - \sum_{k=1}^{p} \hat{\phi}_k \hat{\rho}_d(k)}{2 - \hat{\phi}_1} \right). \tag{14}$$

In the next section, we show that $\hat{\sigma}^2$ is a $\sqrt{N}$-consistent estimator of σ^2.

4 Asymptotic Normality

This section shows that if $m = m(N)$ grows slowly enough in N, the estimators in the last section will be consistent and asymptotically normal. If the number of changepoints m is small relative to N, then the mean shifts should have a negligible impact on the estimated autocovariance of the differences, since $X_t - X_{t-1} = d_t - d_{t-1}$ except at the changepoint times $\tau_1, \ldots, \tau_m$. In particular, to obtain asymptotic normality, we assume that as $N \to \infty$, for some finite B,

- [A.1] $\max_{0 \leq k \leq m(N)} | \mu_{k+1} - \mu_k | \leq B$.
- [A.2] $m(N) = o(\sqrt{N})$.

Condition A.1 imposes existence of some bound on the mean shift sizes and Condition A.2 regulates the number of changepoints that can occur.

We begin with asymptotic normality of the autocorrelations for first-order differences in the general ARMA(p,q) case, which may be of distinct interest. The asymptotic normality of the AR(p) estimators is a corollary to Theorem 1.

Theorem 1 *If $\{X_t\}_{t=1}^N$ obeys (1) with $\{\epsilon_t\}$ satisfying (2) where $\{Z_t\}$ is IID white noise having a finite fourth moment, then for each fixed positive integer k, as $N \to \infty$,*

$$\sqrt{N} \begin{pmatrix} \hat{\rho}_d(1) - \rho_d(1) \\ \vdots \\ \hat{\rho}_d(k) - \rho_d(k) \end{pmatrix} \xrightarrow{\mathbf{D}} \mathbf{N}_k(\mathbf{0}, \mathbf{BWB}^T).$$

Here, the elements in the $(k+1) \times (k+1)$ dimensional $\mathbf{W}$ are from Bartlett's formula for the asymptotic covariance matrix of $(\hat{\rho}_\epsilon(1), \ldots, \hat{\rho}_\epsilon(k+1))^T$, (see Chapter 8 of [19]) and $\mathbf{B}$ is $k \times (k+1)$ dimensional with form

$$\mathbf{B} = \frac{1}{2(1 - \rho_X(1))} \begin{bmatrix} 2 & -1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -1 & 2 & -1 \end{bmatrix}.$$

Proof. We first show that the changepoints have negligible impact on estimated autocorrelations in the limit. To do this, write $d_t = X_t - X_{t-1} = (\epsilon_t - \epsilon_{t-1}) + \delta_t$, with $\delta_t = (\mu_k - \mu_{k-1})I_{[t=\tau_{k+1}]}$, and I_A the indicator of the set A. Letting

$$\tilde{\gamma}_d(h) = \frac{\sum_{t=2}^{N-h}(\epsilon_t - \epsilon_{t-1})(\epsilon_{t+h} - \epsilon_{t+h-1})}{N},$$

then

$$\sqrt{N}\,|\,\hat{\gamma}_d(h) - \tilde{\gamma}_d(h)\,| \le \frac{m}{\sqrt{N}} \left[m^{-1} \sum_{t \in \mathcal{T}} \delta_t(\epsilon_{t+h} - \epsilon_{t+h-1} + \epsilon_{t-h} - \epsilon_{t-h-1} + \delta_t) \right],$$

where $\mathcal{T} = \{\tau_1, \ldots, \tau_m\}$ denote all changepoint times. The term on the right hand side converges to zero if $N^{-1/2}m \to 0$ as $N \to \infty$ (this is Condition A.2) and the sum is bounded in probability (this is guaranteed by Conditions A.1, A.2, and the properties of $\{\epsilon_t\}$). We see that the asymptotic distribution of $\hat{\gamma}_d(h)$ is the same as that of $\tilde{\gamma}_d(h)$.

It is easy to see that

$$\sqrt{N}\,[\tilde{\gamma}_d(h) - (2\hat{\gamma}_\epsilon(h) - \hat{\gamma}_\epsilon(h-1) - \hat{\gamma}_\epsilon(h+1))] = o_P(1), \tag{15}$$

where $o_P(1)$ denotes a term that converges to zero in probability as $N \to \infty$. Using the above and $\hat{\gamma}_d(0)/\gamma_\epsilon(0) \to 2(1 - \rho_\epsilon(1))$ in the almost sure sense, we have

$$\sqrt{N}\,(\hat{\rho}_d(h) - \rho_d(h)) = \frac{\sqrt{N}}{2(1 - \rho_\epsilon(1))}(-1, 2, -1)^T \begin{bmatrix} \hat{\gamma}_\epsilon(h-1) - \gamma_\epsilon(h-1) \\ \hat{\gamma}_\epsilon(h) - \gamma_\epsilon(h) \\ \hat{\gamma}_\epsilon(h+1) - \gamma_\epsilon(h+1) \end{bmatrix} + o_p(1)$$

for each $h = 1, \ldots, k$. Hence,

$$\sqrt{N}(\hat{\boldsymbol{\rho}}_d - \boldsymbol{\rho}_d) = \mathbf{B}\sqrt{N}(\hat{\boldsymbol{\rho}}_\epsilon - \boldsymbol{\rho}_\epsilon) + o_p(1).$$

Theorem 1 now follows from classic results for asymptotic normality for sample autocorrelations of ARMA processes (see Chapter 8 of [19]).

Corollary 2 *Suppose that $\{X_t\}$ follows (1) with $\{\epsilon_t\}$ satisfying (2) with $\{Z_t\}$ IID white noise having a finite fourth moment. For the estimator in (11), as $N \to \infty$,*

$$\sqrt{N} \begin{pmatrix} \hat{\phi}_1 - \phi_1 \\ \vdots \\ \hat{\phi}_p - \phi_p \end{pmatrix} \xrightarrow{\mathbf{D}} \mathbf{N}_p(\mathbf{0}, \boldsymbol{\Sigma}),$$

Here, $\boldsymbol{\Sigma} = \mathbf{MBW}(\mathbf{MB})^{\mathbf{T}}$.

Proof of Corollary 2. Since $\{d_t\}$ is stationary and ergodic, the elements of $\hat{\boldsymbol{M}}$ converge to those in $\boldsymbol{M}$ in the almost sure sense; specifically, (15) gives

$$\sqrt{N} \left(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi} \right) = \mathbf{M}\sqrt{N} \left(\hat{\boldsymbol{\rho}}_d - \boldsymbol{\rho}_d \right) + o_P(1).$$

The conclusion of Corollary 2 now follows.

Theorem 1 and Corollary 2 imply that $\hat{\boldsymbol{\rho}}_d$ and $\hat{\boldsymbol{\phi}}$ are both consistent estimators, so that $\hat{\sigma}^2$ given by (14) is a consistent estimator of the white noise variance.

5 A Simulation Study

A simulation study with $AR(p)$ errors is now conducted. Our Yule-Walker moment estimator based on first-order differencing is now compared to several estimators, including the robust $AR(1)$ estimator of [18], the BIP τ-estimators of [21], and the rolling window methods employed in [23].

The paper [18] studies the $AR(1)$ case and proposes an estimator that is robust to mean shifts:

$$\hat{\phi} := \frac{\left(\underset{1 \le t \le N-2}{\text{median}} \mid X_{t+2} - X_t \mid \right)^2}{\left(\underset{1 \le t \le N-1}{\text{median}} \mid X_{t+1} - X_t \mid \right)^2} - 1. \tag{16}$$

It is not clear how to extend this work to cases where $p > 1$.

The rolling window methods of [23] estimate autocorrelations via window-based methods as follows. For a window length w, with $w \le N$, a moving window scheme generates $N - w + 1$ subsegments, the ith subsegment containing the data at times $i, \ldots, i + w - 1$. Each subsegment is treated as a

stationary series (even though some may contain mean shifts and are thus truly nonstationary) and the time series parameters are estimated in subsegment i from the data in this subsegment only. The final estimates are taken as medians of the estimates over all subsegments. The hope is that most windows will be "changepoint free", and medians over all subsegments will not be heavily influenced by the few windows containing changepoints. Of course, such a scheme may not use all data efficiently in estimation. Moreover, [23] demonstrates that the success of this procedure depends heavily on the choice of w. As we show below, these robust autocovariance estimation methods do not perform particularly well for this problem.

In each simulation, the series length is $N = 1,000$ and m is randomly generated from the discrete uniform distribution Uniform$\{0, 1, \ldots, 10\}$, which roughly corresponds to the changepoint frequency in our data example in the next section. All changepoint times are generated randomly within $\{2, 3, \ldots, N\}$ with equal probability — we do not impose any minimal spacing between successive changepoint times. The segment means μ_i are randomly generated from a Uniform$(-1.5, 1.5)$ distribution. Ten thousand independent runs are conducted for all cases.

We first consider AR(1) errors, simulating ϕ randomly from the Uniform$(-0.95, 0.95)$ distribution and $\{Z_t\}$ as Gaussian white noise with a unit variance. Our Yule-Walker difference estimator in (12) is denoted by `Diff` in future figures. This estimator will be compared to a variety of alternative approaches. The robust AR(1) estimator in (16) is denoted by `AR1seg`. Averaged rolling window estimators, using different window lengths, are denoted by their lengths: `N`, `N/2`, `N/5`, `N/10`, `N/20`, and `N/50`. We also compare to the general ARMA robust estimator of [21] applied to the differenced data, which is denoted by `BIP`. Here, we fit a general ARMA(1,1) model for the errors, which does not take into account that the MA(1) parameter should be -1. This extra flexibility should make the `BIP` method appear better than it truly is. Finally, we include an estimator based on our approach but with the outlying observations in $\{d_t\}$ first removed, which we denote by `Outlier`. Since our method is "corrupted" by non-zero means at the changepoint times, removing outliers (which are likely to occur at the changepoint observations) should improve our approach. For outlier detection, we use a simple nonparametric Tukey fence and acknowledge that other detection schemes could be used.

Our simulation results are summarized in Figure 1. The obvious winner is the Yule-Walker estimator based on first-order differencing. Indeed, this estimator is unbiased and has the smallest variance. The `AR1seg` estimator is unbiased; however, it has a larger variability than the Yule-Walker difference estimators. The performance of the rolling window estimators depends on the choice of the window length, but appears to be inferior to the difference based estimator, even with the optimal window size selected (which is likely somewhere between `N/20` and `N/50` in this simulation). It is hard to decide the optimal window length in practice and smaller window lengths considerably increase computation time. While the general `BIP` robust estimator appears

unbiased, it has a much larger variance than all other estimators. Indeed, this estimator seems to be the worst of all. Our outlier removal approach has a slightly positive bias and slightly larger variance, likely induced by the tendency to remove true observations as outliers.
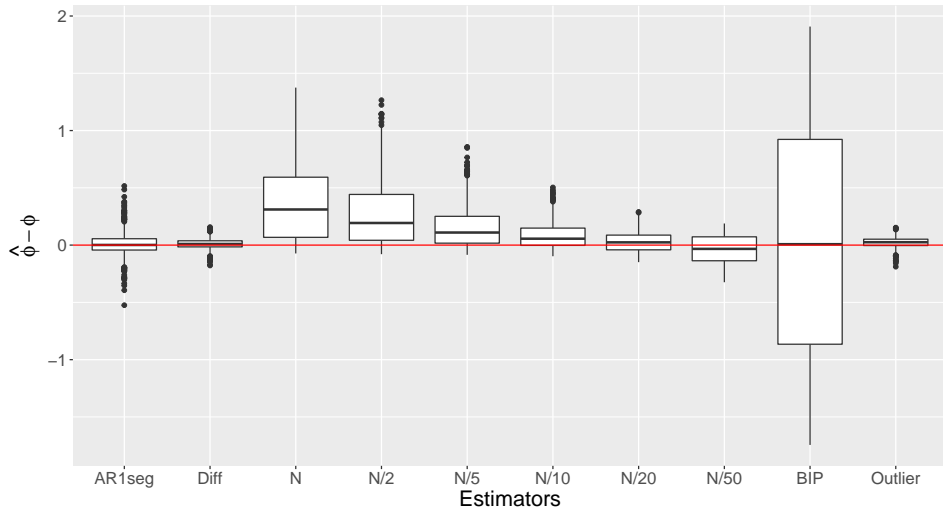

Fig. 1 Box plots of estimates of the AR(1) parameter ϕ. Our differenced based method appears to be unbiased and has the smallest variability; the BIP robust method has the largest variability of all methods.

We now move to AR(2) errors. In each AR(2) simulation, the AR coefficients were uniformly generated from the triangular region guaranteeing model causality: $\phi_1 + \phi_2 < 1$, $\phi_2 - \phi_1 < 1$, and $|\phi_2| < 1$. In these simulations, the changepoint total is fixed at $m = 9$ and all segments have equal lengths. All mean shifts alternate in sign with an absolute magnitude of 2.0, the first shift moving the series upwards. The series length varies with $N \in \{1000, 2000, 5000, 10000, 20000\}$. Since $p > 1$, the `AR1seg` estimator is not applicable. The rolling-window estimator and general `BIP` robust estimator were dropped from consideration due to their poor AR(1) performance and computational time requirements. The simulation results show that estimator bias and variance decreases as the length of the series increases, reinforcing the consistency results in the last section.

Moving to AR(4) simulations, to meet model causality requirements, the AR(4) characteristic polynomial is factored into its four roots, denoted by $1/r_1, 1/r_2, 1/r_3$, and $1/r_4$. That is,

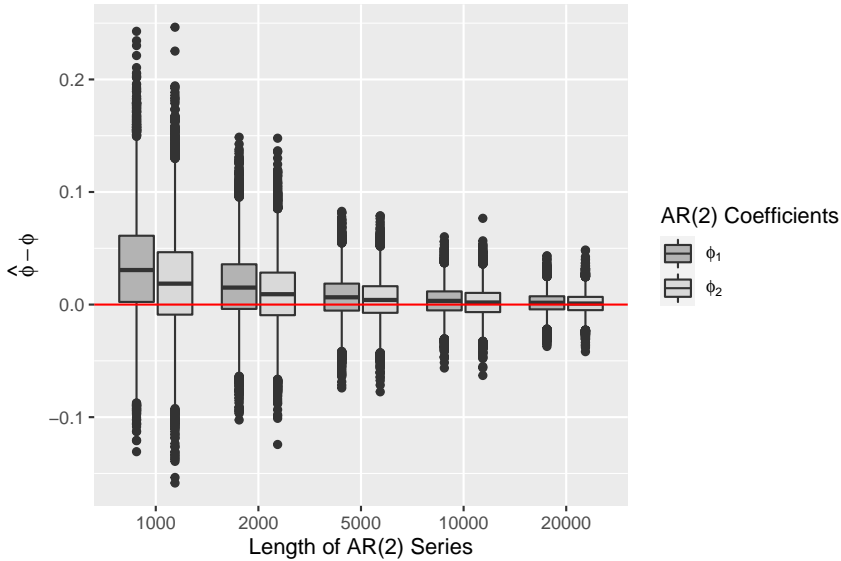$$\phi(z) = (1 - r_1 z)(1 - r_2 z)(1 - r_3 z)(1 - r_4 z).$$

Fig. 2 Box plots of AR(2) coefficient estimates. Variability and bias decrease as the series length increases.

Causality implies that all r_i should lie inside the complex unit circle. To meet this, r_1 and r_2 will be randomly generated from the Uniform$(-0.9, 0.9)$ distribution, and r_3 is a randomly generated complex number with modulus $|r_3| < 0.9$. The root r_4 is taken as the complex conjugate of r_3. This mixes real and complex roots in the AR(4) characteristic polynomial. All other simulation settings are identical to those in the AR(2) case. Figure 3 shows our results, which exhibit the same pattern as the AR(2) case, with decreasing bias and variance as N increases.

Our next simulation returns to the AR(1) setting and conducts a sensitivity analysis to mean shift sizes. Here, estimator accuracy is more greatly influenced by the magnitude of the mean shifts than changepoint locations. We take all mean shifts to have the same size Δ and introduce the signal-to-noise ratio (SNR), defined as the absolute mean shift magnitude over the marginal series standard deviation of X_t:

$$\text{SNR} = \frac{|\Delta|}{\sqrt{\frac{\sigma^2}{1-\phi^2}}}. \tag{17}$$

For simplicity, σ^2 is set to unity. The number of changepoints is fixed at $m = 9$ and their locations are randomly generated over $\{2, \ldots, N\}$ with $N = 1,000$. In each run, the true ϕ is simulated from the Uniform$(-0.95, 0.95)$ distribution. The nine mean shifts alternate signs, with $|\Delta|$ varied in $[0,5]$. Boxplots of the difference between the estimated $\hat{\phi}$ and the true ϕ are presented in Figure 4.

The horizontal line in Figure 4 marks zero bias in $\hat{\phi}$; the solid curve depicts the average differences between $\hat{\phi}$ and ϕ. Obviously, the larger the mean shift magnitude, the more our estimator degrades. This said, in practice, larger
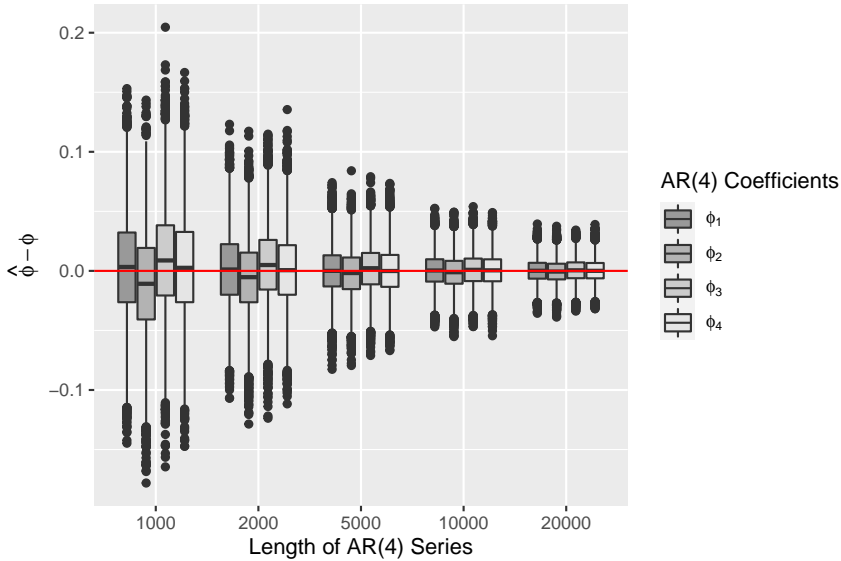
Fig. 3 Box plots of AR(4) coefficient estimates. Again, estimator variabilities and biases decrease with increasing sample size.

mean shift sizes can usually be identified as outliers in the differenced series [24, 25] or can easily be identified in the original series, despite the AR contamination. As such, the essential challenge lies with estimating the AR(p) parameters in the presence of smaller mean shifts.

Two more simulations are included. Our first simulation shows how AR(p) processes can approximate MA(q) errors in changepoint problems. The specifications of the series and changepoints are the same as the first AR(1) case of this section, but the model errors obey the MA(1) model

$$\epsilon_t = Z_t + \theta Z_{t-1},$$

with $\theta = 0.5$. The plot in Figure 5 shows the autocorrelation function of our fitted AR(10) process from one simulation run only. Notice that the fitted AR(10) autocovariance is essentially zero at most lags that exceed unity, indicating the overall quality of the AR(10) approximation (an MA(1) model is characterized by an autocovariance that is non-zero only at lag 1).

Our final simulation considers order selection of p for AR errors by adding the Bayesian Information Criterion (BIC) penalty $(p+1)\ln(N)$ to minus two times the log likelihood of the model. The mean shifts in $\{X_t\}$ "contaminate" the likelihood for $\{X_t\}$ away from a likelihood for an AR series with a fixed (constant) mean. Our remedy here is to demean $\{X_t\}$ before estimating p. While other methods of order estimation are possible, this procedure worked the best amongst several that were experimented with. More specifically, the AR(p) coefficients are first estimated via $\{d_t\}$ for each candidate AR order
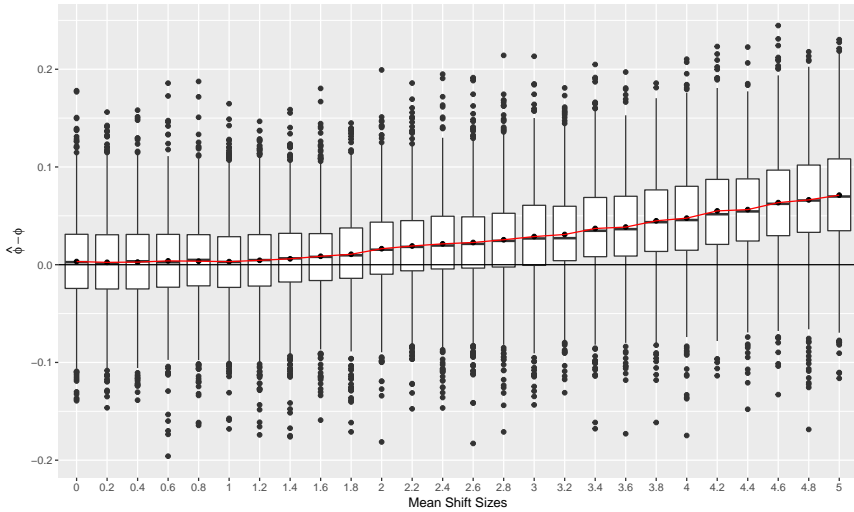
Fig. 4 An AR(1) mean shift size sensitivity plot. The larger the mean shifts are, the more the estimates of ϕ degrade.
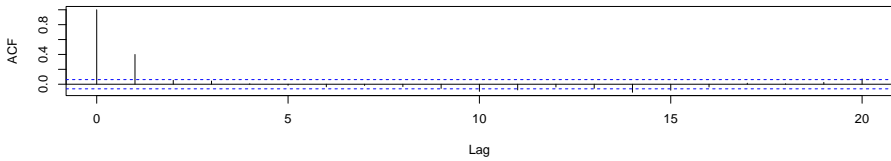

Fig. 5 Autocorrelations of the AR(10) process used to approximate our MA(1) errors. The dashed lines demarcate 95% pointwise confidence thresholds for white noise.

$p \in \{1, 2, \ldots, p_{\max}\}$, where $p_{\max}$ is some preset maximum AR order to consider. Then, one-step-ahead prediction residuals were computed and the changepoint configuration was estimated by some changepoint technique. The pruned exact liner time (PELT) algorithm of [4] was used here. The estimated changepoint configuration is then used to demean $\{X_t\}$. The likelihood and BIC scores are then calculated from the demeaned series for each order p and the order with the smallest penalized likelihood BIC score is selected.

$$\text{BIC}(p) = -2\log(\hat{\sigma}^2) + (p+1)\log(N) \tag{18}$$

In our simulation, $N = 1,000$, nine equally-spaced mean shifts of size 2.5 corrupt the series, and the errors are generated from a causal AR(4) process with coefficients $\boldsymbol{\phi} = (0.3, -0.3, -0.2, -0.1)$. The estimated AR order for $1,000$ simulations is plotted in the Figure 6 histogram. While BIC selects $p = 4$ the majority of the time, it is also prone to overestimation, selecting the order 5

in more than 20% of the runs. AR order overestimation by BIC is classically appreciated in even changepoint-free settings [19].
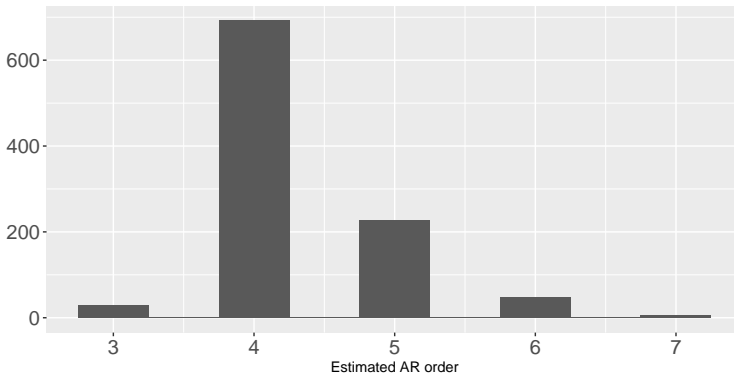

Fig. 6 Estimated AR orders via a Bayesian information criterion penalty. The mode of the histogram is correct at four, but some overestimation of p is also present.

6 Applications

6.1 Changepoints in AR(p) Series

As previously discussed, most changepoint techniques mistakenly flag changepoints when underlying positive dependence is ignored. For example, [26] argues that shifts identified in the London house price series of [27] may be more attributable to the positive correlations in the series than to actual mean shifts. CUSUM based techniques are known to degrade with positive correlation [12]. To remedy this, authors recommend detecting changepoints from estimated versions of the one-step-ahead prediction residuals of the series [14, 28]. This requires estimation of the autocovariance structure of the series in the presence of the unknown changepoints. As such, a major application of our methods serves to decorrelate (pre-whiten) series without any prior knowledge of the changepoint configuration of the series. IID-based changepoint techniques, applied to the estimated one-step-ahead prediction residuals, can then be used to estimate any mean shifts in the series. The Yule-Walker difference estimator proposed here is extensively used in [12] to do just this. In addition, our difference estimator supplies a long-run variance estimate needed in the changepoint methods in [5], [17], and [29].

Table 1 demonstrates the improved performance of two popular multiple changepoint methods, wild binary segmentation (WBS) [2] and PELT [4]. In each run, an AR(1) series of length $N = 500$ is simulated with ϕ fixed within $\{0.25, 0.50, 0.75\}$, and $\sigma^2 = 1$. The series has either no changepoints or three equally spaced changepoints; all mean shift sizes are the same, are denoted by Δ, and are chosen to induce the constant signal-to-noise requirement of

SNR $= 2$ in (17). All simulations are aggregated from $1,000$ independent runs. In Table 1, $\overline{\hat{m}}$ and $SE_{\hat{m}}$ denote the average and standard error of the estimated number of changepoints when WBS and PELT are directly applied to the series. The quantities $\overline{\hat{m}^d}$ and $SE_{\hat{m}^d}$ denote the average and standard error of the estimated number of changepoints from the one-step-ahead prediction residuals after fitting an AR(1) series to the differences by our methods.

	WBS		PELT	
	$\phi = 0.25$			
# of Changepoints	$\overline{\hat{m}}/(SE_{\hat{m}})$	$\overline{\hat{m}^d}/(SE_{\hat{m}^d})$	$\overline{\hat{m}}/(SE_{\hat{m}})$	$\overline{\hat{m}^d}/(SE_{\hat{m}^d})$
Zero	3.85/(2.43)	0.17/(0.54)	0.02/(0.17)	0.00/(0.02)
Three	5.46/(2.01)	3.03/(0.18)	3.07/(0.34)	3.00/(0.03)
	$\phi = 0.5$			
	WBS		PELT	
# of Changepoints	$\overline{\hat{m}}/(SE_{\hat{m}})$	$\overline{\hat{m}^d}/(SE_{\hat{m}^d})$	$\overline{\hat{m}}/(SE_{\hat{m}})$	$\overline{\hat{m}^d}/(SE_{\hat{m}^d})$
Zero	16.03/(3.92)	0.24/(0.70)	1.36/(1.62)	0.00/(0.04)
Three	16.90/(3.81)	3.09/(0.40)	4.28/(1.43)	2.95/(0.37)
	$\phi = 0.75$			
	WBS		PELT	
# of Changepoints	$\overline{\hat{m}}/(SE_{\hat{m}})$	$\overline{\hat{m}^d}/(SE_{\hat{m}^d})$	$\overline{\hat{m}}/(SE_{\hat{m}})$	$\overline{\hat{m}^d}/(SE_{\hat{m}^d})$
Zero	30.77/(3.98)	0.40/(0.96)	12.65/(3.31)	0.01/(0.13)
Three	31.45/(3.86)	2.48/(1.38)	14.33/(3.25)	1.59/(1.44)

Table 1 Results for AR(1) series with three and no changepoints. The thresholding constant used in WBS is $C = 1.3$. Estimation of the changepoint number drastically improves when autocorrelation is taken into account.

It is apparent that IID based WBS and PELT methods overestimate the number of changepoints in a dependent series when positive correlation is ignored; PELT appears to be more resistant to dependence issues than WBS. In contrast, with the help of the proposed Yule-Walker difference estimator and decorrelation techniques, both WBS and PELT become much more accurate.

6.2 New Bedford Precipitation

Annual precipitations from New Bedford and Boston, Massachusetts are studies in [30]. The data are available from https://w2.weather.gov/climate/xmacis.php?wfo=box. The ratio of these series (New Bedford to Boston) is displayed in Figure 7, along with a fitted mean of a model that allows for both multiple changepoints and AR errors. Three documented changepoints, occurring at the years 1886, 1917, and 1967 are indicated. After adjusting for four regime means, Figure 8 shows the sample ACF plot of the precipitation ratio series, suggesting that the series is correlated. The Bayesian Information Criterion estimates $p = 1$ as the AR order. Although this order does not seem to adequately describe all non-zero autocorrelations, we use it anyway to illustrate our points.

The AR(1) parameter estimates fluctuate wildly over distinct methods. Specifically, our difference Yule-Walker estimator and BIP τ-estimators produce antipodal estimates as can be seen in Table 2. Our estimate agrees
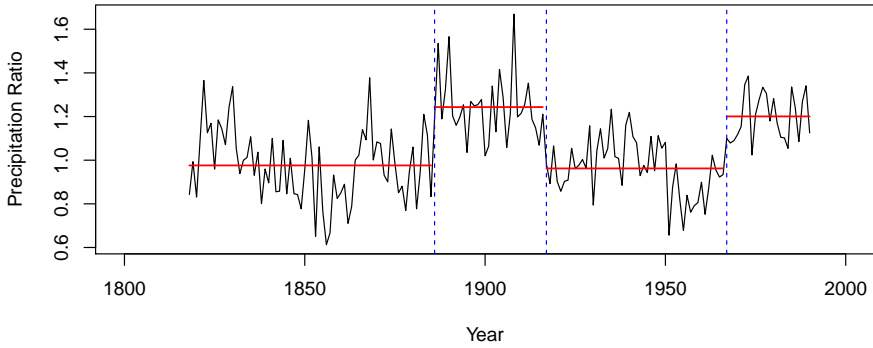
Fig. 7 New Bedford to Boston annual precipitation ratios with three identified change-points.
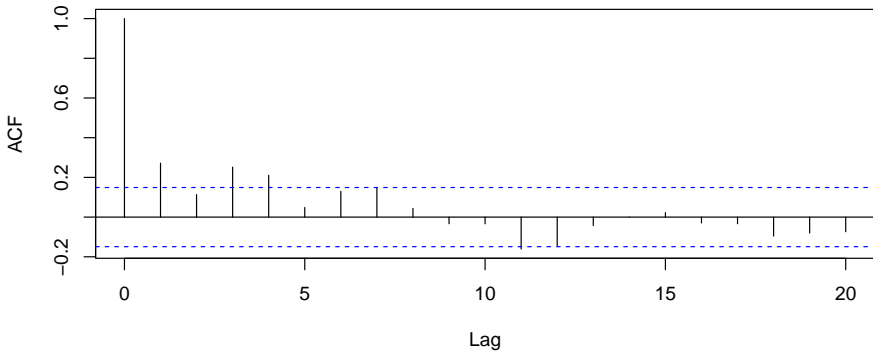

Fig. 8 Sample autocorrelations of the demeaned precipitation ratio series with 95% pointwise confidence bands for zero correlation.

closely with an estimate computed by assuming the three changepoint times are known, but the level of autocorrelation is significantly less than that estimated in a Yule-Walker scheme that ignores all three changepoint times. The results show that one needs to be careful in changepoint problems with correlated data — mean shifts and correlation can inject similar features into time series.

Methods	Estimate $\hat{\phi}$
Yule Walker Estimator (ignoring changepoints)	0.547
BIP τ-estimator	0.990
Yule Walker Estimator (known changepoint times)	0.273
AR1seg	-0.268
Difference Yule-Walker Estimator	0.255

Table 2 $\overline{\text{AR}}(1)$ ϕ estimates for the precipitation ratio series. The individual estimates highly depend on the method.

7 Conclusions

Differencing methods can effectively be used to estimate the autocovariance structure of an AR(p) series corrupted by mean shift changepoints. Our Yule-Walker estimator for autoregressive models is easy to implement, computationally fast, consistent, and asymptotically normal. While the proposed estimator is adversely impacted by large mean shifts, large shifts appear as large outliers in the differenced series and can be removed. When changepoints are present, the difference methods developed here significantly improve changepoint techniques developed for IID errors. The techniques are also applicable if the series has a linear trend (constant across all regimes) with intercept shifts.

References

[1] Page, E.S.: Continuous inspection schemes. Biometrika **41**(1-2), 100–115 (1954)

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

[3] Jackson, B., Scargle, J.D., Barnes, D., Arabhi, S., Alt, A., Gioumousis, P., Gwin, E., Sangtrakulcharoen, P., Tan, L., Tsai, T.T.: An algorithm for optimal partitioning of data on an interval. IEEE Signal Processing Letters **12**(2), 105–108 (2005)

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

[5] Eichinger, B., Kirch, C.: A MOSUM procedure for the estimation of multiple random change points. Bernoulli **24**(1), 526–564 (2018)

[6] Chen, L., Wang, W., Wu, W.B.: Inference of breakpoints in high-dimensional time series. Journal of the American Statistical Association (just-accepted), 1–33 (2021)

[7] Frick, K., Munk, A., Sieling, H.: Multiscale change point inference. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **76**(3), 495–580 (2014)

[8] Hewaarachchi, A.P., Li, Y., Lund, R., Rennie, J.: Homogenization of daily temperature data. Journal of Climate **30**(3), 985–999 (2017)

[9] Norwood, B., Killick, R.: Long memory and changepoint models: A spectral classification procedure. Statistics & Computing **28**(2), 291–302 (2018)

[10] Hall, C.B., Lipton, R.B., Sliwinski, M., Stewart, W.F.: A change point model for estimating the onset of cognitive decline in preclinical alzheimer's disease. Statistics in Medicine **19**(11-12), 1555–1566 (2000)

[11] Chakravarthy, N., Spanias, A., Iasemidis, L.D., Tsakalis, K.: Autoregressive modeling and feature analysis of DNA sequences. EURASIP Journal on Advances in Signal Processing **2004**(1), 1–16 (2004)

[12] Shi, X., Gallagher, C., Lund, R., Killick, R.: A comparison of single and multiple changepoint techniques for time series data. Computational Statistics and Data Analysis **170** (2022). https://doi.org/10.1016/j.csda.2022.107433.

[13] Maidstone, R., Hocking, T., Rigaill, G., Fearnhead, P.: On optimal multiple changepoint algorithms for large data. Statistics and Computing **27**(2), 519–533 (2017)

[14] Robbins, M., Gallagher, C., Lund, R., Aue, A.: Mean shift testing in correlated data. Journal of Time Series Analysis **32**(5), 498–511 (2011)

[15] Dette, H., Eckle, T., Vetter, M.: Multiscale change point detection for dependent data. Scandinavian Journal of Statistics **47**(4), 1243–1274 (2020)

[16] Chan, K.W.: Mean-structure and autocorrelation consistent covariance matrix estimation. Journal of Business & Economic Statistics **40**(1), 201–215 (2022). https://doi.org/10.1080/07350015.2020.1796397

[17] Romano, G., Rigaill, G., Runge, V., Fearnhead, P.: Detecting abrupt

changes in the presence of local fluctuations and autocorrelated noise. Journal of the American Statistical Association (To appear) (2021)

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

[19] Brockwell, P.J., Davis, R.A.: Time Series: Theory and Methods, 2nd edn. Springer, New York City (1991)

[20] Muler, N., Pena, D., Yohai, V.J.: Robust estimation for ARMA models. The Annals of Statistics **37**(2), 816–840 (2009)

[21] Muma, M., Zoubir, A.M.: Bounded Influence Propagation τ-Estimation: A New Robust Method for ARMA Model Estimation. IEEE Transactions on Signal Processing **65**(7), 1712–1727 (2017)

[22] Tecuapetla-Gómez, I., Munk, A.: Autocovariance estimation in regression with a discontinuous signal and M-dependent errors: A difference-based approach. Scandinavian Journal of Statistics **44**(2), 346–368 (2017)

[23] Beaulieu, C., Killick, R.: Distinguishing trends and shifts from memory in climate data. Journal of Climate **31**, 9519–9543 (2018)

[24] Chen, C., Liu, L.-M.: Joint estimation of model parameters and outlier effects in time series. Journal of the American Statistical Association **88**(421), 284–297 (1993)

[25] McQuarrie, A.D., Tsai, C.-L.: Outlier detections in autoregressive models. Journal of Computational and Graphical Statistics **12**(2), 450–471 (2003)

[26] Lund, R., Shi, X.: Commentary on: Detecting possibly frequent changepoints: Wild binary segmentation 2 and steepest-drop model selection. Journal of the Korean Statistical Society **49**, 1090–1095 (2020)

[27] Fryzlewicz, P.: Detecting possibly frequent change-points: Wild binary segmentation 2 and steepest-drop model selection. Journal of the Korean Statistical Society **49**, 1027–1070 (2020)

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

[29] Dette, H., Eckle, T., Vetter, M.: Multiscale change point detection for dependent data. Scandinavian Journal of Statistics **47**(4), 1243–1274 (2020)

[30] Li, S., Lund, R.: Multiple changepoint detection via genetic algorithms. Journal of Climate **25**(2), 674–686 (2012)