Original Article

Journal of Time Series Analysis

First revision received January 2009

Published online in Wiley Online Library: 24 January 2011

(wileyonlinelibrary.com) DOI: 10.1111/j.1467-9892.2010.00707.x

Mean shift testing in correlated data

Michael Robbins[a,*,†], Colin Gallagher[b], Robert Lund[b] and Alexander Aue[c]

Several tests for detecting mean shifts at an unknown time in stationary time series have been proposed, including cumulative sum (CUSUM), Gaussian likelihood ratio (LR), maximum of $F(F_{\max})$ and extreme value statistics. This article reviews these tests, connects them with theoretical results, and compares their finite sample performance via simulation. We propose an adjusted CUSUM statistic which is closely related to the LR test and which links all tests. We find that tests based on CUSUMing estimated one-step-ahead prediction residuals from a fitted autoregressive moving average perform well in general and that the LR and $F_{\max}$ tests (which induce substantial computational complexities) offer only a slight increase in power over the adjusted CUSUM test. We also conclude that CUSUM procedures work slightly better when the changepoint time is located near the centre of the data, but the adjusted CUSUM methods are preferable when the changepoint lies closer to the beginning or end of the data record. Finally, an application is presented to demonstrate the importance of the choice of method.

Keywords: ARMA series; Brownian bridge; changepoints; CUSUM; likelihood ratio.
JEL classification: C22.

1. INTRODUCTION

Identifying and locating structural breaks (changepoints) is a common problem confronting the time-series analyst. For example, mean shifts in temperature series frequently occur when the temperature gauge in a weather station is changed (see Reeves *et al.*, 2007 for a climate overview). Neglecting changepoints can produce radically misleading trend estimates or incorrect short/long memory inferences about the autocovariance structure of the series. Identifying changepoint times in time-series data is hence an important problem.

Changepoints can occur in the mean, variance and/or quantiles of a time series. Outside of speech and financial series, the most commonly encountered problem lies with the detection of an undocumented mean shift. This article summarizes, connects and compares three mean shift test statistics for stationary correlated data: cumulative sum (CUSUM), likelihood ratio (LR) and maximum of F ($F_{\max}$) test statistics. We isolate to the at most one changepoint setting for simplicity; however, the methodology we discuss can be used effectively in the presence of multiple changepoints (Robbins *et al.*, 2011). See Davis *et al.* (2006) and Lu *et al.* (2010) for recent work on multiple changepoint problems with autocorrelated data.

The general model considered here involves a time series $\{X_t\}$ with a stationary autocovariance structure – say $\gamma(h) = \mathrm{Cov}(X_{t+h}, X_t)$ at lag h – with a possible mean shift at an unknown time c. Indexing the observed data from 1 to n, we write

$$X_t = \begin{cases} \mu + \epsilon_t, & \text{for } 1 \leqslant t \leqslant c, \\ \mu + \Delta + \epsilon_t, & \text{for } c < t \leqslant n, \end{cases} \tag{1}$$

where μ is unknown, Δ is the magnitude of the mean shift at the unknown time c, and $\{\epsilon_t\}$ is a zero mean stationary series with autocovariance $\gamma(h)$ at lag h. We wish to test the null hypothesis $H_0 : \Delta = 0$ against the alternative $H_A : \Delta \neq 0$.

The literature on changepoint problems is by now vast. Page (1954, 1955) is widely credited with introducing unknown changepoint problems. Quandt (1958, 1960) extended the setting to linear models and first suggested LR approaches to the problem. Yao and Davis (1984) established asymptotic properties of a LR test for a mean shift in i.i.d. normal data. Gombay and Horváth (1990) quantified the asymptotics of LR changepoint test statistics for i.i.d. data via extreme value asymptotics and convergence to functionals of Brownian motion. MacNeill (1974) discussed the CUSUM-type statistic fundamental to this article and noted the convergence of this statistic to a Brownian bridge in the i.i.d. case. Csörgő and Horváth (1988) extend this convergence to the nonparametric i.i.d. setting and discuss the scaled CUSUM statistic that we will link to LR tests. Donsker's invariance principle (Billingsley, 1995) is the foundation for all results involving convergence to Brownian motion in the i.i.d. case. Under dependence assumptions, results involving CUSUMs of observations (similar to results in the i.i.d. case) are discussed in Antoch *et al.* (1997) and Berkes *et al.* (2009). Davis *et al.* (1995) established convergence of a LR statistic for autoregressive models when all parameters are allowed to change at the changepoint time. Brown *et al.* (1975) introduced statistics based on CUSUMs of residuals in linear models and Bai (1993) and Yu (2007) extended these ideas to residuals of autoregressive moving-average (ARMA) processes. A comprehensive reference on large sample changepoint testing is Csörgő and Horváth (1997). $F_{\max}$ tests are popular

[a]National Institute of Statistical Sciences
[b]Clemson University
[c]University of California, Davis
*Correspondence to: Michael Robbins, National Institute of Statistical Sciences, Research Triangle Park, NC 27709-4006, USA.
†E-mail: robbins@niss.org

changepoint tests in climatology (Lund and Reeves, 2002; Reeves *et al.*, 2007 and the references therein). In many settings, $F_{\max}$ and LR tests are asymptotically equivalent. For example, see Theorem 5 given below.

In this article, we examine the CUSUM, LR and $F_{\max}$ tests for detecting mean shifts in correlated data. The intent is to give practical advice on which methods work best with correlated data. Adjusted CUSUM methods for one-step-ahead prediction residuals are developed and used to link these three tests. We show that CUSUMs of raw data and CUSUMs of ARMA residuals have the same asymptotic behavior (see Theorem 2 below). In Section 3, LR and $F_{\max}$ tests for autoregressive models are connected to the adjusted CUSUM test (see Theorem 5). This enables asymptotic quantification of the LR and $F_{\max}$ test statistics. Throughout, the finite sample performance of the tests is assessed via simulation. The simulations indicate that the adjusted CUSUM statistic applied to one-step-ahead prediction residuals in general offers the most reliable tests. Moreover, we find that

1. For ARMA data, changepoint tests based on CUSUMs of prediction residuals can have significantly better finite sample performance than those based on the observed series.
2. CUSUM tests have slightly higher power than the adjusted CUSUM (and the LR and $F_{\max}$) tests when the changepoint is near the centre of the data record. Otherwise, adjusted CUSUM tests may have significantly higher power.
3. LR and $F_{\max}$ tests where parameters are estimated under the alternative hypothesis introduce substantial computational complexities, but typically only provide a very small increase in power over the adjusted CUSUM test.

The differences in the methods are highlighted in analyses of fish recruitment and southern oscillation index (SOI) series in Section 5. There, we show how changepoint tests based on the raw data and on the residuals can lead to opposite conclusions; that is, the choice of methods matters.

2. CUSUMS OF CORRELATED DATA

If the mean shift in truth occurred at time k, then a statistic comparing $\bar{X}_k = k^{-1}\sum_{t=1}^{k} X_t$ to $\bar{X}_k^* = (n-k)^{-1}\sum_{t=k+1}^{n} X_t$ should reveal differences. The CUSUM statistic,

$$\mathrm{CUSUM}_X(k) = \frac{1}{\sqrt{n}}\left(\sum_{t=1}^{k} X_t - \frac{k}{n}\sum_{t=1}^{n} X_t\right) = \frac{k}{n}\left(1-\frac{k}{n}\right)\left[\sqrt{n}(\bar{X}_k - \bar{X}_k^*)\right],$$

is a scaled difference of these sample means weighting for the respective number of observations. The magnitude of the maximum absolute discrepancy is

$$\max_{1\leqslant k\leqslant n} |\mathrm{CUSUM}_X(k)|,$$

which provides a test statistic, and the (smallest) argument that maximizes $|\mathrm{CUSUM}_X(k)|$ is taken as the estimate of the changepoint time.

To quantify distributional characteristics of $\max_{1\leqslant k\leqslant n}|\mathrm{CUSUM}_X(k)|$, we need additional assumptions on the model in eqn (1). Let $\mathbb{Z}$ denote the set of integers. We impose the causal linear process representation:

$$\epsilon_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}, \quad t \in \mathbb{Z}, \tag{2}$$

on $\{\epsilon_t\}$. Here, $\{Z_t, t \in \mathbb{Z}\}$ is a sequence of i.i.d. random variables (each with zero mean and finite variance σ^2) and the weights $\{\psi_j\}$ satisfy $\sum_{j=1}^{\infty} j|\psi_j| < \infty$.

The asymptotics of $\max_{1\leqslant k\leqslant n}|\mathrm{CUSUM}_X(k)|$ are explained in detail in Csörgő and Horváth (1997). Let $\{B(t), t \in [0,1]\}$ denote a standard Brownian bridge. Define

$$\tau^2 = \lim_{n\to\infty} \frac{1}{n}\mathrm{Var}\left(\sum_{t=1}^{n} \epsilon_t\right) \tag{3}$$

and note that $\tau^2 = \sigma^2$ if $\{\epsilon_t\}$ is i.i.d. Observe that $\tau^2 = 2\pi f(0)$ where

$$f(\eta) = \frac{1}{2\pi}\sum_{h=-\infty}^{\infty} \mathrm{e}^{ih\eta}\gamma(h) = \frac{\sigma^2}{2\pi}|\psi(\mathrm{e}^{-i\eta})|^2, \quad -\pi \leqslant \eta < \pi, \tag{4}$$

is the spectral density at frequency $\eta \in [-\pi, \pi)$. Here, $\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j$ (see, e.g., section 4.4 of Brockwell and Davis 1991).

Theorem 1. *If $\{X_t\}$ satisfies eqns (1) and (2) and H_0 holds, then*

$$\frac{1}{\hat{\tau}}\max_{1\leqslant k\leqslant n} |\mathrm{CUSUM}_X(k)| \xrightarrow{\mathcal{D}} \sup_{0\leqslant t\leqslant 1} |B(t)|, \tag{5}$$

where $\hat{\tau}^2$ is a consistent estimator of τ^2 and $\xrightarrow{\mathcal{D}}$ indicates convergence in distribution.

The above theorem follows from the work of Csörgő and Horváth (1997) (it is essentially Donsker's theorem) and the fact that $\hat{\tau}/\tau \xrightarrow{\mathcal{P}} 1$, where $\xrightarrow{\mathcal{P}}$ indicates convergence in probability. A popular estimator of τ^2 is the nonparametric Bartlett-based expression:

$$\hat{\tau}_1^2 = \frac{1}{n}\sum_{t=1}^{n}(X_t - \bar{X})^2 + 2\sum_{s=1}^{q_n}\left(1 - \frac{s}{q_n+1}\right)\frac{1}{n-s}\sum_{t=1}^{n-s}(X_t - \bar{X})(X_{t+s} - \bar{X}). \tag{6}$$

To ensure consistency of $\hat{\tau}_1^2$, one typically imposes growth conditions on the bandwidth parameter q_n to allow for a slow divergence to infinity as $n \to \infty$. Here, we take $q_n = \lfloor n^{1/3} \rfloor$, which seems to work well in a variety of practical situations and has technical justification (Newey and West, 1987; Andrews, 1991).

Null hypothesis percentiles of the asymptotic distribution are needed to assess statistical significance of the CUSUM statistic. Such percentiles are computed via the series expansion:

$$P(\sup_{0 \leqslant t \leqslant 1} |B(t)| > x) = 2\sum_{k=1}^{\infty}(-1)^{k+1}\mathrm{e}^{-2k^2x^2}, \quad x > 0, \tag{7}$$

(see section 6.10 in Resnick, 2002 e.g.).

2.1. ARMA processes and CUSUMs of residuals

Strong correlation can degrade the test in Theorem 1 (simulations Section 2.2 will support this later). As convergence to limit laws is typically faster with independent data, it may be beneficial to transform strongly correlated data into independent (or nearly independent) data. In that vein, we proceed with the assumption that $\{\epsilon_t\}$ is the unique (in mean square) causal and invertible stationary solution to the ARMA(p, q) difference equation

$$\epsilon_t - \phi_1\epsilon_{t-1} - \cdots - \phi_p\epsilon_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}, \quad t \in \mathbb{Z}, \tag{8}$$

where $\{Z_t\}$ is i.i.d. with $E[Z_t] = 0$ and $E[Z_t^2] = \sigma^2$. We let $\boldsymbol{\phi} = \{\phi_1, \ldots, \phi_p\}'$ and $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_q\}'$.

Causality implies that solutions to eqn (8) have the linear process form in eqn (2), where the weights $\{\psi_j\}$ are obtained from the ARMA coefficients in the usual manner (see Brockwell and Davis 1991, Chapter 3) and satisfy $\sum_{j=0}^{\infty}|\psi_j| < \infty$, thus implying absolute summability of the autocovariances. The spectral density formula for ARMA processes provides $\tau^2 = \sigma^2(1 + \sum_{i=1}^{q}\theta_i)^2/(1 - \sum_{j=1}^{p}\phi_j)^2$. If the ARMA orders p and q are known, then one substitutes $\sqrt{n}$-consistent ARMA parameter estimators $\hat{\boldsymbol{\phi}} = \{\hat{\phi}_1, \ldots, \hat{\phi}_p\}'$ and $\hat{\boldsymbol{\theta}} = \{\hat{\theta}_1, \ldots, \hat{\theta}_q\}'$ to obtain a $\sqrt{n}$-consistent parametric estimator of τ^2:

$$\hat{\tau}_2^2 = 2\pi\hat{f}(0) = \frac{\hat{\sigma}^2(1 + \hat{\theta}_1 + \cdots + \hat{\theta}_q)^2}{(1 - \hat{\phi}_1 - \cdots - \hat{\phi}_p)^2}. \tag{9}$$

In related work, Bai (1993, 1994) also suggests applying CUSUM methods to estimated one-step-ahead prediction residuals; Yu (2007) extends Bai's results to higher moments of the residuals in an effort to detect variance changepoints. The one-step-ahead prediction residual at time t is

$$I_t = \frac{X_t - \hat{X}_t}{\sqrt{\mathrm{Var}(X_t - \hat{X}_t)}}, \tag{10}$$

where $\hat{X}_t$ is the best linear prediction of X_t from linear combinations of a constant and $X_1, \ldots, X_{t-1}$. Here, $\mathrm{Var}(X_t - \hat{X}_t) = E[(X_t - \hat{X}_t)^2]$ is its unconditional variance. Because $\mathrm{Var}(X_t - \hat{X}_t)$ converges very rapidly (geometrically) downwards to σ^2, there is no reason to retain the denominator in eqn (10) in asymptotic analyses. Moreover, p. 265 of Brockwell and Davis (1991) suggests adopting a computationally convenient form of the prediction residuals (see also Bai, 1993, 1994; Yu, 2007). Mimicking eqn (8), these residuals are defined recursively in t via

$$\hat{Z}_t = (X_t - \hat{\mu}) - \hat{\phi}_1(X_{t-1} - \hat{\mu}) - \cdots - \hat{\phi}_p(X_{t-p} - \hat{\mu}) - \hat{\theta}_1\hat{Z}_{t-1} - \cdots - \hat{\theta}_q\hat{Z}_{t-q}; \tag{11}$$

here, $\hat{\mu}$ denotes a $\sqrt{n}$-consistent estimate of μ, and one simply takes $X_t - \hat{\mu} = 0 = \hat{Z}_t$ for $t \leqslant 0$. We use $\hat{\mu}$, $\hat{\boldsymbol{\phi}}$ and $\hat{\boldsymbol{\theta}}$ to represent null hypothesis estimates, which need to be $\sqrt{n}$-consistent only when $\Delta = 0$ and may be calculated using standard ARMA model fitting techniques for homogeneous data. An estimator of σ^2 is $\hat{\sigma}^2 = n^{-1}\sum_{t=1}^{n}\hat{Z}_t^2$. As Yu (2007) shows, no asymptotic loss of precision occurs when using $\{\hat{Z}_t\}$ in lieu of estimated versions of $\{I_t\}$ in a CUSUM procedure.

The CUSUM of residuals will be denoted by

$$\mathrm{CUSUM}_Z(k) = \frac{1}{\sqrt{n}}\left(\sum_{t=1}^{k}\hat{Z}_t - \frac{k}{n}\sum_{t=1}^{n}\hat{Z}_t\right).$$

Using the different equation structure in eqn (8), it is possible to relate $\mathrm{CUSUM}_X(k)$ and $\mathrm{CUSUM}_Z(k)$ directly. In fact, we offer the following result, which is proven in the Appendix.

LEMMA 1. *Suppose that $E[|Z_t|^\nu] < \infty$ for some $\nu \geqslant 2$. Then, under H_0,*

$$\frac{1}{\hat{\sigma}\sqrt{n}}\left(\sum_{t=1}^{k}\hat{Z}_t - \frac{k}{n}\sum_{t=1}^{n}\hat{Z}_t\right) = \frac{(1-\hat{\phi}_1-\cdots-\hat{\phi}_p)}{\hat{\sigma}\sqrt{n}(1+\hat{\theta}_1+\cdots+\hat{\theta}_q)}\left(\sum_{t=1}^{k}X_t - \frac{k}{n}\sum_{t=1}^{n}X_t\right) + \frac{A_k}{\sqrt{n}},$$

where $\max_{1\leqslant k\leqslant n}|A_k| = o_p(n^{1/\nu})$. The explicit form of the remainder A_k is provided in the Appendix.

Lemma 1 has an interesting implication: CUSUM changepoint inferences for ARMA processes will produce the same asymptotic conclusion when applied to either the raw series or the one-step-ahead residuals. The result is stated formally as the following.

THEOREM 2. *When $\{X_t\}$ is a causal and invertible ARMA series with innovations satisfying $E[|Z_t|^\nu] < \infty$ for some $\nu \geqslant 2$, and $\{\hat{Z}_t\}$ are the estimated residuals in eqn (11), then*

$$\frac{1}{\hat{\sigma}}\max_{1\leqslant k\leqslant n}|\mathrm{CUSUM}_Z(k)| - \frac{1}{\hat{\tau}}\max_{1\leqslant k\leqslant n}|\mathrm{CUSUM}_X(k)| = o_p(1),$$

where $\hat{\sigma}^2$ and $\hat{\tau}^2$ are consistent estimates of σ^2 and τ^2 respectively. Moreover,

$$\frac{1}{\hat{\sigma}}\max_{1\leqslant k\leqslant n}|\mathrm{CUSUM}_Z(k)| \xrightarrow{\mathcal{D}} \sup_{0\leqslant t\leqslant 1}|B(t)|. \quad (12)$$

2.2. Comparison of CUSUM techniques via simulation

Here, we examine the performance of the Theorem 1 and 2 statistics via simulation. We refer to tests utilizing eqn (5) as CU_X tests and tests utilizing (12) as CU_Z test. Good tests should have a Type I error probability that is close to the preset Type I error probability α. To explore this, we simulated a variety of time series, all under the null hypothesis of series homogeneity, and calculated the frequency at which each test rejects the null. The empirical rejection proportions are shown in Table 1. Here and in all other simulations presented in this article, we use $\sigma = 1$, we generate Gaussian errors, and we set $q_n = n^{1/3}$ when eqn (6) is used; also, all simulation results are based off of 10,000 replicate datasets per model, and we impose a target Type I error of $\alpha = 0.05$ throughout.

The Table 1 Type I errors vary wildly for the CU_X tests (regardless of how τ is estimated and even when it is taken as known). For the AR(1) and MA(1) simulations, values of ϕ_1 or θ_1 close to unity produce Type I errors so far from 0.05 that we cannot recommend using these tests. By contrast, the Type I error of the CU_Z test is consistently close to 0.05.

The poor performance of all tests in the ARMA(2,2) model when $\phi_1 = 0.5$, $\phi_2 = -0.2$ and $\theta_1 = -0.45$, $\theta_2 = -0.5$ is attributed to a near lack of invertibility; in fact, the MA(2) polynomial has a root at approximately 1.034. Models that are close to being non-invertible have a large $\tau^{-2} = [2\pi f(0)]^{-1}$. The results suggest that $\hat{\tau}_2$ provides a more accurate approximation of τ than $\hat{\tau}_1$; however, even when τ is taken as known, the CU_X tests may display an unacceptable Type I error. Thus, not all performance issues involve estimation of τ.

Table 1. CUSUM Type I error estimations for AR(1) models, MA(1) models, ARMA(1,1) models and ARMA(2,2) models with $n = 1000$ and $\alpha = 0.05$

AR(1) Models					MA(1) Models				
ϕ_1	CU_X^a	CU_X^b	CU_X^c	CU_Z	θ_1	CU_X^a	CU_X^b	CU_X^c	CU_Z
−0.95	0.0924	0.0001	0.0888	0.0442	−0.95	1.0000	0.0000	0.9985	0.0348
−0.9	0.0702	0.0014	0.0747	0.0486	−0.9	0.7733	0.0000	0.7605	0.0412
−0.5	0.0555	0.0286	0.0501	0.0449	−0.5	0.0559	0.0152	0.0597	0.0464
−0.1	0.0449	0.0391	0.0438	0.0431	−0.1	0.0465	0.0386	0.0437	0.0428
0.1	0.0427	0.0438	0.0439	0.0446	0.1	0.0457	0.0436	0.0460	0.0466
0.5	0.0438	0.0590	0.0359	0.0407	0.5	0.0438	0.0450	0.0410	0.0437
0.9	0.0250	0.3130	0.0214	0.0412	0.9	0.0468	0.0466	0.0420	0.0440
0.95	0.0192	0.5777	0.0091	0.0324	0.95	0.0450	0.0468	0.0398	0.0430
ARMA(1,1) Models					**ARMA(2,2) Models**				
ϕ_1	θ_1	CU_X^b	CU_X^c	CU_Z	$\{\phi_1, \phi_2\}$	$\{\theta_1, \theta_2\}$	CU_X^b	CU_X^c	CU_Z
0.5	−0.95	0.0000	0.7282	0.0502	{0.6, 0.35}	{0.6, −0.3}	0.7000	0.0041	0.0240
0.5	−0.9	0.0000	0.2465	0.0378	{0.6, 0.3}	{0.5, −0.2}	0.4154	0.0162	0.0344
0.5	−0.1	0.0564	0.0334	0.0386	{0.6, −0.1}	{−0.6, 0.3}	0.0584	0.0361	0.0396
0.1	−0.5	0.0178	0.0560	0.0454	{0.5, −0.2}	{−0.45,−0.5}	0.0000	0.9693	0.1415
0.9	−0.5	0.3006	0.0211	0.0397	{0.5, −0.2}	{−0.4, −0.5}	0.0000	0.7762	0.0530
0.95	−0.5	0.5594	0.0104	0.0346	{0.2, −0.5}	{−0.45, −0.05}	0.0009	0.0860	0.0466

[a] τ is taken as known.
[b] τ is estimated with $\hat{\tau}_1$.
[c] τ is estimated with $\hat{\tau}_2$.

Table 2. CUSUM Type I error estimations for AR(1) models with various n for $\phi_1 = 0.5$ and $\phi_1 = -0.5$ when $\alpha = 0.05$

	$\phi_1 = 0.5$				$\phi_1 = -0.5$			
n	$CU_X{}^a$	$CU_X{}^b$	$CU_X{}^c$	CU_Z	$CU_X{}^a$	$CU_X{}^b$	$CU_X{}^c$	CU_Z
50	0.0169	0.0135	0.0028	0.0096	0.0560	0.0034	0.0475	0.0241
100	0.0242	0.0388	0.0125	0.0246	0.0566	0.0100	0.0505	0.0333
200	0.0308	0.0523	0.0252	0.0342	0.0526	0.0167	0.0498	0.0386
400	0.0354	0.0590	0.0325	0.0397	0.0495	0.0206	0.0468	0.0392
800	0.0355	0.0579	0.0356	0.0406	0.0486	0.0264	0.0498	0.0439

[a] τ is taken as known.
[b] τ is estimated with $\hat{\tau}_1$.
[c] τ is estimated with $\hat{\tau}_2$.

Table 3. Power comparisons of CUSUM tests for various ARMA models with $\Delta = 0.15$, $n = 1000$ and $\alpha = 0.05$

$\boldsymbol{\phi}'$	$\boldsymbol{\theta}'$	$CU_X{}^a$	$CU_X{}^b$	CU_Z
$\{-0.2\}$	–	0.6871	0.7230	0.7189
$\{0.2\}$	–	0.3808	0.3679	0.3746
$\{0.4\}$	–	0.2581	0.2186	0.2300
–	$\{-0.4\}$	0.9060	0.9510	0.9445
–	$\{-0.2\}$	0.7218	0.7616	0.7566
–	$\{0.2\}$	0.4027	0.4011	0.4054
$\{-0.2\}$	$\{0.2\}$	0.5371	0.5299	0.5329
$\{0.2\}$	$\{-0.2\}$	0.5381	0.5306	0.5302
$\{0.4\}$	$\{-0.4\}$	0.5500	0.5394	0.5423
$\{0.3, -0.1\}$	$\{-0.1, -0.1\}$	0.5076	0.5499	0.5445
$\{-0.1, -0.1\}$	$\{0.3, -0.1\}$	0.5198	0.5599	0.5560
$\{0.3, 0.1\}$	$\{-0.1, -0.3\}$	0.4944	0.5428	0.5361

[a] τ is estimated with $\hat{\tau}_1$.
[b] τ is estimated with $\hat{\tau}_2$.

The instability of CUSUM tests is due to slow convergence of the partial sums sequence to a Brownian motion, especially when the AR(p) or MA(q) polynomial has a root close to the unit circle.

To further investigate convergence rates, we ran the tests for various sample sizes. Table 2 contains estimated p-values for each test for an AR(1) with $\phi_1 = 0.5$ and $\phi_1 = -0.5$ and for $n = 50, 100, 200, 400, 800$. The results suggest that the CU_Z appears to be conservative for small n. Also, the results indicate that the differences between tests can be even more noticeable for small n – conclusions drawn from the CU_X may be questionable even for moderate autocorrelation when the sample size is small.

To investigate if the CU_X test offers power increases over the CU_Z test, simulations similar to those in Tables 1 and 2 were run for moderate correlation structures with $n = 1000$, except that a mean shift of magnitude $\Delta = 0.15$ at time $c = n/2 = 500$ was added to the series. The results are shown in Table 3, where one sees that CU_Z tests are just as powerful as CU_X tests.

Of course, the ARMA orders p and q are rarely known in application; hence, CU_X methods are attractive as nonparametric tests. We also examined the ability of the CU_Z test to detect mean shifts when the ARMA orders were misspecified. Our results, which we summarize with the following example, indicate that the test works well when the ARMA orders are overestimated. When data are simulated from an ARMA(2,2) model with $\boldsymbol{\phi}' = \{0.4, 0.3\}$ and $\boldsymbol{\theta}' = \{0.1, -0.6\}$ and the resulting data are fitted with an ARMA(1,1), the Type I error probability is estimated as 0.1359 with the CU_Z test where we use a target of $\alpha = 0.05$. However, when these same series are fitted with an ARMA(4,4) model, the Type I errors become 0.0504 and 0.0445 [compared to 0.0444 and 0.0386 when using the correct ARMA(2,2) orders]. We also note that we observed no significant loss in power in residuals tests where the ARMA order was overestimated.

Our simulation study yields a clear theme for the practitioner: whenever possible, use residuals-based mean shift tests.

3. THE ADJUSTED CUSUM AND RELATED TESTS

Issues arise with the CUSUM test at the data boundaries. In particular, the limiting Brownian bridge is tied down at $t = 0$ and $t = 1$ [meaning $B(0) = B(1) = 0$] and hampers the ability of the test to detect mean shifts occurring near the beginning or end of the data. Many authors address this problem by scaling $\text{CUSUM}_X(k)$ by a weight function $w(k/n)$. Specifically, let w be a non-zero function defined on (0,1), increasing in a neighbourhood of zero, decreasing in a neighbourhood of unity, and satisfying & $\inf_{\zeta \leqslant t \leqslant 1-\zeta} w(t) > 0$ whenever $0 < \zeta < 1/2$. Define

$$I(w,\zeta) = \int_0^1 \frac{1}{t(1-t)} \exp\left(-\frac{\zeta w^2(t)}{t(1-t)}\right) dt.$$

Following theorem 2.1.1 of Csörgő and Horváth (1997), one obtains the following.

THEOREM 3. *Let $\hat{\tau}$ and $\hat{\sigma}$ be consistent estimators of τ and σ respectively. Then under H_0,*

$$\max_{1 \leqslant k \leqslant n} \frac{|\text{CUSUM}_X(k)|}{w(k/n)\hat{\tau}} \xrightarrow{\mathcal{D}} \sup_{0<t<1} \frac{|B(t)|}{w(t)} \quad \text{and} \quad \max_{1 \leqslant k \leqslant n} \frac{|\text{CUSUM}_Z(k)|}{w(k/n)\hat{\sigma}} \xrightarrow{\mathcal{D}} \sup_{0<t<1} \frac{|B(t)|}{w(t)}$$

if and only if $I(w, \zeta) < \infty$ for some $\zeta > 0$.

A natural weight choice is $w(t) = [t(1-t)]^{\gamma}$ for non-negative γ (Csörgő and Horváth, 1997). When $0 < \gamma < 1/2$, the conditions for Theorem 3 are satisfied, although null hypothesis percentiles akin to those in eqn (7) are not readily available. We will not consider such tests in our ensuing discussion. To use $\gamma \geqslant 1/2$, we must crop the set of admissible changepoint times. Below, we link LR and $F_{\max}$ tests to a weighted CUSUM with $w(t) = \sqrt{t(1-t)}$. Thereby, a limit result is needed for $w(t) = \sqrt{t(1-t)}$. We let

$$\lambda_X(k) = \frac{\text{CUSUM}_X^2(k)}{\frac{k}{n}\left(1-\frac{k}{n}\right)} \quad \text{and} \quad \lambda_Z(k) = \frac{\text{CUSUM}_Z^2(k)}{\frac{k}{n}\left(1-\frac{k}{n}\right)}, \tag{13}$$

and apply Theorem 3 to obtain the following key result.

THEOREM 4. *Suppose that $0 < \ell < h < 1$ and that $\hat{\tau}$ and $\hat{\sigma}$ are consistent estimators of τ and σ respectively. Then under H_0,*

$$\frac{1}{\hat{\tau}^2} \max_{\ell \leqslant \frac{k}{n} \leqslant h} \lambda_X(k) \xrightarrow{\mathcal{D}} \sup_{\ell<t<h} \frac{B^2(t)}{t(1-t)} \tag{14}$$

and

$$\frac{1}{\hat{\sigma}^2} \max_{\ell \leqslant \frac{k}{n} \leqslant h} \lambda_Z(k) \xrightarrow{\mathcal{D}} \sup_{\ell<t<h} \frac{B^2(t)}{t(1-t)}. \tag{15}$$

We refer to tests using statistics in Theorem 4 as *adjusted CUSUM* tests, the modifier 'adjusted' referring to the inclusion of the weighting function seen in eqn (13). One can find p-values for this test via

$$P\left(\sup_{\ell<t<h} \frac{B^2(t)}{t(1-t)} > x\right) \approx \sqrt{\frac{xe^{-x}}{2\pi}}\left\{\left(1-\frac{1}{x}\right)\log\left(\frac{(1-\ell)h}{\ell(1-h)}\right)+\frac{4}{x}\right\}$$

as $x \to \infty$, which is provided in Csörgő and Horváth (1997).

3.1. LR and $F_{\max}$ tests for AR(p)

Thus far, all second-order parameters have been estimated under the null hypothesis. While such estimates are relatively easy to calculate, they are not consistent under the alternative. A better approach would jointly estimate mean shift and autocovariance parameters at each admissible changepoint time k and construct test statistics from this information. Such a procedure is computationally intensive and considerably more involved to asymptotically quantify in generality. Because of this, we limit our discussion on such methods here to AR(p) processes. In this subsection, two procedures that utilize parameters estimated under the alternative, LR and $F_{\max}$ tests, are studied.

Assume that $\{\epsilon_t\}$ follows a causal autoregression (AR) of order p. For convenience with start-up effects, we assume that X_t is observed at times $t = -p+1, \ldots, n$. Considering a LR test for a mean shift, let $\hat{\sigma}^2 = n^{-1}\sum_{t=1}^{n} \hat{Z}_t^2$ be the estimator of the error variance under H_0, where $\hat{Z}_t$ is as defined in eqn (11) with $\hat{\mu}$ and $\hat{\boldsymbol{\phi}}$ representing null hypothesis conditional Gaussian likelihood estimators of μ and $\boldsymbol{\phi}$. (Note that one need not necessarily assume Gaussian innovations.)

Let $\hat{\boldsymbol{\phi}}_k = \{\hat{\phi}_{1,k}, \ldots, \hat{\phi}_{p,k}\}'$ be the conditional Gaussian likelihood estimators of $\boldsymbol{\phi}$ based on a likelihood function that allows for a mean shift at time k. For example, the prediction residual variance is estimated with

$$\hat{\sigma}_k^2 = \frac{1}{n}\sum_{t=1}^{n}[(X_t-\hat{\mu}_t)-\hat{\phi}_{1,k}(X_{t-1}-\hat{\mu}_{t-1})-\cdots-\hat{\phi}_{p,k}(X_{t-p}-\hat{\mu}_{t-p})]^2,$$

where

$$\hat{\mu}_t = \begin{cases} \hat{\mu}_k & \text{for } -p+1 \leqslant t \leqslant k, \\ \hat{\mu}_k^* & \text{for } k+1 \leqslant t \leqslant n, \end{cases} \tag{16}$$

with $\hat{\mu}_k$ denoting the likelihood estimator of μ and $\hat{\mu}_k^*$ denoting the likelihood estimator of $\mu^* = \mu + \Delta$ while allowing for a mean shift at time k. The likelihood estimators of $\hat{\mu}$, $\hat{\mu}_k$ and $\hat{\mu}_k^*$ differ from their sample mean counterparts $\bar{X}$, $\bar{X}_k$ and $\bar{X}_k^*$; see (3.103) in Shumway and Stoffer (2006) for $p = 1$ and Chapter 3 therein for generalities.

The Gaussian LR statistic for gauging a mean shift at time k is

$$\Lambda_k = \left(\frac{\hat{\sigma}_k^2}{\hat{\sigma}^2}\right)^{n/2},$$

where $\max_{\ell \leqslant \frac{k}{n} \leqslant h}\{-2\log\Lambda_k\}$ is used as the test statistic. Likewise, the F-statistic for a mean shift at time k is

$$F_k = \frac{SSE_0 - SSE_k}{SSE_k/(n-2)}.$$

where $SSE_0 = n\hat{\sigma}^2$ and $SSE_k = n\hat{\sigma}_k^2$. We connect $-2\log\Lambda_k$ and F_k via $\lambda_X(k)$:

$$\max_{\ell \leqslant \frac{k}{n} \leqslant h}\{-2\log\Lambda_k\} - \frac{1}{\hat{\tau}_2^2}\max_{\ell \leqslant \frac{k}{n} \leqslant h}\lambda_X(k) = o_p(1) \tag{17}$$

and

$$\max_{\ell \leqslant \frac{k}{n} \leqslant h} F_k - \frac{1}{\hat{\tau}_2^2}\max_{\ell \leqslant \frac{k}{n} \leqslant h}\lambda_X(k) = o_p(1), \tag{18}$$

where $\hat{\tau}_2^2$ is the null hypothesis estimator in eqn (9). Formulae (17) and (18) follow from the following, which is proven in the Appendix.

Lemma 2. *Suppose that $E[|Z_t|^\nu] < \infty$ for some $\nu \geqslant 2$. If $\{\epsilon_t\}$ is a causal AR(p) process, then under H_0,*

$$n(\hat{\sigma}^2 - \hat{\sigma}_k^2) = (1 - \hat{\phi}_1 - \cdots - \hat{\phi}_p)^2\lambda_X(k) + C_k,$$

where $\max_{\ell \leqslant \frac{k}{n} \leqslant h}|C_k| = o_p(n^{1/\nu - 1/2})$ *with* $0 < \ell < h < 1$.

Lemma 2 implies that the ratio of the null and alternative error variance estimators converges to unity, which yields eqn (18). To extract eqn (17) from Lemma 2, write $2\log\Lambda_k = n\log[1 - (\hat{\sigma}^2 - \hat{\sigma}_k^2)/\hat{\sigma}^2]$ and then Taylor expand the logarithm. Theorem 4 now gives the following result.

Theorem 5. *If the conditions of Lemma 2 hold, then*

$$\max_{\ell \leqslant \frac{k}{n} \leqslant h}\{-2\log\Lambda_k\} \xrightarrow{\mathcal{D}} \sup_{\ell < t < h}\frac{B^2(t)}{t(1-t)}, \tag{19}$$

and

$$\max_{\ell \leqslant \frac{k}{n} \leqslant h} F_k \xrightarrow{\mathcal{D}} \sup_{\ell < t < h}\frac{B^2(t)}{t(1-t)}. \tag{20}$$

One drawback of adjusted CUSUM methods lies with the need to truncate the admissible set of changepoints times to obtain a proper limit distribution of the supremum. A cancer patient seeking to detect the disease as soon as possible after onset would not consider truncating data boundaries. For such purpose, we would suggest a sequential analysis approach along the lines of Mei (2006). Another avenue of attack lies through extreme value scalings. As suggested earlier, if we let $T_n = \max_{1 \leqslant k \leqslant n}\lambda_X(k)$, then $T_n \to \infty$ almost surely as $n \to \infty$. One way to address this problem is to find constants a_n and b_n such that $a_nT_n - b_n$ converges to a Gumbel distribution asymptotically. Berkes *et al.* (2009) provide such results for the $\lambda_X(k)$ statistic. However, as is commonly conceded throughout the literature (see Csörgő and Horváth, 1997, p. 25, e.g.), statistics that employ convergence to extreme value distributions can exhibit significantly slower convergence rates than those that utilize quantiles based on Brownian motion.

3.2. Comparisons of the adjusted CUSUM and related rests

We compare adjusted CUSUM, LR and $F_{\max}$ statistics via simulation in this subsection. We call tests based on eqns (14) and (15) as λ_X and λ_Z tests respectively (both are adjusted CUSUM methods). The LR and $F_{\max}$ statistics are found in eqns (19) and (20) respectively. Here, we use $\ell = 1 - h = 0.05$ and $\alpha = 0.05$.

Simulations akin to those in Table 1 were conducted to compare λ_X and λ_Z tests. The conclusions are similar to those drawn from Table 1: residuals-based methods outperform non-residual methods. Because of this, we only consider residual tests in the ensuing discussion.

The λ_Z test enjoys practical advantages over the LR and $F_{\max}$ tests. For instance, LR and $F_{\max}$ tests are more computationally intensive. The Type I error rates in Table 4 indicate that the asymptotics for the LR and $F_{\max}$ tests are slower to 'kick in' than

Table 4. Type I error rates for the λ_Z, LR and F_{max} tests for various AR(1) models when $n = 200$ and $n = 1000$

	$n = 200$			$n = 1000$		
ϕ_1	λ_Z	LR	F_{max}	λ_Z	LR	F_{max}
0.9	0.0212	0.3237	0.3403	0.0329	0.3463	0.3501
0.7	0.0250	0.1348	0.1418	0.0354	0.0912	0.0930
0.5	0.0248	0.0754	0.0799	0.0401	0.0566	0.0577
0.3	0.0274	0.0524	0.0565	0.0384	0.0490	0.0496
0.1	0.0276	0.0418	0.0450	0.0409	0.0503	0.0510
−0.1	0.0330	0.0445	0.0469	0.0417	0.0413	0.0418
−0.3	0.0246	0.0370	0.0384	0.0372	0.0318	0.0424
−0.5	0.0348	0.0357	0.0376	0.0433	0.0433	0.0441
−0.7	0.0338	0.0329	0.0356	0.0432	0.0393	0.0400
−0.9	0.0372	0.0342	0.0362	0.0437	0.0382	0.0386

Table 5. Power approximations for the λ_Z, LR and F_{max} tests for various AR(1) models when $n = 200$ with c=20 and $\Delta = 1.0$ (Case 1) and when $n = 1000$ with $c = 500$ and $\Delta = 0.15$ (Case 2)

	Case 1			Case 2		
ϕ_1	λ_Z	LR	F_{max}	λ_Z	LR	F_{max}
0.5	0.1866	0.3427	0.3514	0.1107	0.1493	0.1498
0.3	0.4657	0.5743	0.5835	0.2052	0.2298	0.2316
0.1	0.7737	0.8106	0.8167	0.3307	0.3469	0.3492
−0.1	0.9501	0.9521	0.9543	0.4949	0.5016	0.5040
−0.3	0.9926	0.9915	0.9922	0.6616	0.6642	0.6654
−0.5	0.9995	0.9993	0.9994	0.8031	0.8030	0.8051

those for the λ_Z test. Another advantage of the λ_Z test is that it may be applied to processes with a moving average (MA) component.

Table 5 compares powers of the λ_Z, LR and F_{max} tests under the alternative hypothesis. Here, the results are mixed. One may believe that the LR and F_{max} tests would be more powerful than the λ_Z test, especially when the null hypothesis estimates of second-order parameters are biased, which is likely the case when the magnitude of the mean shift is large, when autocorrelations are strong, or when n is small. However, most large mean shifts will likely be detected by the λ_Z test. Also, when n is small or when the autocorrelation is strong, the LR and F_{max} tests may have too large of a Type I error. Although the results in Table 5 demonstrate situations in which the LR and F_{max} tests provide a significant increase in power, it does not seem worthwhile to tune the time-series parameter estimators to the changepoint time in general circumstances.

We also ran simulations to compare the utility of the λ_X and λ_Z tests against that of their extreme value analogues (as mentioned at the end of Section 3.1). The extreme value tests proved to be overly conservative for all practical sample sizes. Taking $\alpha = 0.05$, an extreme value test based off of the $\lambda_X(k)$-statistic while estimating τ with eqn (6) had an approximate Type I error rate of 0.0153 for AR(1) data with $\phi_1 = 0.3$ and with $n = 12{,}800$ (whereas the λ_X test observed a rate of 0.0526 for the same data).

At this point, we have now justified bullet points 1 and 3 in the Introduction. Our next section considers bullet point 2.

4. POWER COMPARISONS: CUSUM VS. ADJUSTED CUSUM

Our objective in this section is to compare the power of the residual versions of the CUSUM and adjusted CUSUM tests. If a mean shift is deemed to occur at time c, we will report the value of c/n. In accordance with eqn (1), the magnitude of the mean shift is denoted by Δ. In this subsection, we also use $\ell = 1 - h = 0.05$ as needed and we set a target Type I error rate of $\alpha = 0.05$ throughout.

First, consider an AR(1) series with $\varphi_1 = 0.5$. For $\Delta = 0.5$, Figure 1 plots the empirical power of the residual based tests as a function of c/n for $n = 500$,1000 and 2000.

Figure 1 shows powers that increase with n for all tests and decreases as c/n obtains close to 1 or 0. When the mean shift occurs near the endpoints, the adjusted CUSUM statistics are more powerful; by contrast, when the mean shift occurs in the middle of the data set, the CUSUM test is more powerful. Around $c/n = 0.25$, all tests have approximately the same power.

To investigate how the changepoint time, the model and the magnitude of the mean shift affect test power, we express the CUSUM of the data in terms of a CUSUM of zero mean random variables. When a mean shift occurs at time c, some computations provide

$$\frac{\mathrm{CUSUM}_X(k)}{\tau} = \frac{1}{\tau\sqrt{n}}\left(\sum_{i=1}^{k} Y_i - \frac{k}{n}\sum_{i=1}^{n} Y_i\right) - \sqrt{n}\cdot\min\left\{\frac{k}{n},\frac{c}{n}\right\}\left(1-\max\left\{\frac{k}{n},\frac{c}{n}\right\}\right)\frac{\Delta}{\tau}, \tag{21}$$

where $Y_i = X_i - \mu$ if $i \leqslant c$ and $Y_i = X_i - (\mu + \Delta)$ if $i > c$. Arguments akin to those in the proof of Lemma 1 establish a relationship similar to eqn (21) for CUSUMs of residuals. From this, we see that the asymptotic power of a CUSUM test is a function of

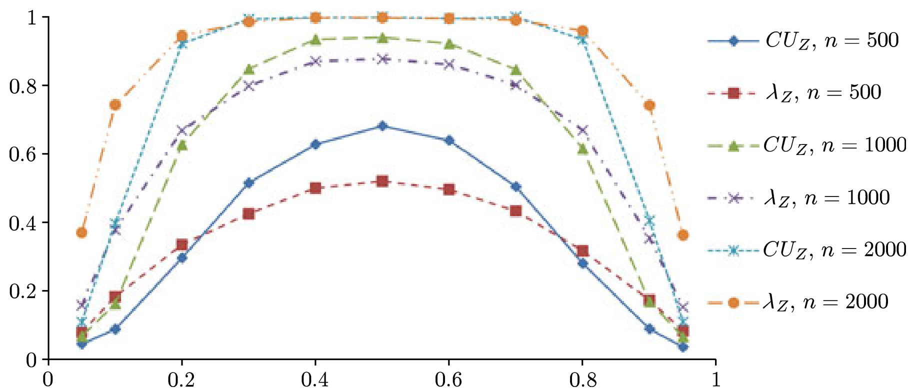

Figure 1. Graph of $\frac{c}{n}$ (horizontal axis) against power (vertical axis) with $\Delta = 0.5$ for an AR(1) with $\phi_1 = 0.5$

$$\delta_1 = \sqrt{n} \cdot \frac{c}{n}\left(1 - \frac{c}{n}\right)\frac{|\Delta|}{\tau}, \tag{22}$$

and the asymptotic power of an adjusted CUSUM test is a function of

$$\delta_2 = \sqrt{n} \cdot \sqrt{\frac{c}{n}\left(1 - \frac{c}{n}\right)}\frac{|\Delta|}{\tau}. \tag{23}$$

Table 6 explores the power claims in eqns (22) and (23) in more detail for the CU_Z and λ_Z tests. The estimated power of these tests is simulated for various series lengths, ARMA models, and mean shift times and magnitudes. We consider two cases: the first picks Δ so that δ_1 is always 1.2 and the second picks Δ so that δ_2 is always 3.5. The Table 6 powers for Case 1 are for CUSUM tests and those for Case 2 are for adjusted CUSUM tests. The powers in Table 6 appear roughly constant across the two columns, as they should be if eqns (22) and (23) hold. However, the last two rows show substantially smaller powers. The reduction of power here is again attributed to near unit roots of the AR or MA polynomials.

Finally, we examine the difference in power between the CUSUM and adjusted CUSUM tests in more detail. Figure 2 plots empirical powers with δ_1 or δ_2 held constant (and hence the powers for the CU_Z or λ_Z tests should be relatively constant) for various values of c/n. Because these powers are *a priori* symmetric about $c/n = 0.5$, we isolate on cases where $c/n \leqslant 0.5$. The results in Figure 2 show that the adjusted CUSUM tests are more powerful than CUSUM tests in some settings; however, when the CUSUM tests are more powerful, the difference is never substantial.

Overall, we suggest that the practitioner uses the adjusted CUSUM test as a first choice if the location of the changepoint is 'uniformly unknown'. Should a hypothesized changepoint lie closer to the centre of the data record, one may improve the power of the adjusted CUSUM test by increasing ℓ and $1-h$.

5. DATA APPLICATIONS

Theorems 2 and 5 demonstrate the asymptotic equivalence of the different methods. However, our simulation studies suggested that different methods may yield different conclusions in finite samples, which is a point that we now illustrate using data. Here, we apply the changepoint methods to two related data sets, the classical SOI and its related fish recruitment series. These series are

Table 6. The power of the CU_Z test when restricting $\delta_1 = 1.2$ (Case 1) and the power of the λ_Z test when restricting $\delta_2 = 3.5$ (Case 2)

				Case 1		Case 2	
ϕ'	θ'	n	c/n	Δ	Power	Δ	Power
$\{-0.5\}$	–	500	0.50	0.1431	0.5539	0.2087	0.7815
$\{0.1\}$	–	1000	0.40	−0.1757	0.5599	−0.2510	0.7894
$\{0.7\}$	–	2000	0.30	0.4259	0.5661	0.5693	0.7978
–	$\{-0.5\}$	500	0.20	−0.1677	0.5287	−0.1957	0.7388
–	$\{0.1\}$	1000	0.10	0.4638	0.5850	0.4058	0.7909
–	$\{0.7\}$	2000	0.05	−0.9603	0.6225	−0.6105	0.7981
$\{0.2\}$	$\{0.4\}$	1000	0.95	1.3981	0.5053	0.8887	0.7720
$\{-0.4\}$	$\{-0.3\}$	500	0.90	−0.2981	0.4860	−0.2609	0.7457
$\{0.5\}$	$\{-0.1\}$	2000	0.80	0.3019	0.5698	0.3522	0.7932
$\{0.1, -0.6\}$	$\{-0.8, 0.3\}$	2000	0.70	−0.0426	0.5731	−0.0569	0.7962
$\{0.9, -0.2\}$	$\{0.3, 0.2\}$	1000	0.60	0.7906	0.5474	1.1296	0.7917
$\{0.3, -0.6\}$	$\{0.7, -0.5\}$	500	0.50	−0.1982	0.5212	−0.2890	0.7569
$\{-0.6, 0.1\}$	$\{-0.5, -0.4\}$	500	0.70	−0.0170	0.2871	−0.0228	0.5151
$\{0.5, 0.4\}$	$\{0.1, 0.6\}$	1000	0.60	2.6879	0.4260	3.8407	0.6602

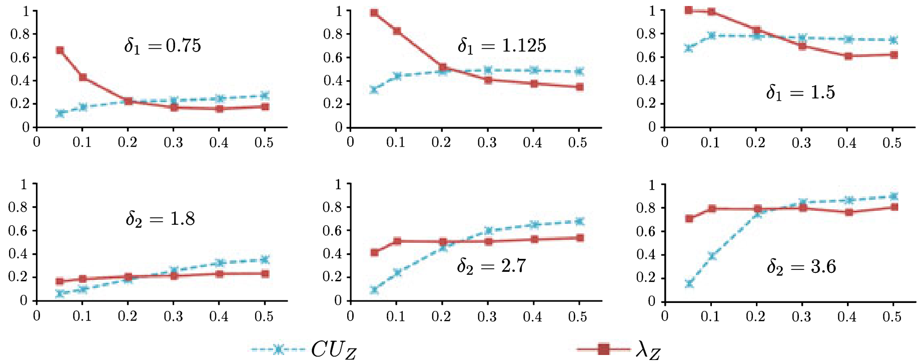

Figure 2. Power (vertical axes) vs. $\frac{c}{n}$ (horizontal axes) while holding δ_1 or δ_2 constant for an ARMA(1,1) with $\varphi_1 = 0.5$ and $\theta_1 = 0.5$ ($n = 1000$)

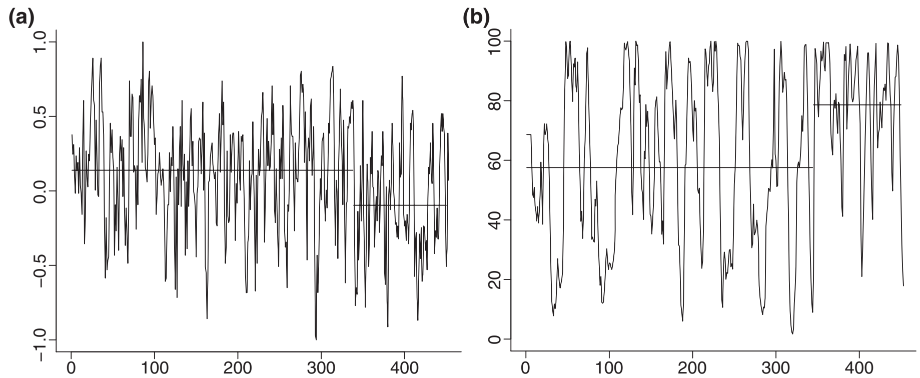

Figure 3. Time-series plots of SOI (a) and recruitment (b)

plotted in Figure 3 and are discussed further in Example 1.5 of Shumway and Stoffer (2006). Each series contains $n = 453$ monthly observations ranging from 1950 to 1987. The SOI series measures pressure differences between Darwin and Tahiti in the Pacific Ocean; the recruitment data count the number of new fish in a fish population. The SOI series measures the strength of the El-Nino effect and is known to influence weather patterns on a global scale.

Sample autocorrelations and partial autocorrelation suggest that an AR(2) model is reasonable for both the SOI and recruitment series respectively. With these model orders, the following changepoint statistics were computed: CU_X with the Bartlett estimator (6), CU_X with τ_2^2 estimated via eqn (9), CU_Z, the adjusted versions λ_X based on the Bartlett estimator $\hat{\tau}_1^2$, λ_X based on the estimator $\hat{\tau}_2^2$, λ_Z, and finally LR and $F_{\max}$.

The test statistics and p-values are summarized in Table 7 and vary considerably. It is clear that the choice of changepoint detection method matters greatly. Non-residual based tests produce smaller p-values than residual based tests. LR and $F_{\max}$ tests always reject. Despite the varying conclusions, all SOI test statistics are maximized at an estimated changepoint time of $\hat{c} = 339$ (ca. 1978), while most recruitment statistics are maximized 6 months later at $\hat{c} = 345$. A cross-correlation analysis reveals that SOI tends to lead recruitment by half a year (see Shumway and Stoffer, 2006 for details). The segmented levels in Figure 3 graphically portray estimated means before and after the estimated mean shift time.

Table 7. Changepoint test results for the SOI data (a), and the recruitment data (b) when an AR(2) is fit to both series

	SOI data			Recruitment data		
	Test stat.	p-value	$\hat{c}$	Test stat.	p-value	$\hat{c}$
$CU_X{}^a$	1.4733	0.0260	339	1.1895	0.1180	345
$CU_X{}^b$	1.1896	0.1179	339	0.8513	0.4632	345
CU_Z	1.2288	0.0976	339	0.8373	0.4848	344
$\lambda_X{}^a$	11.5264	0.0244	339	7.7923	0.1278	345
$\lambda_X{}^b$	7.5143	0.1440	339	3.9918	0.5866	345
λ_Z	8.0184	0.1159	339	3.8371	0.6192	344
LR	10.0815	0.0467	339	17.0518	0.0019	345
$F_{\max}$	10.1495	0.0453	339	17.3001	0.0017	345

[a] τ is estimated with $\hat{\tau}_1$.
[b] τ is estimated with $\hat{\tau}_2$.

With the changepoint times estimated at 339 and 345, AR parameter estimates are $\hat{\phi}_1 = 0.5767 \pm 0.0921$ and $\hat{\phi}_2 = 0.0018 \pm 0.0923$ for the SOI data and $\hat{\phi}_1 = 1.3508 \pm 0.0817$ and $\hat{\phi}_2 = -0.4647 \pm 0.0825$ for the recruitment data (these intervals have a 95% confidence level). These estimates suggest strong positive autocorrelation in both series. Given that the simulations in Tables 1 and 4 imply that the CUSUM and adjusted CUSUM tests have inflated Type I error probabilities in settings with strong correlation, the residuals-based tests are viewed as more reliable. As such, we conclude that neither series shows sufficient evidence of a changepoint at the 5% significance level. We have been unable to find any physical reason for changepoints circa 1978.

We noticed that the estimate for ϕ_2 in the SOI dataset is not significantly different from 0, therefore, we ran the changepoint tests while fitting an AR(1) to these data. The results were nearly identical to those seen in Table 7.

6. COMMENTS AND CONCLUSIONS

The results show that applying CUSUM methods to stationary correlated series to find undocumented mean shifts can produce spurious conclusions if care is not taken. The Type I errors in Table 1 for such tests are far from their intended values, even for sample sizes on the orders of thousands. The situation improves when one applies changepoint tests to the one-step-ahead prediction residuals of an ARMA model fitted to the data. With residual tests, it is shown that an adjusted CUSUM test is best at detecting mean shifts occurring near the data boundaries, while CUSUM tests are slightly more powerful when the changepoint occurs near the centre of the record. An undesirable feature of the proposed adjusted CUSUM test is that one must crop endpoint data. Phrased another way, it is difficult to detect changepoints occurring near the data boundaries. To avoid this, one could perform the test without cropping and use a bootstrapping technique to determine statistical significance.

APPENDIX

Proof of Lemma 1. Let $\epsilon_t = X_t - \mu$, where $\{\epsilon_t\}$ satisfies eqn (8) and set $\hat{\epsilon}_t = X_t - \hat{\mu}$. An application of eqn (11) gives

$$(1 + \hat{\theta}_1 + \cdots + \hat{\theta}_q)\sum_{t=1}^{k}\hat{Z}_t - (1 - \hat{\phi}_1 - \cdots - \hat{\phi}_p)\sum_{t=1}^{k}\hat{\epsilon}_t$$
$$= \sum_{\ell=1}^{q}\hat{Z}_{k-\ell+1}\sum_{j=\ell}^{q}\hat{\theta}_j. + \sum_{\ell=1}^{p}\hat{\epsilon}_{k-\ell+1}\sum_{j=\ell}^{p}\hat{\phi}_j.$$

Calling the bottom line in the above equation R_k and applying the triangle inequality yields

$$|R_k| \leqslant q^2 \max_{1\leqslant i\leqslant q}|\hat{\theta}_i| \max_{1\leqslant t\leqslant n}|\hat{Z}_t| + p^2 \max_{1\leqslant i\leqslant p}|\hat{\phi}_i| \max_{1\leqslant t\leqslant n}|\hat{\epsilon}_t|.$$

As the fitted model is causal and invertible, $Z_t = \sum_{j=0}^{\infty}\pi_j(\phi,\theta)\epsilon_{t-j}$, where $\pi_j = \pi_j(\phi,\theta)$ and $\hat{\pi}_j = \pi_j(\hat{\phi},\hat{\theta})$ are absolutely summable in j. Moreover, the definition of the prediction residual gives $\hat{Z}_t = \sum_{j=0}^{t-1}\hat{\pi}_j\hat{\epsilon}_{t-j}$. Hence,

$$|\hat{Z}_t| \leqslant \max_{1\leqslant t\leqslant n}|\hat{\epsilon}_t|\sum_{j=0}^{\infty}|\hat{\pi}_j|.$$

For any $\epsilon > 0$, there is some $K > 0$ and $0 < \beta < 1$ so that $|\hat{\pi}_j - \pi_j| \leqslant \epsilon K j \beta^{j-1}$ for all $j \geqslant 0$ whenever $|\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}| < \epsilon$ and $|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}| < \epsilon$. This follows from lemma 4.2.1 of Csörgő and Horváth (1997) and yields, by the consistency of $\hat{\boldsymbol{\phi}}$ and $\hat{\boldsymbol{\theta}}$, that $\sum_{j=0}^{\infty}|\hat{\pi}_j| \xrightarrow{\mathcal{P}} \sum_{j=0}^{\infty}|\pi_j|$. Consequently,

$$\max_{1\leqslant k\leqslant n}|R_k| = \mathcal{O}_p\left(\max_{1\leqslant t\leqslant n}|\hat{\epsilon}_t|\right).$$

It remains to estimate $\max_t|\hat{\epsilon}_t|$. We note that

$$\max_{1\leqslant t\leqslant n}|X_t| = o_p(n^{1/\nu}) \tag{24}$$

for some $\nu \geqslant 2$, which is justified below. Using eqn (24) and the weak consistency of $\hat{\mu}$ for μ, we see that $\max_t|\hat{\epsilon}_t| = o_p(n^{1/\nu})$. Combining these now gives

$$\left|\frac{\text{CUSUM}_Z(k)}{\hat{\sigma}} - \frac{\text{CUSUM}_X(k)}{\hat{\tau}_2}\right| \leqslant \frac{|R_k| + \frac{k}{n}|R_n|}{\hat{\sigma}\sqrt{n}|1 + \hat{\theta}_1 + \cdots + \hat{\theta}_q|} = \frac{|A_k|}{\sqrt{n}}.$$

We note that $\max_{1\leqslant k\leqslant n}|A_k| = o_p(n^{1/\nu})$, which follows from eqn (24), the weak consistency of the second-order parameter estimates and the fact that $|R_k| + \frac{k}{n}|R_n| \leqslant 2\max_{1\leqslant k\leqslant n}|R_k|$.

To verify eqn (24), use strict stationarity of $\{X_t\}$ to obtain $P(\max(X_1, \ldots, X_n) > \epsilon n^{1/\nu}) \leqslant nP(X_1 > \epsilon n^{1/\nu})$ for each $\epsilon > 0$. Now combine this with

$$
\begin{aligned}
nP(X_1 > \epsilon n^{1/\nu}) &\leqslant nP(|X_1|^\nu > n\epsilon^\nu) \leqslant n\int_{n\epsilon^\nu}^{\infty} \mathrm{d}F_\nu(y) \\
&\leqslant n\int_{n\epsilon^\nu}^{\infty} \frac{y}{n\epsilon^\nu}\mathrm{d}F_\nu(y) \leqslant \epsilon^{-\nu}\int_{n\epsilon^\nu}^{\infty} y\mathrm{d}F_\nu(y) \\
&= \epsilon^{-\nu}E[|X_1|^\nu I_{[|X_1|^\nu > n\epsilon^\nu]}],
\end{aligned}
$$

which tends to zero as $n \to \infty$ for any fixed $\epsilon > 0$ by the assumption that $E[|X_1|^\nu] < \infty$, to complete the proof. Here, $F_\nu(x) = P(|X_1|^\nu \leqslant x)$. □

Proof of Lemma 2. Set $\hat{\epsilon}_t = X_t - \hat{\mu}$, where $\hat{\mu}$ is the conditional maximum-likelihood estimator (MLE) of μ under H_0, and $\tilde{\epsilon}_t = X_t - \hat{\mu}_t$ for all t, where $\hat{\mu}_t$ is the conditional MLE allowing for a mean shift at lag k defined in eqn (16). Then, let $\hat{\gamma}(j) = n^{-1}\sum_{t=1}^{n}\hat{\epsilon}_{t-j}\hat{\epsilon}_t$ and $\hat{\gamma}_k(j) = n^{-1}\sum_{t=1}^{n}\tilde{\epsilon}_{t-j}\tilde{\epsilon}_t$ be estimators of $\gamma(j)$ under the null and alternative hypotheses allowing a mean shift at lag k respectively. Set $\hat{\Gamma} = \{\hat{\gamma}(i-j)\}_{i,j=0}^{p-1}$, $\hat{\boldsymbol{\gamma}} = \{\hat{\gamma}(1), \ldots, \hat{\gamma}(p)\}'$, $\hat{\Gamma}_k = \{\hat{\gamma}_k(i-j)\}_{i,j=0}^{p-1}$ and $\hat{\boldsymbol{\gamma}}_k = \{\hat{\gamma}_k(1), \ldots, \hat{\gamma}_k(p)\}'$. Extensive calculations then yield

$$
n(\hat{\sigma}^2 - \hat{\sigma}_k^2) = \sum_{i=0}^{p}\hat{\pi}_i\sum_{j=0}^{p}\hat{\pi}_j\sum_{t=1}^{n}(\hat{\epsilon}_{t-i}\hat{\epsilon}_{t-j} - \tilde{\epsilon}_{t-i}\tilde{\epsilon}_{t-j}) + (\hat{\boldsymbol{\phi}} - \hat{\boldsymbol{\phi}}_k)'n\hat{\Gamma}_k(\hat{\boldsymbol{\phi}} - \hat{\boldsymbol{\phi}}_k), \tag{25}
$$

where we employ the usual notation for invertible time series by setting $\hat{\pi}_0 = 1$ and $\hat{\pi}_i = -\hat{\phi}_i$ for all $i = 1, \ldots, p$ and $k = 1, \ldots, n$. Recall that $\hat{\boldsymbol{\phi}}$ and $\hat{\boldsymbol{\phi}}_k$ are null and alternative hypothesis MLEs, respectively, of $\boldsymbol{\phi}$. Derivation of eqn (25) uses $\hat{\boldsymbol{\gamma}}_k - \hat{\Gamma}_k\hat{\boldsymbol{\phi}}_k = \mathbf{0}$, which follows from the normal equations for autoregressive processes.

We consider the first term on the right-hand side of eqn (25). Since $\hat{\epsilon}_{t-i}\hat{\epsilon}_{t-j} - \tilde{\epsilon}_{t-i}\tilde{\epsilon}_{t-j}$ is symmetric in i and j, we assume without loss of generality that $0 \leqslant i \leqslant j \leqslant p$. Then, some lengthy computations yield that

$$
\begin{aligned}
\sum_{t=1}^{n}(\hat{\epsilon}_{t-i}\hat{\epsilon}_{t-j} - \tilde{\epsilon}_{t-i}\tilde{\epsilon}_{t-j}) &= \sum_{t=1}^{n}\left[(\hat{\mu}^2 - \hat{\mu}_t^2) - (\hat{\mu} - \hat{\mu}_t)X_t - (\hat{\mu} - \hat{\mu}_t)X_t\right] \\
&+ (\hat{\mu}_k - \hat{\mu})\left(\sum_{t=1-i}^{0}X_t + \sum_{t=1-j}^{0}X_t\right) \\
&+ (\hat{\mu} - \hat{\mu}_k^*)\left(\sum_{t=n-i+1}^{n}X_t + \sum_{t=n-j+1}^{n}X_t\right) \\
&+ (\hat{\mu}_k^* - \hat{\mu}_k)\left[j\hat{\mu}_k^* + i\hat{\mu}_k + \sum_{t=k+1}^{k+j-i}(X_{t-(j-i)} - X_t)\right],
\end{aligned} \tag{26}
$$

where the number of terms in each summation other than the first depends only on i and j (some of the summations may equal 0 if $i = 0$ and/or $i = j$). Further computations show

$$
\begin{aligned}
\sum_{t=1}^{n}\left[(\hat{\mu}^2 - \hat{\mu}_t^2) - (\hat{\mu} - \hat{\mu}_t)X_t - (\hat{\mu} - \hat{\mu}_t)X_t\right] &= -n\bar{X}^2 + k\bar{X}_k^2 + (n-k)(\bar{X}_k^*)^2 + n(\hat{\mu} - \bar{X})^2 \\
&\quad - k(\hat{\mu}_k - \bar{X}_k)^2 - (n-k)(\hat{\mu}_k^* - \bar{X}_k^*)^2.
\end{aligned} \tag{27}
$$

Next, we examine the rate of convergence of the conditional MLEs to their sample mean counterparts. To this end, taking partial derivatives of the null log likelihood with respect to μ and setting equal to zero yields

$$
(1 - \phi_1 - \cdots - \phi_p)\left(\sum_{t=1}^{n}X_t - n\mu\right) = \sum_{j=1}^{p}\phi_j\left(\sum_{t=1-j}^{0}X_t - \sum_{t=n-j+1}^{n}X_t\right),
$$

which in turn gives the bound

$$
|\bar{X} - \hat{\mu}| \leqslant \frac{1}{n}\frac{\max\{|\hat{\phi}_1|, \ldots, |\hat{\phi}_p|\}}{1 - \hat{\phi}_1 - \ldots - \hat{\phi}_p}\sum_{j=0}^{p-1}(p-j)|X_{-j} - X_{n-j}| = \mathcal{O}_p(n^{-1}). \tag{28}
$$

Similarly, setting the derivatives of the log likelihood under the alternative with resepect to μ_k and μ_k^* equal to zero and arguing as above yields

$$
\max_{\ell \leqslant \frac{k}{n} \leqslant h}|\bar{X}_k - \hat{\mu}_k| = o_p(n^{1/\nu - 1}) \quad \text{and} \quad \max_{\ell \leqslant \frac{k}{n} \leqslant h}|\bar{X}_k^* - \hat{\mu}_k^*| = o_p(n^{1/\nu - 1}), \tag{29}
$$

where we have used eqn (24). Calculations provide

$$
n^{-1/2} \cdot \mathrm{CUSUM}_X(k) = \frac{k}{n}(\bar{X}_k - \bar{X}) = \frac{k}{n}\left(1 - \frac{k}{n}\right)(\bar{X}_k - \bar{X}_k^*) = \left(1 - \frac{k}{n}\right)(\bar{X} - \bar{X}_k^*).
$$

Applying eqn (28), (29) and Donsker's theorem to the above, we see

$$\max_{\ell\leqslant\frac{k}{n}\leqslant h}|\hat{\mu}-\hat{\mu}_k| = \mathcal{O}_p(n^{-1/2}) \quad \text{and} \quad \max_{\ell\leqslant\frac{k}{n}\leqslant h}\left|\hat{\mu}-\hat{\mu}_k^*\right| = \mathcal{O}_p(n^{-1/2}). \tag{30}$$

Donsker's theorem and eqn (29) also give

$$\max_{\ell\leqslant\frac{k}{n}\leqslant h}|\hat{\mu}_k-\mu| = \mathcal{O}_p(n^{-1/2}) \quad \text{and} \quad \max_{\ell\leqslant\frac{k}{n}\leqslant h}|\hat{\mu}_k^*-\mu| = \mathcal{O}_p(n^{-1/2}). \tag{31}$$

Finally, eqns (26), (27) and additional calculations provide

$$\sum_{t=1}^{n}(\hat{\epsilon}_{t-i}\hat{\epsilon}_{t-j}-\tilde{\epsilon}_{t-i}\tilde{\epsilon}_{t-j}) = k\bar{X}_k^2+(n-k)(\bar{X}_k^*)^2-n\bar{X}^2+B_k = \lambda_X(k)+B_k, \tag{32}$$

where $\max_{\ell\leqslant\frac{k}{n}\leqslant h}|B_k| = o_p(n^{1/\nu-1/2})$ for some $\nu \geqslant 2$, which follows from eqns (24), (28), (29), (30) and (31).

It remains to show that the second term on the right-hand side of eqn (25) is asymptotically negligible. Since $\max_{\ell\leqslant\frac{k}{n}\leqslant h}\lambda_X(k) = \mathcal{O}_p(1)$, eqn (32) yields

$$\max_{\ell\leqslant\frac{k}{n}\leqslant h}|\hat{\gamma}(j)-\hat{\gamma}_k(j)| = \mathcal{O}_p(n^{-1}) \tag{33}$$

for all $0 \leqslant j \leqslant p$. Recognizing the equivalence of moment and MLEs in AR settings, this gives, when $\hat{\boldsymbol{\phi}} = \hat{\Gamma}^{-1}\hat{\boldsymbol{\gamma}}$ is also applied, the bound

$$\max_{\ell\leqslant\frac{k}{n}\leqslant h}\left|\hat{\phi}_j-\hat{\phi}_{j,k}\right| = \mathcal{O}_p(n^{-1}) \tag{34}$$

for the MLEs of the autoregressive parameters, $j = 0, 1, \ldots, p$. From the ergodic theorem (Theorem 24.1, Billingsley, 1995) along with eqn (33) we see

$$\max_{\ell\leqslant\frac{k}{n}\leqslant h}\hat{\gamma}_{k(j)} = \mathcal{O}_p(1) \tag{35}$$

for $j = 0, \ldots, p-1$. Combining eqns (32), (34) and (35), we can simplify eqn (25) to

$$n(\hat{\sigma}^2-\hat{\sigma}_k^2) = (\hat{\pi}_0+\hat{\pi}_1+\ldots+\hat{\pi}_p)^2\lambda_X(k)+C_k = (1-\hat{\phi}_1-\ldots-\hat{\phi}_p)^2\lambda_X(k)+C_k,$$

where $\max_{\ell\leqslant\frac{k}{n}\leqslant h}|C_k| = o_p(n^{1/\nu-1/2})$, which proves Lemma 2. □

Acknowledgements

The authors thank the referee, whose comments and suggestions improved the content and presentation of this article. This work was done while Michael Robbins was a PhD student at Clemson University. Robert Lund's research was supported by National Science Foundataion Grant DMS 0905570. Alexander Aue's research was supported by National Science Foundation Grant DMS 0905400.

REFERENCES

Andrews, D. W. K. (1991) Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* **59**, 817–58.
Antoch, J., Horváth, L. and Hušková, M. (1997) Effect of dependence on statistics for determination of change. *Journal of Statistical Planning and Inference* **60**, 291–310.
Bai, J. (1993) On the partial sums of residuals in autoregressive and moving average models. *Journal of Time Series Analysis* **14**, 247–60.
Bai, J. (1994) Weak convergence of the sequential empirical processes of residuals in ARMA models. *The Annals of Statistics* **4**, 2051–61.
Berkes, I., Gombay, E. and Horváth, L. (2009) Testing for changes in the covariance structure of linear processes. *Journal of Statistical Planning and Inference* **139**, 2044–63.
Billingsley, P. (1995) *Probability and Measure*, 3rd edn. New York: John Wiley and Sons.
Brockwell, P. R. and Davis, R. A. (1991) *Time Series: Theory and Methods*, 2nd edn. New York: Springer-Verlag.
Brown, R. L., Durbin, J. and Evans, J. M. (1975) Techniques for testing the constancy of regression relationships over time (with discussion). *Journal of the Royal Statistical Society, Series B* **37**, 149–92.
Csörgő, M. and Horváth, L. (1988) Nonparametric methods for changepoint problems. In *Quality Control and Reliability*, (eds P. R. Krishnaiah and C. R. Rao) Vol. **7**. Amsterdam, North Holland: Handbook of Statistics, pp. 403–25.
Csörgő, M. and Horváth, L. (1997) *Limit Theorems in Change-Point Analysis*. Chichester: John Wiley and Sons.
Davis, R. A., Huang, D. and Yao, Y.-C. (1995) Testing for a change in the parameter values and order of an autoregressive model. *Annals of Statistics* **23**, 282–304.
Davis, R. A., Lee, T. C. M. and Rodriguez-Yam, G. A. (2006) Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association* **101**, 223–39.
Gombay, E. and Horváth, L. (1990) Asymptotic distributions of maximum likelihood tests for change in the mean. *Biometrika* **77**, 411–4.
Lu, Q., Lund, R. B. and Lee, T. C. M. (2010) An MDL approach to the climate segmentation problem. *Annals of Applied Statistics* **4**, 299–319.
Lund, R. B. and Reeves, J. (2002) Detection of undocumented changepoints – a revision of the two-phase regression model. *Journal of Climate* **17**, 2547–54.

MacNeill, I. B. (1974) Tests for change of parameter at unknown times and distributions of some related functionals of Brownian motion. *Annals of Statistics* **2**, 950–62.

Mei, Y. (2006) Sequential change-point detection when the unknown parameters are present in the pre-change distribution. *Annals of Statistics* **34**, 92–122.

Newey, W. K. and West, K. D. (1987) A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* **55**, 703–8.

Page, E. S. (1954) Continuous inspection schemes. *Biometrika* **41**, 100–5.

Page, E. S. (1955) A test for a change in a parameter occurring at an unknown point. *Biometrika* **42**, 523–7.

Quandt, R. E. (1958) The estimation of the parameters of a linear regression system obeying two separate regimes. *Journal of the American Statistical Association* **53**, 873–80.

Quandt, R. E. (1960) Tests of the hypothesis that a linear regression system obeys two separate regimes. *Journal of the American Statistical Association* **55**, 324–30.

Reeves, J., Chen, J., Wang, X. L., Lund, R. B. and Lu, Q. (2007) A review and comparison of changepoint detection techniques for climate data. *Journal of Applied Meteorology and Climatology* **46**, 900–15.

Resnick, S. I. (2002) *Adventures in Stochastic Processes*. Boston: Birkhäuser.

Robbins, M. W., Lund, R. B., Gallagher, C. M. and Lu, Q. (2011) Changepoints in the North Atlantic tropical cyclone record. *Journal of the American Statistical Association*. Forthcoming.

Shumway, R. H. and Stoffer, D. S. (2006) *Time Series Analysis and Its Applications*. New York: Springer Verlag.

Yao, Y.-C. and Davis, R. A. (1984) The asymptotic behavior of the likelihood ratio statistic for testing a shift in mean in a sequence of independent normal variates. *Sankhyā Series A* **48**, 339–53.

Yu, H. (2007) High moment partial sum processes of residuals in ARMA models and their applications. *Journal of Time Series Analysis* **28**, 72–91.