CHANGEPOINT DETECTION FOR DEPENDENT GAUSSIAN SEQUENCES

W. ELLENS [•,⋆], J. KUHN [•,⋆], M. MANDJES [•], P. ŻURANIEWSKI [⋆,◦]

ABSTRACT. In this paper easily applicable techniques are devised for detecting changepoints in autocorrelated Gaussian sequences. Our method proceeds by sequential evaluation of a CUSUM-type test statistic, which is compared to a predefined threshold. We assume that data is tested in sliding windows of fixed size. The distinguishing feature of this work is that, based on large deviations theory, we derive rather explicit equations that determine the threshold in such a way that the false alarm probability per window is approximately kept at the desired level. This criterion – as opposed to the usual average run length – allows to restrict not only the average number of false alarms but also their variability. Illustrative examples are provided, including the detection of a shift in mean in ARMA processes. The procedures are validated by means of a broad set of simulation experiments, and overall perform well.

• Korteweg-de Vries Institute for Mathematics, University of Amsterdam, the Netherlands.

⋆ TNO Performance of Networks and Systems, Delft, the Netherlands.

◦ Department of Applied Mathematics, AGH University of Science and Technology, Kraków, Poland

M. Mandjes is also with EURANDOM (Eindhoven University of Technology, the Netherlands) and CWI (Amsterdam, the Netherlands).

Email: wendy.ellens@tno.nl,j.kuhn@uva.nl,m.r.h.mandjes@uva.nl,piotr.zuraniewski@tno.nl

1. INTRODUCTION

The ability to detect *changepoints* in data sequences (corresponding to a change in the underlying probability distribution) is of great practical importance, and one of the main concerns in statistical process control (SPC). In numerous application domains one is faced with problems of this nature. To mention but a few examples, changepoint techniques have been used in finance [10], electrocardiogram analysis [16, 17], climate change [4] and communication networks [9, 38].

In the basic changepoint detection problem the goal is to detect a changepoint in a sequence of independent observations of some quality variable of interest. For continuous data it is

Date: February 8, 2016.

Key words and phrases. Changepoint detection, CUSUM, multivariate normal distribution, ARMA processes, large deviations theory, likelihood ratio.

usually assumed that the data is independent and normally distributed [35], and the change of interest is often a shift in the mean value. The goal is to detect it as soon as possible, while at the same time limiting the number of false alarms. Theoretical background on changepoint detection can be found in the books [2, 28]. The survey [35] gives an overview on SPC from a practical perspective.

A commonly used technique in changepoint detection is that of *Cumulative Sum* (CUSUM) [26]. For independent data, the CUSUM statistic for detecting a change from the in-control parameter value to a pre-specified alternative can be expressed in terms of cumulative sums of log-likelihood ratio (LLR) increments. The monitoring is stopped and an alarm is issued as soon as the test statistic exceeds some predefined threshold.

In the literature the question of how the threshold should be chosen is often disregarded, and when it is not, then typically the threshold has been selected based on a condition on the *average run length* (ARL), the expected time till the first false alarm [2]. Since the ARL is simply the average of the stopping time, an obvious drawback of this approach is that it does not allow to restrict the variability of the false alarms. This can be crucial when thinking of applications in networks: Imagine, for example, one were to monitor patients' health data in a hospital (thus, testing multiple independent data streams in parallel). Then a high variance of false alarms could lead to a scenario where the capacity of the hospital staff is exceeded because a large number of false alarms (next to actual ones) occurred at the same time.

Therefore, instead of the ARL, in this paper we restrict the probability of raising a false alarm in any given window. This ensures that the false alarm probability is low locally (and is thus also still low on average). Furthermore, this approach circumvents an issue pointed out in [25], namely that the ARL is not always finite, and thus not in general an informative criterion.

In his influential paper [19], Lai proposed two other false alarm criteria as an alternative to the ARL, which allow to limit the variance of the number of false alarms. It turns out that asymptotically, for window-limited detection, our criterion and one of Lai's are similar (see Section 4). With respect to Lai's second criterion (which was coined *maximum local false alarm probability* in [36]), our method has the advantage of simplicity. In fact, it was stated in [36, Ch. 8] that the practical implementation of this criterion is difficult because no closed-form expressions or even bounds are available that would allow the selection of the threshold.

Previous results on how to select the threshold usually restrict the data points to be independent. For example, under this assumption the conceivable fact is proven that (under an

appropriate scaling) a functional central limit theorem (CLT) holds, meaning that the cumulative random walk process converges to a Brownian motion. This result enables us to assess the test's false alarm probability [33]. Apart from the CLT regime, asymptotic expressions for the false alarm probability have been derived under a large deviations scaling as well, see e.g. [8, Ch. VI.E] and [13]. Because these asymptotic expressions are available in closed form, choosing the threshold based on these results is relatively easy, yet ensures that the false alarm probability is limited.

The analysis complicates significantly, however, if the observations do *not* correspond to independent variables. This situation is highly relevant, as in many practical situations the observations constituting the data sequence cannot be assumed independent. In the networking context, we refer to, e.g., the nice (unpublished) overview [39] for an extensive treatment of traffic characteristics in communication networks; notably, it has been found that there are non-negligible correlations over broad ranges of time scales.

This motivates that in the current paper we focus on Gaussian processes that exhibit serial dependence. An important class of Gaussian processes that include dependence is that of the so-called autoregressive moving-average (short: ARMA) processes [5, 6], which we consider as a more specific example. For the class of ARMA processes Johnson and Bagshaw [18] established the convergence to Brownian motion, thus enabling the type I error (false alarm) analysis of a CUSUM-type procedure. Alternative tests under the CLT scaling were described extensively by Czörgő and Horváth [11, Ch. IV], with a focus on a Brownian-bridge based test statistic (see also [3]). Basseville and Nikiforov [2, Ch. 7] discuss testing procedures for dependent Gaussian processes that rely on a whitening transformation of the data sequence. A similar avenue is taken in [14] and [29] for the problem of mean shift detection in ARMA processes. Besides these works, upper bounds have been provided for more general scenarios, where the Gaussianity assumption is relaxed (see e.g. [19] and [37]).

The current paper contributes to the theory on changepoint techniques for serially correlated data. We develop a window-limited testing procedure with LLR test statistic (in the spirit of the CUSUM method), and provide a method for selecting the threshold (function) such that the probability of raising a false alarm is low in every given window of data points, as motivated above. An advantage of testing data in windows rather than keeping the entire history

of observations is that a change can be detected more quickly since it has a bigger impact relative to the (fewer) previous observations within the current window. Furthermore, the usual assumption of stationary data is less restrictive in this case.

While previous (asymptotic) work on CUSUM for dependent data has primarily focused on the CLT regime, in the present paper we consider a large-deviations (short: LD) setting. More specifically, we construct LD-based CUSUM-type changepoint detection tests for dependent normal data, covering also the class of (Gaussian) ARMA processes. Since LD theory [8, 12] focuses on the rare-event setting, this framework is particularly suitable for the problem at hand as the probability of raising a false alarm is required to be low.

An additional attractive feature of applying LD theory here is that it nicely facilitates the analysis of hypothesis testing with multiple alternatives. In the changepoint detection problem we have to consider a union of hypotheses corresponding to a change in a parameter value *at some point* in the dataset. In the LD regime the probability of such a union of events essentially coincides with the probability of the most likely event among them; this phenomenon is usually referred to as the *principle of the largest term* [15]. We therefore obtain a threshold *function* rather than a single value as is usually assumed (see [2, 36]), ensuring that the probability of raising a false alarm is essentially equally likely irrespective of the location of the changepoint. We provide a numerical example in Section 5 that indicates that choosing a threshold *function* is indeed favourable.

In that section, we also discuss a number of relevant cases in greater detail: a change in the mean (with the correlations held fixed), a change in variance (for independent observations), and a change of the 'scale' of the process (that is, the means blow up by a factor f, the covariance matrix by a factor f^2). In these cases we obtain particularly simple equations for the threshold function, see Eqs. (12), (14) and (15), respectively. The change in scale example was considered in more detail in [21] in a multidimensional setting; it has applications in the context of communication networks where a change in scale may result from an increase of the number of users.

The paper is organized as follows. In Section 2 we provide preliminaries on CUSUM, reviewing the *independent case* in the LD scaling. Then Section 3 provides a series of useful computations for likelihood ratio tests related to multivariate normal distributions, which are used in Section 4 to develop changepoint detection tests for *dependent* data, and includes the aforementioned more specific examples. Section 5 presents an extensive simulation study so as to

assess the performance of the tests; these experiments confirm that the proposed procedure works well in a broad range scenarios.

2. CUMULATIVE SUM: PRELIMINARIES

Consider a representative window of observations $X_1, X_2, \ldots, X_n$, during which potentially a changepoint occurs. In this section we assume that the X_i are independent, but we do not assume anything about their distribution. Later in this paper we look at situations in which the X_i may be dependent, but follow a normal distribution. In probabilistic terms a *changepoint*, to be considered as a change in the statistical law of the underlying random variable, can be described as follows.

- Under the null-hypothesis (H_0) the X_i ($i = 1, \ldots, n$) are independent and identically distributed (i.i.d.) realizations of a random variable with density $f(\cdot)$.
- Under the alternative hypothesis (H_1) up to $k-1$ the observations are i.i.d. samples from a distribution with density $f(\cdot)$, while from observation k on they are i.i.d. with a *different* density $g(\cdot)$ (for some k ranging between 1 and n).

In other words: under the null-hypothesis there has *not* been a changepoint, while under the alternative hypothesis the process changes. Observe that this setup is not a simple binary hypothesis testing problem, as the alternative is essentially a *union* of hypotheses. More precisely: with $H_1(k)$ corresponds to having a changepoint at k, we can write H_1 as the union of the $H_1(k)$, with $k = 1, \ldots, n$.

A changepoint detection test, that is, a test that determines whether to accept the null hypothesis or to reject it — in which case it issues an alarm —- aims at keeping the probability of a type I error (a false alarm) limited. On the other hand, the test should be such that the detection probability is as high as possible, in other words, it should minimize the type II error probability while maintaining the false alarm rate at a given low level.

The technique we describe in this section, known as CUSUM, has been proposed [26] to identify parameter changes from the in-control value to a pre-specified alternative. Since in practice the parameter after the change is typically unknown, it is commonly replaced by its maximum likelihood estimator (resulting in the generalized likelihood ratio (GLR) test), or by some smallest tolerable value [1]. Also a combination of multiple testing procedures is possible, as, for example, proposed in [40]. Since this question is not in the scope of the current paper, in the following description of the CUSUM method we assume that the alternative is specified — we roughly follow the setup presented in [33, Ch. II.6].

Consider first the common likelihood test for H_0 versus $\mathrm{H}_1(k)$. Evidently, the statistic to be considered is

$$\bar{S}_k := \left(\prod_{i=k}^{n} g(X_i)\right) \Bigg/ \left(\prod_{i=k}^{n} f(X_i)\right);$$

it turns out, though, that it is more practical to work with the corresponding *log*-likelihood:

$$S_k := \sum_{i=k}^{n} \log\left(\frac{g(X_i)}{f(X_i)}\right).$$

To deal with the fact that H_1 equals the union of the $\mathrm{H}_1(k)$, we have to verify whether there is a $k \in \{1, \ldots, n\}$ such that S_k exceeds a certain critical value. As a result, the statistic for the *composite* test (that is, H_0 versus H_1) is

$$t_n := \max_{k\in\{1,\ldots,n\}} S_k = T_n - \min_{k\in\{1,\ldots,n\}} T_{k-1}, \tag{1}$$

with T_k denoting the cumulative sum $\sum_{i=1}^{k} \log\big[g(X_i)/f(X_i)\big]$; the null-hypothesis is rejected if t_n exceeds some critical level b.

Observe from the above that the test statistic can be written in terms of the cumulative sums T_k (corresponding to increments that are distributed as $g(X_i)/f(X_i)$), which explains the name of the test. Also, note that the statistic (1) represents the height of the random walk T_k relative to the minimum that was achieved so far; in this sense, there is a close connection to an associated (discrete-time) *queueing* process, as described in, e.g., [33]. CUSUM has certain optimality problems in terms of the tradeoff mentioned above (timely detection versus low rate of false alarms, that is), as established in a Bayesian framework in [31, 32], whereas [20, 27] address this property in the non-Bayesian setting.

We now scale the threshold b by n, and focus on asymptotics for large n; this limiting regime is usually referred to as the *large deviations regime* [8, 12, 22]. More specifically, we analyze the probability of issuing a false alarm (type I error), that is, $\mathbb{P}_0(t_n \geq nb)$. Here $\mathbb{P}_0$ corresponds to probability under H_0 and $\mathbb{E}_0$ is the associated expectation. We roughly follow the setup of [8, Ch. VI.E]. Under H_0, due to reversibility arguments,

$$t_n = T_n - \min_{k\in\{1,\ldots,n\}} T_{k-1} = \max_{k\in\{1,\ldots,n\}} (T_n - T_{k-1}) \stackrel{\mathrm{d}}{=} \max_{k\in\{1,\ldots,n\}} T_k,$$

so that the probability of our interest can be rewritten as

$$\mathbb{P}_0(t_n \geq nb) = \mathbb{P}_0\left(\exists k \in \{1, \ldots, n\} : T_k \geq nb\right).$$

Due to $n^{-1} \cdot \log n \to 0$ and

$$\max_{k\in\{1,\ldots,n\}} \mathbb{P}_0\left(T_k \geq nb\right) \leq \mathbb{P}_0\left(\exists k \in \{1,\ldots,n\} : T_k \geq nb\right) \leq n \cdot \max_{k\in\{1,\ldots,n\}} \mathbb{P}_0\left(T_k \geq nb\right),$$

we have the following expression for the so-called *decay rate*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}_0(t_n \geq nb) = \max_{\lambda\in(0,1]} \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}_0\left(\frac{T_{n\lambda}}{n} \geq b\right)$$

(realize that $n\lambda$ is not necessarily integer, so there is mild abuse of notation in the previous display); in words, this means that the decay rate of the union of all n events coincides with the decay rate of the most likely event among these (the so-called 'principle of the largest term'; see [15]). Relying on Cramér's theorem [8, Ch. II.A], we can rewrite the above decay rate to

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}_0(t_n \geq nb) = \max_{\lambda\in(0,1]} \lim_{n\to\infty} \frac{\lambda}{n\lambda} \log \mathbb{P}_0\left(\frac{T_{n\lambda}}{n\lambda} \geq \frac{b}{\lambda}\right) = \max_{\lambda\in(0,1]} \left(-\lambda \sup_{\theta} \left(\theta \frac{b}{\lambda} - \log M(\theta)\right)\right);$$

here $M(\theta)$ is the moment generating function (under H_0) of $\log\left[g(X_i)/f(X_i)\right]$:

$$M(\theta) = \mathbb{E}_0 \exp\left(\theta \log \frac{g(X_i)}{f(X_i)}\right) = \mathbb{E}_0 \left(\frac{g(X_i)}{f(X_i)}\right)^{\theta} = \int_{-\infty}^{\infty} (g(x))^{\theta} (f(x))^{1-\theta} \mathrm{d}x.$$

We can then set b such that the decay rate under study equals some predefined (negative) constant $-\gamma$ (where $\gamma > 0$). In principle, however, there is no need to take a *constant* b; we could pick a *function* $b(\lambda)$ instead. It can be seen that, in terms of optimizing the type II error performance, it is optimal to choose this function $b(\lambda)$ such that

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}_0\left(\frac{T_{n\lambda}}{n} \geq b(\lambda)\right) = -\lambda \sup_{\theta} \left(\theta \frac{b(\lambda)}{\lambda} - \log M(\theta)\right)$$

is constant in $\lambda \in (0, 1]$ (and equaling $-\gamma$). Intuitively, this choice entails that for any point $n\lambda$ in time, issuing an alarm (which is done if $T_n - T_{n\lambda-1}$ exceeds $nb(1 - \lambda + 1/n)$) is essentially equally likely if there is no changepoint.

In the setup described above the individual observations X_i are assumed to be independent. The main objective of the paper is to develop a machinery that can deal with *dependent* data. As mentioned earlier, we focus on the case that the data stem from a multivariate normal distribution. To this end, we first work out the likelihood ratio test of a single multivariate normal distribution against another one in Section 3, which is used in Section 4 to develop a changepoint detection procedure for dependent normal data.

3. LIKELIHOOD RATIO TEST FOR MULTIVARIATE NORMAL DATA

As we saw in the previous section, the CUSUM method is in essence a sequentially applied LLR hypothesis test. We therefore first consider the situation that under H_0 the data $X_1, \ldots, X_n$ has a normal distribution with mean $\bar{\mu}$ under H_0 and mean $\bar{\nu}$ under H_1. That is, in this section we assume that there is no changepoint (or, equivalently, that the change has occurred already at the first observation within the considered window). The results of this section will be used in Section 4 to develop a procedure to find a change *somewhere* in the sequence.

It is immediately seen that, without loss of generality, we can pick $\bar{\mu} = 0$ (by subtracting it from $\bar{\nu}, X_1, \ldots, X_n$). Because we wish to explicitly allow for correlated data points, we further assume that the vector of observations $\boldsymbol{X} = (X_1, \ldots, X_n)$ stems from an n-dimensional multivariate normal distribution with mean $\boldsymbol{\mu}_n \equiv \boldsymbol{\mu}$ and covariance matrix $\Sigma_n \equiv \Sigma$ (which thus does not need to be diagonal), denoted by $\mathscr{N}(\boldsymbol{\mu}, \Sigma)$, while under H_1 they stem from $\mathscr{N}(\boldsymbol{\nu}, T)$.

We let $f_n(\cdot)$ and $g_n(\cdot)$ be the corresponding n-dimensional densities, that is,

$$f_n(\boldsymbol{x}) = (2\pi)^{-n/2} \,|\, \Sigma \,|^{-1/2} \, \exp\left(-\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{x}\right),$$

and

$$g_n(\boldsymbol{x}) = (2\pi)^{-n/2} \,|\, T \,|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\nu})^{\mathrm{T}}T^{-1}(\boldsymbol{x}-\boldsymbol{\nu})\right).$$

Observe that $\boldsymbol{\mu}$ and $\boldsymbol{\nu} \in \mathbb{R}^n$, while Σ and T are positive-definite matrices of dimension $n \times n$. In this section, we first develop a large-deviations based likelihood ratio test for distinguishing $g_n(\cdot)$ from $f_n(\cdot)$, and then specialize to a series of relevant special cases.

A LLR hypothesis test features the test statistic

$$\mathscr{L}_n(\boldsymbol{X}) = \log\left(\frac{g_n(\boldsymbol{X})}{f_n(\boldsymbol{X})}\right),$$

which can be evaluated as

$$\mathscr{L}_n(\boldsymbol{X}) = \frac{1}{2}\log|\,\Sigma\,| - \frac{1}{2}\log|\,T\,| + \frac{1}{2}\boldsymbol{X}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{X} - \frac{1}{2}(\boldsymbol{X}-\boldsymbol{\nu})^{\mathrm{T}}T^{-1}(\boldsymbol{X}-\boldsymbol{\nu}). \tag{2}$$

To determine the critical value nb above which the null hypothesis is rejected, we wish to evaluate the type I error probability $\mathbb{P}_0(\mathscr{L}_n(\boldsymbol{X}) \geq nb)$, where $b > \mathbb{E}_0\mathscr{L}_n(\boldsymbol{X})/n$. It turns out to be hard to evaluate this probability explicitly, but we can derive an accurate approximation based on large deviations theory. Relying on the Gärtner-Ellis theorem [8, 12], the following equation holds for the decay rate

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}_0(\mathscr{L}_n(\boldsymbol{X}) \geq nb) = -\mathscr{I}(b),$$

where $\mathscr{I}(b)$ denotes the associated Legendre transform

$$\mathscr{I}(b) := \sup_{\theta} \left(\theta b - \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}_0 \exp(\theta \mathscr{L}_n(\boldsymbol{X})) \right), \tag{3}$$

given that the limiting log-moment generating function exists. This leads to the approximation

$$\mathbb{P}_0(\mathscr{L}_n(\boldsymbol{X}) \geq nb) \approx e^{-n\mathscr{I}(b)}.$$

To use this approximation, we first compute the moment generating function $\mathbb{E}_0 \exp(\theta \mathscr{L}_n(\boldsymbol{X}))$ in more explicit terms. It is clear that

$$\mathbb{E}_0 \exp(\theta \mathscr{L}_n(\boldsymbol{X})) = (2\pi)^{-n/2} \,|\, \Sigma \,|^{-1/2} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp(\theta \mathscr{L}_n(\boldsymbol{x})) \exp\left(-\frac{1}{2} \boldsymbol{x}^{\mathrm{T}} \Sigma^{-1} \boldsymbol{x} \right) \mathrm{d}x_1 \cdots \mathrm{d}x_n.$$

Then notice that

$$\theta \mathscr{L}_n(\boldsymbol{x}) - \frac{1}{2} \boldsymbol{x}^{\mathrm{T}} \Sigma^{-1} \boldsymbol{x} = \frac{\theta}{2} \log \frac{|\,\Sigma\,|}{|\,T\,|} - \frac{1}{2} \boldsymbol{x}^{\mathrm{T}} (\theta T^{-1} + (1-\theta)\Sigma^{-1}) \boldsymbol{x} + \theta \boldsymbol{\nu}^{\mathrm{T}} T^{-1} \boldsymbol{x} - \frac{\theta}{2} \boldsymbol{\nu}^{\mathrm{T}} T^{-1} \boldsymbol{\nu}. \tag{4}$$

Now realize that $\theta T^{-1} + (1-\theta)\Sigma^{-1}$ is positive-definite; let $B^{\mathrm{T}} B$ be the corresponding Cholesky decomposition. As a next step, we perform the substitution $\boldsymbol{y} = B\boldsymbol{x}$, so that

$$\mathrm{d}x_1 \cdots \mathrm{d}x_n = |\, B^{-1} \,| \, \mathrm{d}y_1 \cdots \mathrm{d}y_n = \frac{1}{|\, \theta T^{-1} + (1-\theta)\Sigma^{-1} \,|^{1/2}} \, \mathrm{d}y_1 \cdots \mathrm{d}y_n.$$

Then Expression (4) can be rewritten as

$$\frac{\theta}{2} \log \frac{|\,\Sigma\,|}{|\,T\,|} - \frac{1}{2} \boldsymbol{y}^{\mathrm{T}} \boldsymbol{y} + \theta \boldsymbol{\nu}^{\mathrm{T}} T^{-1} B^{-1} \boldsymbol{y} - \frac{\theta}{2} \boldsymbol{\nu}^{\mathrm{T}} T^{-1} \boldsymbol{\nu},$$

which equals

$$\begin{aligned} &\frac{\theta}{2} \log \frac{|\,\Sigma\,|}{|\,T\,|} - \frac{1}{2} \left(\boldsymbol{y} - \theta (B^{-1})^{\mathrm{T}} T^{-1} \boldsymbol{\nu} \right)^{\mathrm{T}} \left(\boldsymbol{y} - \theta (B^{-1})^{\mathrm{T}} T^{-1} \boldsymbol{\nu} \right) \\ &\quad - \quad \frac{\theta}{2} \boldsymbol{\nu}^{\mathrm{T}} T^{-1} \boldsymbol{\nu} + \frac{\theta^2}{2} \boldsymbol{\nu}^{\mathrm{T}} T^{-1} \left(\theta T^{-1} + (1-\theta)\Sigma^{-1} \right)^{-1} T^{-1} \boldsymbol{\nu}. \end{aligned}$$

Recognizing a multivariate normal density, we conclude that the moment generating function $\mathbb{E}_0 \exp(\theta \mathscr{L}_n(\boldsymbol{X}))$ equals, with I_n denoting an $n \times n$ identity matrix,

$$\begin{aligned}\mathbb{E}_0 \exp(\theta \mathscr{L}_n(\boldsymbol{X})) &= \left(\frac{|\,\Sigma\,|}{|\,T\,|}\right)^{\theta/2} \frac{|\,\Sigma\,|^{-1/2}}{|\,\theta T^{-1} + (1-\theta)\Sigma^{-1}\,|^{1/2}} \\ &\quad \times \exp\left(-\frac{\theta}{2}\boldsymbol{\nu}^{\mathrm{T}} T^{-1}\boldsymbol{\nu} + \frac{\theta^2}{2}\boldsymbol{\nu}^{\mathrm{T}} T^{-1}\left(\theta T^{-1} + (1-\theta)\Sigma^{-1}\right)^{-1} T^{-1}\boldsymbol{\nu}\right) \\ &= \left(\frac{|\,\Sigma\,|}{|\,T\,|}\right)^{\theta/2} \frac{1}{|\,\theta T^{-1}\Sigma + (1-\theta) I_n\,|^{1/2}} \\ &\quad \times \exp\left(-\frac{\theta}{2}\boldsymbol{\nu}^{\mathrm{T}} T^{-1}\boldsymbol{\nu} + \frac{\theta^2}{2}\boldsymbol{\nu}^{\mathrm{T}} T^{-1}\left(\theta T^{-1} + (1-\theta)\Sigma^{-1}\right)^{-1} T^{-1}\boldsymbol{\nu}\right). \end{aligned} \tag{5}$$

The above analysis gives, in principle, a technique to calculate $\mathscr{I}(b)$, and hence, a technique to approximate the type I error probability. This allows us to determine the critical value b. In specific cases, the computations can be made more explicit. Below we treat two of those special cases. In Section 3.1 we work out the moment generating function (5) and find the Legendre transform (3) for a test designed to decide between two different means, while for the special case of independent data (5) is simplified in Section 3.2 .

3.1. **Special case I: difference in mean for dependent data.** In the first special case we focus on, there is only a difference in the means of the multivariate normal distributions, that is, the covariance matrix is left unchanged: $\Sigma = T$. It means that

$$\mathbb{E}_0 \exp(\theta \mathscr{L}_n(\boldsymbol{X})) = \exp\left(-\frac{\theta}{2}\boldsymbol{\nu}^{\mathrm{T}} T^{-1}\boldsymbol{\nu} + \frac{\theta^2}{2}\boldsymbol{\nu}^{\mathrm{T}} T^{-1}\boldsymbol{\nu}\right).$$

As a consequence — defining $\mathscr{I}_n(b) := n\mathscr{I}(b)$ — we have

$$\mathscr{I}_n(b) = \sup_\theta \left(n\theta b + \frac{\theta}{2}\boldsymbol{\nu}^{\mathrm{T}} T^{-1}\boldsymbol{\nu} - \frac{\theta^2}{2}\boldsymbol{\nu}^{\mathrm{T}} T^{-1}\boldsymbol{\nu}\right).$$

The supremum can be determined explicitly, leading to

$$\mathscr{I}_n(b) = \frac{(nb + \frac{1}{2}\boldsymbol{\nu}^{\mathrm{T}} T^{-1}\boldsymbol{\nu})^2}{2\boldsymbol{\nu}^{\mathrm{T}} T^{-1}\boldsymbol{\nu}}. \tag{6}$$

We will use this result in Section 4.1 to develop a changepoint detection test to find a change in the mean of a dependent (multivariate normal) sequence.

3.2. **Special case II: difference in mean and variance for independent data.** In the second special case we have that there is a difference in both mean and covariance matrix of the multivariate normal distributions, but in such a way that the covariance matrices Σ and T correspond to independent random variables. In this setting Σ is the diagonal matrix with the

vector $\boldsymbol{\sigma}^2$ on the diagonal (to be denoted by $\mathrm{diag}(\boldsymbol{\sigma}^2)$), while $T = \mathrm{diag}(\boldsymbol{\tau}^2)$. It is a matter of elementary calculus to verify that

$$\begin{aligned}\mathbb{E}_0 \exp(\theta \mathscr{L}_n(\boldsymbol{X})) &= \prod_{i=1}^n \left(\frac{\sigma_i}{\tau_i}\right)^\theta \times \prod_{i=1}^n \left(\theta \frac{\sigma_i^2}{\tau_i^2} + (1-\theta)\right)^{-1/2} \\ &\quad \times \exp\left(-\frac{\theta}{2}\sum_{i=1}^n \frac{\nu_i^2}{\tau_i^2} + \frac{\theta^2}{2}\sum_{i=1}^n \frac{\nu_i^2 \sigma_i^2/\tau_i^2}{\theta\sigma_i^2 + (1-\theta)\tau_i^2}\right). \end{aligned} \tag{7}$$

The above result is used in Section 4.2 for a test that detects a change in variance somewhere in a sequence of independent normally distributed data.

4. CHANGEPOINT DETECTION TESTS FOR DEPENDENT DATA

We now propose a series of changepoint detection tests, in line with the one presented for an i.i.d. sequence in [8, Ch. VI.E] (discussed in Section 2 of this paper). The idea is that H_0 corresponds to a model $\mathbb{P}_0$, whereas under H_1 there is a shift of the model $\mathbb{P}_0$ to $\mathbb{P}_1$ at the $(n\beta+1)$-th observation, for some $\beta \in [0,1)$ such that $n\beta$ is integer-valued. In line with [8, Ch. VI.E, Eq. (43)] we reject H_0 if

$$\max_{\beta\in[0,1)} \left(\frac{1}{n}\mathscr{L}_{n,\beta}(\boldsymbol{X}) - b(\beta)\right) := \max_{\beta\in[0,1)} \left(\frac{1}{n}\log\left(\frac{g_{n,\beta}(\boldsymbol{X})}{f_n(\boldsymbol{X})}\right) - b(\beta)\right) > 0, \tag{8}$$

where the density $g_{n,\beta}(\cdot)$ corresponds to H_1 with a change at time $n\beta+1$, and $b(\cdot)$ is a function specified below. Large-deviations theory enables us to compute

$$\lim_{n\to\infty} \frac{1}{n}\log \mathbb{P}_0\left(\max_{\beta\in[0,1)} \big(\mathscr{L}_{n,\beta}(\boldsymbol{X}) - b(\beta)\big) > 0\right),$$

using the machinery of Section 3. To optimize the type II error rate performance [8, Ch. VI.E, p. 113], $b(\cdot)$ should be chosen such that the decay rate satisfies

$$-\mathscr{I}(b(\beta)) = \lim_{n\to\infty} \frac{1}{n}\log \mathbb{P}_0\big(\mathscr{L}_{n,\beta}(\boldsymbol{X}) - b(\beta) > 0\big) = -\gamma \tag{9}$$

for a uniform positive γ, across all $\beta \in [0,1)$; this enables us to determine $b(\beta)$. In practice the observations arrive one by one and at every new observation X_m the changepoint detection test is then performed on the sequence of the n most recent observations $(X_{m-n+1}, \ldots X_m)$. An alarm is issued at time m if the test statistic $\mathscr{L}_{n,\beta}(\boldsymbol{X})$ exceeds the threshold $b(\beta)$ for any $\beta \in [0,1)$. The goal is to detect a changepoint as soon as possible, while at the same time keeping the number of false alarms limited. We explain the details of how to carry out the testing in more detail in the remainder of this section, and also provide numerical examples

in Section 5. In the following, we again use $X_1, \ldots, X_n$ to represent the observations of the current window (thus, dropping the enumeration of the windows by m).

Note that traditionally in changepoint detection the ARL — the expected time till the first false alarm — has been used to design procedures with a limited number of false alarms. However, the ARL criterion does not ensure that the number of false alarms is low for every window (see also the discussion in [19]), and furthermore, it may not always be applicable (see [25] for examples where the ARL becomes infinite). Our methodology in the current paper circumvents these issues.

It should be pointed out that the proposed procedure is essentially a (window-limited) CUSUM-type detection procedure. In [19] CUSUM is proven to be *asymptotically optimal* (as $\alpha \to 0$) in that it achieves the lowest possible detection delay provided that $\sup_{m\geq 1} \mathbb{P}_0(m \leq \tau \leq m+n) \leq \alpha$, where τ denotes the stopping time of the CUSUM-type procedure. It turns out that the distribution of τ is approximately exponential [36, Ch. 8]. Therefore, from the memoryless-property of the exponential distribution, we have that

$$\sup_{m\geq 1} \mathbb{P}_0(m \leq \tau \leq m-1+n) \approx \mathbb{P}_0(1 \leq \tau \leq n)\,.$$

Since we consider a window-limited detection procedure, where $\tau < n$ is not considered, we thus impose (9) rather than the criterion proposed by Lai in [19]. Furthermore, as (9) limits the false alarm probability for any given window, the resulting *average* false alarm rate (averaged over all windows that do not include the changepoint) will also be limited to the same level.

We now perform the computation of (9) and the determination of the critical function $b(\beta)$ for various specific models. In [8, Ch. VI.E Example 3] the critical function is determined for a change in mean in a sequence of independent normally distributed observations. In Section 4.1 we look at a change in mean somewhere in a (dependent) multivariate normal sequence (using the result of Section 3.1), in Section 4.2 we consider a change in variance for independent normally distributed sequences (using the result of Section 3.2) and Section 4.3 treats the case of a change in scale of a (dependent) multivariate normal sequence.

4.1. **Test 1: change in mean for dependent data.** In this section we show how to compute the critical function $b(\beta)$ when testing for a change in the mean of a dependent sequence. We derive an explicit expression for $b(\beta)$ for the case of autoregressive-moving-average (ARMA) processes.

We are in the setting that $\Sigma = T$, and that we want to detect a change in mean at some index $n\beta + 1$, for $\beta \in [0, 1)$. Without loss of generality we consider a change from mean 0 to some other value, say $\bar{\nu}$. In line with the above, we wish to find a function $b(\beta)$ such that (9) holds for $\beta \in [0, 1)$, for a given $\gamma > 0$. We can apply formula (6), with the first $n\beta$ entries of $\boldsymbol{\nu}$ equal to 0 and the last $n(1-\beta)$ equal to $\bar{\nu}$. Defining

$$t_{n,\beta} := \sum_{i=n\beta+1}^{n} \sum_{j=n\beta+1}^{n} (T^{-1})_{i,j},$$

we obtain

$$-\gamma = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}_0 \left(\frac{1}{n} \log \left(\frac{g_{n,\beta}(\boldsymbol{X})}{f(\boldsymbol{X})} \right) \geq b(\beta) \right) = -\mathscr{I}(b(\beta)) = -\lim_{n\to\infty} \frac{1}{2} \frac{(nb(\beta) + \frac{1}{2}\bar{\nu}^2 t_{n,\beta})^2}{n\bar{\nu}^2 t_{n,\beta}}.$$

As an example we could consider $\boldsymbol{X}$ corresponding to an *autoregressive process of order 1* (usually abbreviated to AR(1)). This is a stationary process (with mean c) obeying the recursion

$$X_i - c = \varrho(X_{i-1} - c) + \varepsilon_i,$$

where the ε_i s are i.i.d. samples from a zero-mean normal distribution with variance σ^2 (where we assume $|\varrho| < 1$). It is known that

$$T = \frac{\sigma^2}{1-\varrho^2} \begin{pmatrix} 1 & \varrho & \varrho^2 & \varrho^3 & \cdots & \varrho^{n-1} \\ \varrho & 1 & \varrho & \varrho^2 & \cdots & \varrho^{n-2} \\ \varrho^2 & \varrho & 1 & \varrho & \cdots & \varrho^{n-3} \\ \varrho^3 & \varrho^2 & \varrho & 1 & \cdots & \varrho^{n-4} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \varrho^{n-1} & \varrho^{n-2} & \varrho^{n-3} & \varrho^{n-4} & \cdots & 1 \end{pmatrix}.$$

It is elementary to verify that

$$T^{-1} = \frac{1}{\sigma^2} \begin{pmatrix} 1 & -\varrho & 0 & 0 & \cdots & 0 \\ -\varrho & 1+\varrho^2 & -\varrho & 0 & \cdots & 0 \\ 0 & -\varrho & 1+\varrho^2 & -\varrho & \cdots & 0 \\ 0 & 0 & -\varrho & 1+\varrho^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{pmatrix}.$$

It follows that (realizing that there are roughly n diagonal entries of value $1+\varrho^2$, and that there are roughly $2n$ entries of value $-\varrho$ above and below the diagonal),

$$\lim_{n\to\infty} \frac{t_{n,\beta}}{n(1-\beta)} = \frac{1}{\sigma^2}\left(1\cdot(1+\varrho^2)+2\cdot(-\varrho)\right) = \left(\frac{1-\varrho}{\sigma}\right)^2,$$

and hence

$$b(\beta) = \bar{\nu}\left(\frac{1-\varrho}{\sigma}\right)\sqrt{2\gamma(1-\beta)} - \frac{1}{2}\bar{\nu}^2\left(\frac{1-\varrho}{\sigma}\right)^2(1-\beta). \tag{10}$$

Compared to the function $b(\beta)$ that was derived for the unit-variance i.i.d. case [8, Ch. VI.E, p. 113], $\bar{\nu}$ needs to be replaced by $\bar{\nu}(1-\varrho)/\sigma$, in order to account for the dependence between the observations, and the value of the variance. For $\varrho = 0$ and $\sigma^2 = 1$, the two functions obviously match.

Also in case that T^{-1} cannot be computed explicitly, we can still find the limiting value of $t_{n,\beta}/(n(1-\beta))$. We now consider the general ARMA(p,q) model, defined as a stationary model with mean value c obeying [6]

$$X_i - c = \varepsilon_i + \sum_{j=1}^{p}\varrho_j(X_{i-j}-c) + \sum_{j=1}^{q}\vartheta_j\varepsilon_{i-j}, \tag{11}$$

for $p, q \in \mathbb{N}$, where we assume that the roots of the AR polynomial lie outside the unit circle. Again we assume that the ε_i are i.i.d. samples from a zero-mean normal distribution with variance σ^2.

The following lemma implies that the limiting value of $t_{n,\beta}/(n(1-\beta))$ does not depend on β, or, put differently, that $t_{n,\beta}$ grows essentially linear in $n(1-\beta)$; cf. [29, Eq. (9)].

Lemma 1. *For $\boldsymbol{X}$ obeying an* ARMA*(p,q) model, and $\beta \in [0,1)$,*

$$\mathscr{T}_\beta := \lim_{n\to\infty}\frac{t_{n,\beta}}{n(1-\beta)} = \left(\frac{1-\sum_{j=1}^{p}\varrho_j}{\sigma\left(1+\sum_{j=1}^{q}\vartheta_j\right)}\right)^2 =: \mathscr{T}.$$

The proof can be found in Appendix A. The immediate consequence of the lemma is that

$$-\gamma = \lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}_0\left(\frac{1}{n}\log\left(\frac{g_{n,\beta}(\boldsymbol{X})}{f(\boldsymbol{X})}\right)\geq b(\beta)\right) = -\frac{1}{2}\frac{\left(b(\beta)+\frac{1}{2}\bar{\nu}^2\mathscr{T}(1-\beta)\right)^2}{\bar{\nu}^2\mathscr{T}(1-\beta)},$$

and

$$b(\beta) = \bar{\nu}\sqrt{2\mathscr{T}\gamma(1-\beta)} - \frac{1}{2}\bar{\nu}^2\mathscr{T}(1-\beta). \tag{12}$$

We have seen that for AR(1) processes $\mathscr{T} = ((1-\varrho)/\sigma)^2$. From Lemma 1 it follows that for an MA(1) process with parameter ϑ it holds that $\mathscr{T} = 1/(\sigma(1+\vartheta))^2$ and

$$b(\beta) = \bar{\nu}\left(\frac{1}{\sigma(1+\vartheta)}\right)\sqrt{2\gamma(1-\beta)} - \frac{1}{2}\bar{\nu}^2\left(\frac{1}{\sigma(1+\vartheta)}\right)^2(1-\beta). \tag{13}$$

4.2. **Test 2: change in variance for independent data.** We now consider the case in which there is no change in mean, where under H_0 all observations are independent and normally distributed with variance σ^2 while under H_1 the variance changes from σ^2 to τ^2 at some specific moment. We set $\boldsymbol{\nu} = \mathbf{0}$, $\Sigma = \sigma^2 I_n$, and T is an $n \times n$ diagonal matrix with σ^2 at the first $m = \beta n$ diagonal positions ($\beta \in [0,1)$), and τ^2 at the other diagonal positions. Note that this corresponds to a change in variance at time $\beta n + 1$. Filling out (7), we get

$$\begin{aligned}\Lambda_\beta(\theta) &:= \frac{1}{n}\log \mathbb{E}_0 \exp(\theta \mathscr{L}_n(\boldsymbol{X})) \\ &= \theta(1-\beta)\log\frac{\sigma}{\tau} + \frac{1}{2}(1-\beta)\log\tau^2 - \frac{1}{2}(1-\beta)\log\left(\theta\sigma^2 + (1-\theta)\tau^2\right)\end{aligned}$$

Now let us compute $\mathscr{I}(b(\beta)) = \sup_\theta(\theta b(\beta) - \Lambda_\beta(\theta))$. Writing $A_1 + A_2\theta = \theta\sigma^2 + (1-\theta)\tau^2$, the optimizing θ satisfies

$$b(\beta) = (1-\beta)\left(\log\frac{\sigma}{\tau} - \frac{\frac{1}{2}A_2}{A_1 + A_2\theta}\right),$$

which can be solved, giving

$$\theta = -\frac{\frac{1}{2}(1-\beta)}{b(\beta) - (1-\beta)\log(\sigma/\tau)} - \frac{\tau^2}{\sigma^2 - \tau^2},$$

so that $b(\beta)$ can be evaluated numerically from

$$\gamma = (1-\beta)\left(-\frac{1}{2} - \frac{\tau^2}{\sigma^2-\tau^2}\left(\frac{b(\beta)}{1-\beta} - \log\frac{\sigma}{\tau}\right) - \frac{1}{2}\log\left(\frac{-2\tau^2}{\sigma^2-\tau^2}\left(\frac{b(\beta)}{1-\beta} - \log\frac{\sigma}{\tau}\right)\right)\right). \tag{14}$$

4.3. **Test 3: change in scale for dependent data.** We now consider the more general situation in which the typical deviations of the process are inflated by a factor f. This type of change has applications in the context of communication networks; for details we refer to [21]. More specifically, we concentrate on the case we have that after time $n\beta$ the mean $\bar{\mu}$ changes into $f\bar{\mu}$, while the covariance matrix becomes $f^2\Sigma$. Again, we can shift space so that the first $n\beta$ entries of the alternative mean $\boldsymbol{\nu}$ equal 0 and the last $n(1-\beta)$ equal $\bar{\nu} = f\bar{\mu} - \bar{\mu}$. We suppose that $\boldsymbol{X}$ corresponds to a stationary sequence of random variables with possibly 'weak dependence' (as defined in [7, Ch. IV]); ARMA(p,q) processes fall in this class. In this section, we assume that the change is introduced abruptly. By this we mean that the memory of observations

is not kept after the change which thus results in a new stationary process that is independent from the process before the change. Because of this, the statistic $\mathscr{L}_{n,\beta}(\boldsymbol{X})$ of (2) becomes $\mathscr{L}_{n,\beta}(\check{\boldsymbol{X}}) = \log\left[g_{n,\beta}(\check{\boldsymbol{X}})/f_n(\check{\boldsymbol{X}})\right]$, where $\check{\boldsymbol{X}} = (X_{n\beta+1},\ldots,X_n)$. This, using the notation of Section 3, reduces to

$$\begin{aligned}\frac{1}{2}\log|\Sigma_{n(1-\beta)}| - \frac{1}{2}\log f^{2n(1-\beta)}|\Sigma_{n(1-\beta)}| + \frac{1}{2}\check{\boldsymbol{X}}^{\mathrm{T}}\Sigma^{-1}_{n(1-\beta)}\check{\boldsymbol{X}} \\ -\frac{1}{2f^2}(\check{\boldsymbol{X}}-\boldsymbol{\nu}_{n(1-\beta)})^{\mathrm{T}}\Sigma^{-1}_{n(1-\beta)}(\check{\boldsymbol{X}}-\boldsymbol{\nu}_{n(1-\beta)}) \\ = \quad -n(1-\beta)\log f + \frac{1}{2}\check{\boldsymbol{X}}^{\mathrm{T}}\Sigma^{-1}_{n(1-\beta)}\check{\boldsymbol{X}} - \frac{1}{2f^2}(\check{\boldsymbol{X}}-\boldsymbol{\nu}_{n(1-\beta)})^{\mathrm{T}}\Sigma^{-1}_{n(1-\beta)}(\check{\boldsymbol{X}}-\boldsymbol{\nu}_{n(1-\beta)}).\end{aligned}$$

Using (5), it is not hard to verify that the moment generating function $\mathbb{E}_0\exp(\theta\mathscr{L}_{n,\beta}(\check{\boldsymbol{X}}))$ of our test statistic equals

$$f^{-\theta(1-\beta)n}\left(\sqrt{\theta/f^2+(1-\theta)}\right)^{-(1-\beta)n}\times\exp\left(-\frac{\theta s_{n,\beta}}{2f^2}\bar{\nu}^2+\frac{\theta^2 s_{n,\beta}}{2(\theta f^2+(1-\theta)f^4)}\bar{\nu}^2\right),$$

with

$$s_{n,\beta} := \sum_{i=n\beta+1}^{n}\sum_{i=n\beta+1}^{n}(\Sigma^{-1})_{i,j},$$

where we recall that $s_{n,\beta}$ is essentially linear in n and thus the limiting log-moment generating function exists. The standard machinery now enables us to derive $b(\beta)$.

A simplification can be made in case $\bar{\nu}=0$. This situation occurs when there is no change in mean, while the covariance matrix is multiplied by f^2. Then $b(\beta)$ follows from

$$\gamma = \mathscr{I}(b(\beta)) = \sup_\theta\left(\theta b(\beta)+\theta(1-\beta)\log f+\frac{1-\beta}{2}\log\left(\frac{\theta}{f^2}+(1-\theta)\right)\right).$$

The optimizing θ is

$$-\left(\frac{\frac{1}{2}(1-\beta)}{b(\beta)+(1-\beta)\log f}+\frac{1}{1/f^2-1}\right),$$

so that $b(\beta)$ can be evaluated numerically from

$$\gamma = (1-\beta)\left(-\frac{1}{2}-\frac{1}{1/f^2-1}\left(\frac{b(\beta)}{1-\beta}+\log f\right)-\frac{1}{2}\log\left(\frac{-2}{1/f^2-1}\left(\frac{b(\beta)}{1-\beta}+\log f\right)\right)\right). \quad (15)$$

Note that the last equation of Section 4.2 follows directly from the above equation when f is replaced by τ/σ.

5. Numerical evaluation

In Section 4 we have developed changepoint detection tests for dependent sequences. In this section, we evaluate the performance of our proposed method. To this end, we perform

a number of simulation experiments. This set-up facilitates evaluating the sensitivity of the procedure, as it enables us to assess its performance in a broad range of scenarios, both in terms of the underlying model, and in terms of the type of change that has taken place in the sequence of observations (in relation to the type of change the sequence is tested against).

We start by explaining the 'basic experiment', various variations of which are studied throughout this section. In the basic experiment we simulate an ARMA process with a change from mean 0 to mean 3 and apply the changepoint detection test of Section 4.1. (A numerical evaluation for the change in scale test of Section 4.3 was carried out in [21].) More specifically, in the basic experiment we carry out the following procedure:

- In every run we simulate a stationary AR(1) or MA(1) time series of length 200 that obeys the recursion given in (11) with mean $c = 0$ up to observation 99 and mean $c = 3$ afterwards, thus having a changepoint at observation 100. The standard deviation σ of the ε_i is set to 1.[1]
- We then consider windows of size 50 that we shift along the time series and we test each window for a change in mean. Thus, for window 1 we test observations 1 up to 50 for a changepoint, for window 2 we test observations 2 up to 51 for a changepoint and we continue this procedure up to window 151 which consists of observations 151 up to 200. Note that the first window in which the changepoint is contained is window number 51.
- In order to test for a change in mean within a certain window, we determine whether Inequality (8) holds true. To this end, first, the test statistic $\mathscr{L}_{50,\beta}(\boldsymbol{X}) = \log\left[g_{50,\beta}(\boldsymbol{X})/f_{50}(\boldsymbol{X})\right]$ is computed according to (2). Here ν_i is 0 for $i < 100$ and ν_i is 3 for $i \geq 100$, the covariance matrix $\Sigma = T$ of an ARMA process is computed using the algorithm developed in [24] and $\boldsymbol{X}$ is simulated as described above. Second, the threshold function $b(\beta)$ is computed using (10) for an AR(1) and (13) for an MA(1) process. The significance level α is put to 0.01, so that γ in these equations can be found from $e^{-50\,\gamma} = 0.01$. Third, we calculate $\frac{1}{50}\mathscr{L}_{50,\beta}(\boldsymbol{X}) - b(\beta)$ for $\beta = \frac{i}{50}, i = 0, \ldots, 49$. If the maximum of this difference (taken over β) is bigger than zero, we raise an alarm. Otherwise we conclude that there is no changepoint in the current window. We repeat this step for all windows. All the steps above are repeated 300 times.

[1] In this experiment — consistent with the assumptions in Section 4.1 — the memory $X_{100-1}, \varepsilon_{100-1}$ is used as the initial condition for the observation after the change. The transition from the original to the changed process is therefore smooth — as opposed to the abrupt change assumed in Section 4.3.

- As soon as we know for each window whether an alarm is raised or not, the performance of the test is evaluated by the following metrics.
  - For every window number the *alarm ratio* is calculated as the number of alarms for that window in 300 runs divided by 300. Note that the alarm ratio for the windows 1 up to 50 gives the *false alarm ratio* per window while for the windows 51 up to 151 it gives the *detection ratio*.
  - The *detection delay* is calculated as the time of detection minus the true changepoint. We define the time of detection as the number of the first observation for which we know that a change has happened, that the last observation of the first window in which an alarm was raised after the changepoint occured. For instance, if the changepoint is first detected at time 104 (i.e. the first alarm after the change is raised for window number 55), the delay is 4. We repeat this procedure 300 times, and take the mean of the detection delay over the runs.

In the next two sections we discuss the results of the above described experiment, focusing on the alarm ratio in Section 5.1 and on the detection delay in Section 5.2. In Section 5.3 we compare the performance of the test for different sizes of the mean shift in order to assess how small of a change in the mean value can be detected. We also examine the sensitivity to the alternative mean chosen in the test setup. We do so by evaluating the performance when testing against a change in mean that is larger than the change we simulate.

We remark that our straightforward implementation of the procedure in Matlab was executed in 0.1 ms per window. At the same time, in practice a new window will probably be considered only after aggregating a reasonable amount of traffic (which could even be in the order of minutes) in a time bin. In that case 0.1 ms (or even several seconds) of calculation time is fast enough to qualify it as (quasi) on-line. Further improvements can be achieved, for example, by using approximations for the inverse covariance matrix (see, e.g., [30]).

5.1. **Alarm ratio.** In this section we analyze the performance of our changepoint detection method by calculating the ratio of (false) alarms as defined above. We will see that for practically relevant coefficients of the AR(1) and MA(1) processes, the number of false alarms is low. For those coefficients that correspond to a high number of false alarms we explain the reason and describe ways to improve the results.

As examples we consider an AR and an MA process both with coefficient 0.5, see Figs. 1– 2. The dots depict the alarm ratios that we obtained, while the vertical line highlights the earliest window where we could have detected the changepoint.

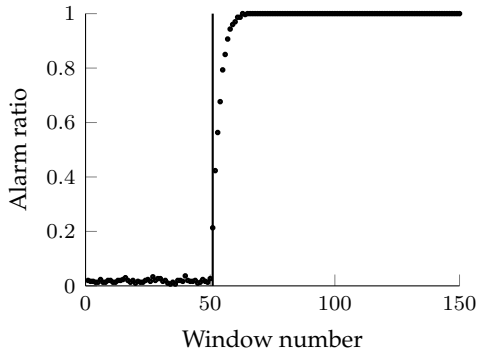

FIGURE 1. Alarm ratio per window for an AR(1) with coefficient 0.5 and a changepoint at observation 100.

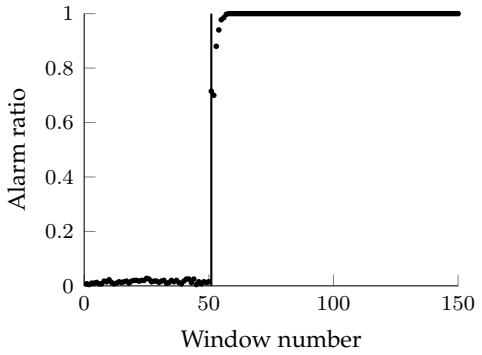

FIGURE 2. Alarm ratio per window for an MA(1) with coefficient 0.5 and a changepoint at observation 100.

The picture reveals that we have very few false alarms, their ratio being in the order of 0.01 (as intended since we chose a significance level of 0.01). At the same time, we have achieved the desirable property that the changepoint is detected almost instantly; there is only a small delay. It is noted that MA(1) processes fluctuate more frequently than AR(1) processes; this may explain the fact that the changepoint is detected earlier for MA(1) than for AR(1) when both have coefficient 0.5. We come back to the detection delay in Section 5.2.

Above we put the coefficients of the MA(1) and AR(1) processes equal to 0.5. Now, we want to compare false alarm ratios for a range of different coefficients. To that end we take the mean of the alarm ratios up to the first window where the changepoint is visible; thus, including only windows where every alarm is a false alarm. In this way we obtain Fig. 3, which shows that for coefficients between -0.3 and 0.6 we obtain an excellent performance in terms of false alarms. The cases for which the method does not perform well yet can be improved; later in this section we point out how the procedure can be adapted to obtain the improved curve shown in Fig. 4. Furthermore, we remark that the proposed method does not have to be used as the only detector but rather can be combined with some other sensors in the effort to reduce the false alarm rate to the acceptable level.

We now provide an intuitive explanation as to *why* our testing procedure tends to perform inadequately for specific parameter values, as we observed in Fig. 3. It turns out that the limiting value of $t_{n,\beta}/(n(1-\beta))$, as given in Lemma 1, is approached slowly for negative coefficients, especially when β is big. This effect is illustrated in Figs. 5–6 below, where n is plotted against the difference of $t_{n,\beta}/n(1-\beta)$ and the corresponding limit value. As examples we chose a process that showed a good test performance in terms of false alarms (viz. an AR(1)

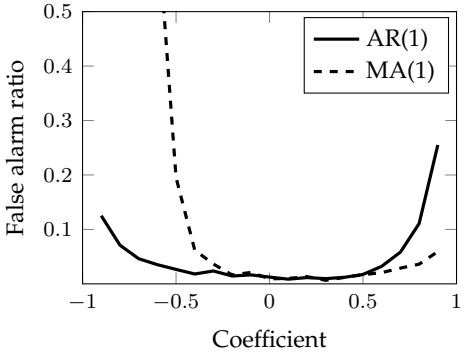

FIGURE 3. False alarms for a range of different coefficients, basic experiment.

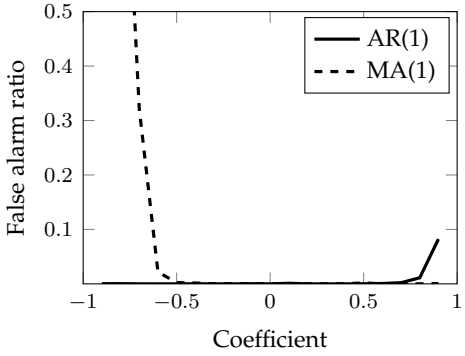

FIGURE 4. False alarms for a range of different coefficients, adjusted experiment.

with coefficient 0.5) in Fig. 5, as well as a process with a very high false alarm rate (viz. an MA(1) with coefficient -0.9) in Fig. 6.

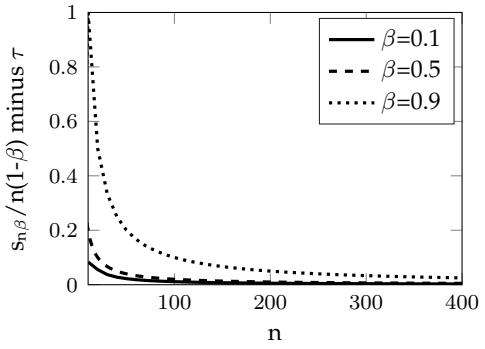

FIGURE 5. Difference of $t_{n,\beta}/(n(1-\beta))$ and $\mathscr{T}$ for an AR(1) with coefficient 0.5.

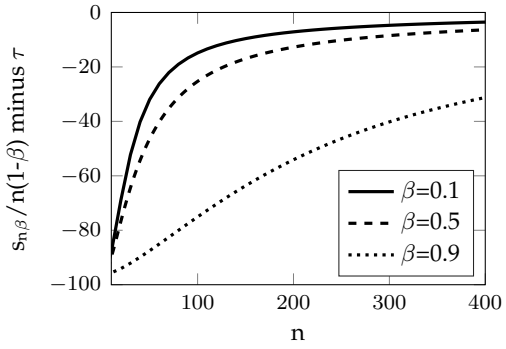

FIGURE 6. Difference of $t_{n,\beta}/(n(1-\beta))$ and $\mathscr{T}$ for an MA(1) with coefficient -0.9.

We conclude from Figs. 5–6 that for the negatively correlated MA process we are still far away from the limiting value when n is 400, while for the AR process the limiting value is approximated reasonably well already when n is 50 (which corresponds to the chosen window size of 50).

In case we do want to handle processes with a high negative correlation we can improve the false alarm rate by adapting our procedure as described in the following paragraphs. As a leading example we consider an MA(1) process with coefficient -0.6 (see Fig. 7). One obvious possibility to control the number of false alarms is to lower the significance level α (see Fig. 8).

We can further improve the performance of our testing procedure in terms of false alarms by using a concept similar to the 'tuning procedure' proposed in [23, Section 5]. The main idea behind it is the following. We observed that most false alarms were raised because of a suspected changepoint at the *end* of the window, that is, for large β. (This problem is well

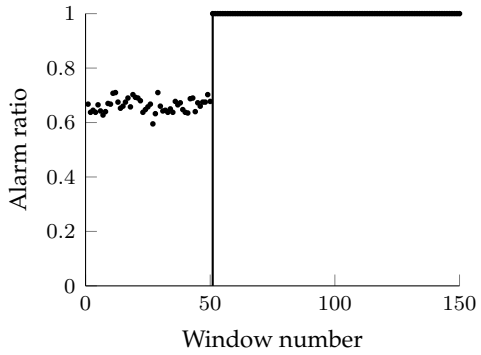

FIGURE 7. Alarm ratio per window for an MA(1) with coefficient -0.6, $\alpha = 0.01$.

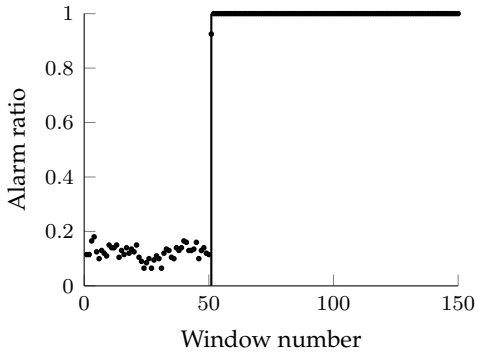

FIGURE 8. Alarm ratio per window for an MA(1) with coefficient -0.6, $\alpha = 0.0001$.

known for LLR tests, see [11]). A simple method to reduce the false alarm rate substantially is to ignore changepoints that correspond to β larger than, say, 0.95 (see Fig. 9); we call this adaptation 'tuning'. Note that even though we observed that most false alarms occur at the end of the window, tuning also neglects 'real' changepoints if they correspond to $\beta > 0.95$, and can therefore cause a delayed detection. However, the graph indicates that in the case of an MA(1) with coefficient -0.6 this approach works remarkably well.

Fig. 10 shows that we obtain an even better result if we in addition increase the window size to 100.[2]

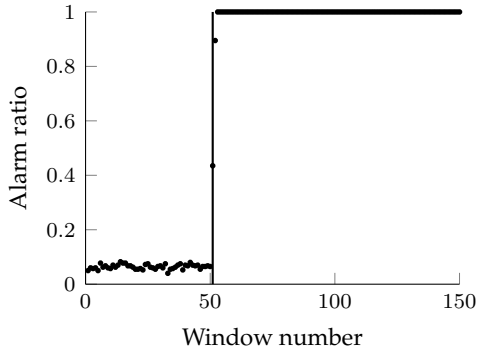

FIGURE 9. Alarm ratio per window for an MA(1) with coefficient -0.6, $\alpha = 0.0001$, when tuning is applied and the window size is 50.

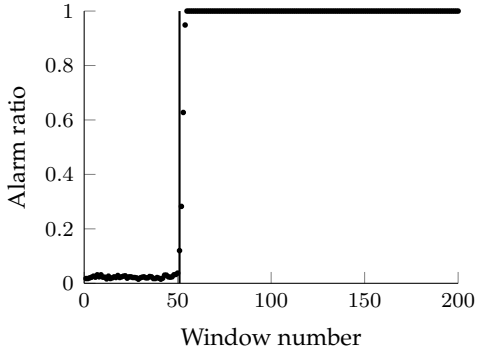

FIGURE 10. Alarm ratio per window for an MA(1) with coefficient -0.6, $\alpha = 0.0001$, tuning is applied, the window size is 100.

Using these three adjustments — that is: (i) a lower significance level of $\alpha = 0.0001$, (ii) application of tuning, and (iii) a larger window of length 100 — the false alarm performance is substantially better for most coefficients; compare Fig. 4 with Fig. 3. However, for MA(1) processes with a very high negative correlation (close to -1, that is) the window size of 100 is still

[2]To account for the larger window size, in this figure the length of the time series is 300 and the change takes place at time 150.

too small — as can be expected from Fig. 6. In all other cases the false alarm rate is now close to zero.

Note that improving the false alarm rate can lead to a lower detection ratio. However, considering the alarm ratios after the changepoint in Figs. 7–10, it is seen that the negative impact of the above adjustments is minor. In some cases a small additional detection delay is introduced, but we always detect the changepoint even when we apply the adjustments. We will see in Section 5.2 that the negative impact on the delay is smallest for very negative MA coefficients, which is exactly the case in which we have the largest number of false alarms (see Fig. 3), and hence for which the adjustments are most needed. Of course, these results depend also on the magnitude of the new mean after the changepoint. When the mean after the changepoint is large, the adjustment settings can be applied more generally, because the delay decreases (see Section 5.3).

5.2. **Detection delay.** After having evaluated how many false alarms are raised before the change, we now wish to assess how fast a changepoint is detected once it occurred. We will see that the delay is low for most AR and MA coefficients. When using the adjusted settings (to decrease the false alarm ratio), the delay increases, but is still quite low for negative coefficients and very low for MA processes with a very negative coefficient. However, using the adjusted settings for positively correlated processes, increases the detection delay significantly.

In Fig. 11 we plot the detection delay, which we define as the difference of the detection time and the true changepoint. We do so for a range of different coefficients of the AR and MA processes. For comparison, we have included the delays resulting from testing with a single value threshold that was chosen by simulation in such a way that the false alarm rate (approximately) equals the false alarm rate obtained in Fig. 3. Fig. 11 confirms that the changepoint is detected almost immediately for most coefficients. The larger delay for the experiment with simulation-based threshold indicates that a single value threshold can be inferior to a threshold function.

Fig. 11 also demonstrates that we detect the changepoint earlier for coefficients that correspond to a higher false alarm ratio. A notable exception is the case of an AR(1) process with a large positive coefficient where both the false alarm ratio (recall Fig. 3) and the detection delay are larger. AR(1) processes with a high positive correlation tend to behave rather erratically. Therefore, the change is visible later, and moreover, larger jumps have to be tolerated. As an example we may look at a realization of an AR(1) process with coefficient 0.9, with a large change from mean 0 to mean 5 at observation 100. The first alarm after the changepoint

is raised at window 56, meaning that we locate the changepoint at observation 105. This delay is in line with Fig. 13; actually, by just looking at the process, it is not clear where to locate the changepoint.

When using the adjusted settings, we detect the changepoint later (compare Fig. 11 to Fig. 12). When the mean after the change is 3, in the AR case the alarm is raised about 4 up to 5 observations late for negative and small positive coefficients. For bigger AR coefficients the delay increases sharply. In case of an MA process and a change in mean of 3 we are between 4 and 6 observations late for coefficients larger than -0.3. For smaller coefficients, the delay is smaller. In short, the adjusted settings have fewest impact on the detection delay for very negative MA coefficients while the impact is high for very positive AR coefficients.

We will see in Section 5.3 that when the mean after the change is larger, overall the detection delay decreases and thus the negative impact of using the adjusted settings is smaller. When exactly to apply the adjusted settings depends on the requirements on the false alarm ratio and the detection delay, which differ from application to application. In general, the settings are suited to MA processes with a very negative coefficient and to negatively correlated AR processes or positively correlated MA processes when the change in mean is large (much larger than the standard deviation). When applying the adjusted settings, one should be aware of an increased detection delay for positively correlated AR processes.

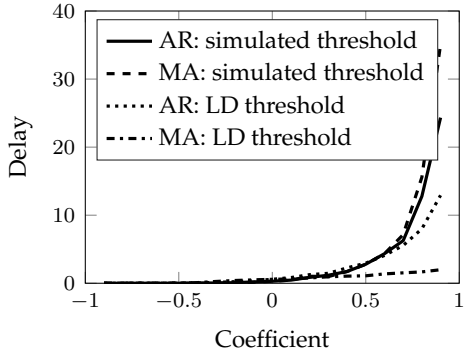

FIGURE 11. Detection delay, basic experiment, change to mean 3; as well as delays obtained with a simulation-based threshold.

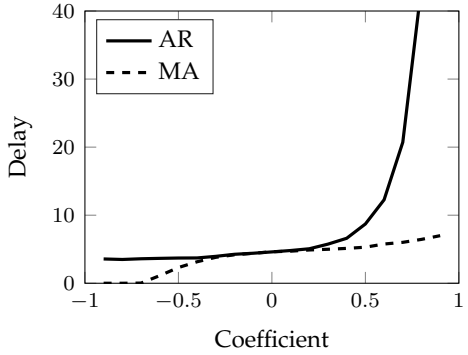

FIGURE 12. Detection delay, adjusted experiment, change to mean 3.

5.3. **Sensitivity analysis.** In the above experiments, we chose a shift size $\bar{\nu}$ and assessed the test's performance for this shift. In the current section we analyze how this performance (in terms of false alarms and detection delay) is affected by the specific value of $\bar{\nu}$. We will see that – in accordance with our intuition – the delay decreases when the change in mean is larger.

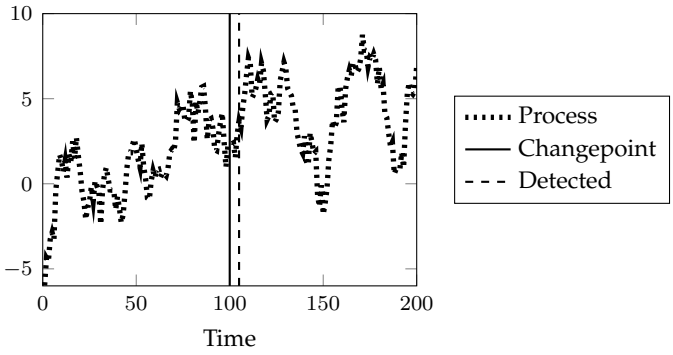

FIGURE 13. Realization of an AR(1) process with coefficient 0.9 and a change from mean 0 to 5 at observation 100.

This may allow us to apply the adjusted settings introduced in Section 5.1 more generally when the change in mean is large. For the most relevant scenarios (with moderate correlation), the performance in terms of false alarms is good for a broad range of values of $\bar{\nu}$.

In addition, in our experiments so far, we ran tests in which the mean after the changepoint coincided with the mean we test for. Of course, we would like to have some 'robustness'; for that reason we also study in this section the test's performance in case the mean after the changepoint differs from the one that we test for. It turns out that, except for very high positive correlations, the tests are robust against a smaller change than tested for; the detection delay increases slowly when the simulated change becomes smaller.

► *Varying the size of the change, testing for the mean that we simulated*. We run the basic experiment, but now we vary the size of the mean shift. Importantly, in these experiments the mean after the changepoint coincides with the mean we test for. Figs. 14–17 describe the tradeoff between an early detection and a low false alarm ratio. As expected, we see that in general it holds that how bigger the change in mean, the smaller the detection delay. The results for the false alarm ratio are somewhat more complicated:

- For large positive coefficients, we note that the larger the mean the *lower* the number of false alarms. It seems logical that a shift in mean is harder to detect as long as this shift is within the range of the fluctuations typical for the unchanged process. Accordingly, the further $\bar{\nu}$ exceeds this range the less false alarms we obtain.
- Surprisingly, for very negative coefficients we see that the opposite: the larger the mean, the *higher* the number of false alarms. For an MA process, the false alarm ratio increases much more sharply than for an AR process. To understand this recall that the limit value $\mathscr{T}$ of $t_{n,\beta}/n(1-\beta)$ from Lemma 1 is used to compute the threshold function in (12). As we saw from Fig. 6, for negative MA coefficients $\mathscr{T}$ is substantially *larger*

than $t_{n,\beta}/n(1-\beta)$ when n is small. This, in combination with $\bar{\nu} > 1$, makes the threshold function more negative than it should be — the larger $\bar{\nu}$, the more pronounced this effect.

- When the AR or MA coefficient is close to zero, neither of the above described effects has a strong impact and the false alarms are systematically low in this case.

To summarize, what we have seen is that — as we expected — detection gets easier as the mean after the change $\bar{\nu}$ increases. As long as the mean is larger than, say 1 or 1.5 (one or one and halve times the standard deviation of the process), the delay seems acceptable. Concerning the false alarm ratio we have that, for the most relevant case of moderate correlations (AR and MA coefficients close to zero), the false alarm ratio is low (close to the target of 0.01) for all $\bar{\nu}$. For highly positively correlated processes the ratio of false alarms is low enough if the change in mean is reasonably large (at least 3, i.e. much larger than the standard deviation of the process). When the correlation is highly negative, the false positive ratio is only low for AR processes with a small change in mean (close to the standard deviation). However, the performance of negatively correlated (AR with large mean change and MA) processes can be improved by using the adjustment settings introduced in Section 5.1.

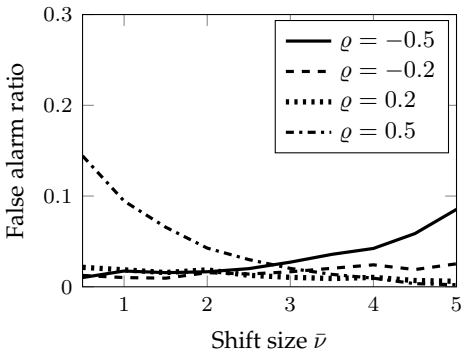

FIGURE 14. False alarms for different sizes of the mean shift, AR case.

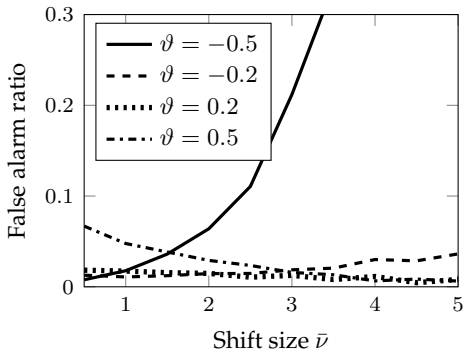

FIGURE 15. False alarms for different sizes of the mean shift, MA case.

► *Varying the simulated change in mean, while testing for mean 5.* We now again vary the simulated mean after the changepoint, but keep the mean that we use in the test setup fixed at 5.

We would expect false alarm rates not to be affected when varying the simulated mean after the changepoint, because false alarms occur before the changepoint. Indeed, we obtain false alarm rates that remain constant for the means we simulated. For coefficients ≥ -0.3, the false alarm ratio is close to 0.01, as we aimed for. Consistently with the earlier results, the false alarm ratio is higher for very high coefficients.

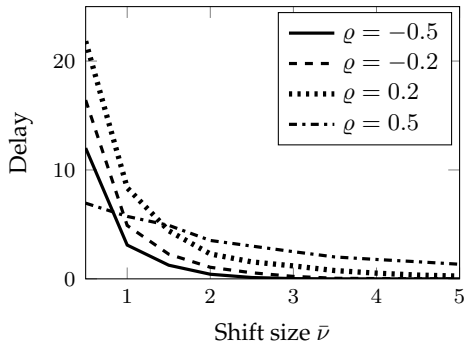

FIGURE 16. Detection delay for different sizes of the mean shift, AR case.

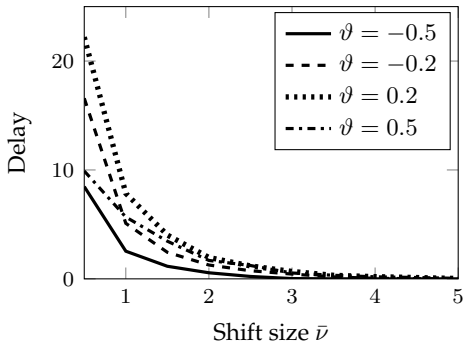

FIGURE 17. Detection delay for different sizes of the mean shift, MA case.

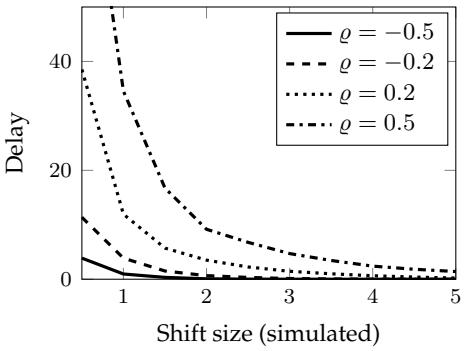

FIGURE 18. Detection delay for different simulated mean shifts in the AR case. Always test with shift size $\bar{\nu} = 5$.

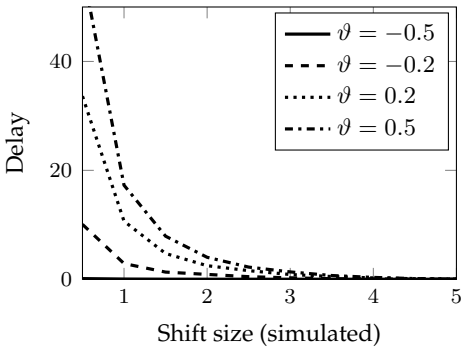

FIGURE 19. Detection delay for different simulated mean shifts in the MA case. Always test with shift size $\bar{\nu} = 5$.

We expect the detection delay to increase for a wrongly specified test, where the mean we test for is larger than the actual change. Figs. 18–19 show that the simulated results correspond to this expectation. Nevertheless, it turns out that a change in mean smaller than specified in the test, is tolerated quite well, particularly when the AR or MA coefficient is small.

6. DISCUSSION AND CONCLUDING REMARKS

In this paper we have developed CUSUM-type changepoint detection tests for dependent Gaussian data sequences. The paper includes the setting in which the underlying dataset follows an ARMA structure, a versatile class of models that has been frequently used to describe traffic streams (and other networking related time series). The changepoint tests consist of a log-likelihood test statistic in the spirit of CUSUM, and the corresponding threshold derived from a large-deviations approximation to the false alarm probability. In the literature such LD-based CUSUM-type tests have so far predominantly focused on procedures for detecting a change in mean in a sequence of independent observations. We have extended the application

of this type of test to the case of detecting (1) a change in mean in correlated normal data, (2) a change in variance in independent normal data and (3) a change in scale (that is, the process blows up by a factor) in correlated normal data. Furthermore, the false alarm criterion we employed ensures that the false alarm rate is low for every given window, thus allowing for a low variability of the number of false alarms.

We have demonstrated our changepoint detection test in a number of examples where we tested AR(1) and MA(1) processes against a change in mean. These simulations have shown that the test performs well (in terms of false alarm ratio and detection delay) for AR(1) and MA(1) coefficients between -0.3 and 0.6, as long as the change in mean is larger than the standard deviation of the process. In case of a strong negative correlation or a large change in mean, adaptation of the test settings is possible to further reduce the number of false alarms with minor negative influence on the detection delay. Moreover, the test performance seems to be rather resilient with respect to misspecification of the change size (as used in the test set-up).

Various next steps could be thought of. A detailed (empirical) comparison to the performance that is achieved under the ARL criterion is in place. Further, the tests should be modified such that they can be applied to detect a change in the correlation structure within a data sequence. Moreover, other light-tailed distributions may be considered.

APPENDIX A. PROOF OF LEMMA 1

We first study $v(n) := \mathbb{V}\mathrm{ar}\, S_n$, with $S_n = X_1 + \cdots + X_n$. It follows that

$$S_n - nc = \sum_{i=1}^{n} \varepsilon_i + \sum_{i=1}^{n} \sum_{j=1}^{p} \varrho_j \left(X_{i-j} - c\right) + \sum_{i=1}^{n} \sum_{j=1}^{q} \vartheta_j \varepsilon_{i-j}.$$

From this point on we take, without loss of generality, $c = 0$. Recognizing S_n in the right-hand side, bringing all terms involving S_n to the left-hand side, and taking the variance of both sides, it is now elementary to show that

$$\frac{v(n)}{n} \to \left(\frac{\sigma \left(1 + \sum_{j=1}^{q} \vartheta_j\right)}{1 - \sum_{j=1}^{p} \varrho_j} \right)^2 ; \tag{16}$$

this identity can alternatively be deduced relying on the spectral density formula for ARMA processes [29].

Based on 'Gärtner-Ellis', with $\pi_n := \mathbb{P}_0 \left(S_n \geq n\right)$,

$$\lim_{n \to \infty} \frac{1}{n} \log \pi_n = -\frac{1}{2s^2},$$

where s^2 is the limiting value of $v(n)/n$ (which we assume to exist). On the other hand, based on (a discrete-skeleton version of) 'Schilder' [22, Section 4.2], recalling that $T \equiv T_n$ is the covariance matrix of the X_i,

$$\lim_{\varepsilon\downarrow 0}\lim_{n\to\infty}\frac{1}{n}\log\pi_n(\varepsilon) = -\frac{1}{2}\lim_{n\to\infty}\frac{1}{n}\cdot \mathbf{1}T_n^{-1}\mathbf{1} = -\frac{1}{2}\mathscr{T}_0,$$

with $\pi_n(\varepsilon) := \mathbb{P}_0\left(\forall i \in \{1,\dots,n\} : S_i \in (i(1-\varepsilon), i(1+\varepsilon)), S_n \geq n\right)$. We want to prove that

$$\lim_{n\to\infty}\frac{1}{n}\log\pi_n = \lim_{\varepsilon\downarrow 0}\lim_{n\to\infty}\frac{1}{n}\log\pi_n(\varepsilon), \tag{17}$$

because if this holds, then the claim of the lemma is an immediate consequence of the fact that $s^{-2} = \mathscr{T}_0$. Equation (17) can be proved in three steps.

- We first observe that, due to 'Schilder',

$$\lim_{n\to\infty}\frac{1}{n}\log\pi_n = \lim_{n\to\infty}\frac{1}{n}\left(-\inf_{\boldsymbol{x}\in\mathscr{A}_n}\frac{1}{2}\boldsymbol{x}T_n^{-1}\boldsymbol{x}\right), \tag{18}$$

  with $\mathscr{A}_n := \{\boldsymbol{x} \mid \sum_{i=1}^n x_i \geq n\}$. It is known [22, Section 6.1] that the optimizing $\boldsymbol{x}$, say $\boldsymbol{x}^\star$, is such that

$$\sum_{j=1}^i x_j^\star \equiv \sum_{j=1}^i x_j^\star(n) = \frac{\mathbb{C}\mathrm{ov}\left(S_i, S_n\right)}{v(n)}\cdot n = \frac{v(n)+v(i)-v(n-i)}{2v(n)}\cdot n.$$

  It now follows from (16) that

$$\lim_{n\to\infty}\sum_{j=1}^i x_j^\star(n) = \lim_{n\to\infty}\frac{ns^2+is^2-(n-i)s^2}{2ns^2}\cdot n = i.$$

- Due to the very same line of reasoning, we also have that

$$\lim_{n\to\infty}\frac{1}{n}\log\pi_n(\varepsilon) = \lim_{n\to\infty}\frac{1}{n}\left(-\inf_{\boldsymbol{x}\in\mathscr{B}_n}\frac{1}{2}\boldsymbol{x}T_n^{-1}\boldsymbol{x}\right), \tag{19}$$

  with, for $\varepsilon > 0$,

$$\mathscr{B}_n(\varepsilon) := \left\{\boldsymbol{x} \;\middle|\; \forall i \in \{1,\dots,n\} : \sum_{j=1}^i x_j \in (i(1-\varepsilon), i(1+\varepsilon)), \sum_{j=1}^n x_j \geq n\right\}.$$

- Obviously, we have that $\mathscr{B}_n(\varepsilon) \subseteq \mathscr{A}_n$ for all $\varepsilon > 0$. By construction $\boldsymbol{x}^\star$ lies in $\mathscr{A}_n$, but, due to the fact that $\lim_{n\to\infty}\sum_{j=1}^i x_j^\star(n) = i$, we also have that $\boldsymbol{x}^\star$ lies in $\mathscr{B}_n(\varepsilon)$ (as $n\to\infty$). As a consequence, Expressions (18) and (19) coincide.

Now let $\varepsilon \downarrow 0$, and conclude that $s^{-2} = \mathscr{T}_0$, as claimed. □

REFERENCES

[1] M. BASSEVILLE (1988). Detecting changes in signals and systems–a survey. *Automatica*, **24**, pp. 309–326.

[2] M. BASSEVILLE, I.V. NIKIFOROV (1993). *Detection of abrupt changes: theory and application.* Englewood Cliffs, NJ: Prentice Hall

[3] J. ANTOCH, M. HUŠKOVÁ, and Z. PRÁŠKOVÁ (1997), Effect of dependence on statistics for determination of change. *J. Statist. Plann. Inf.* **60**, pp. 291–310.

[4] T. BARNETT, D. PIERCE, and R. SCHNUR (2011). Detection of anthropogenic climate change in the world's oceans. *Science*, **292**, pp. 270-274.

[5] G. BOX and G. JENKINS (1970). *Time Series Analysis Forecasting and Control.* Holden Day, San Francisco.

[6] P. BROCKWELL and R. DAVIS (2002). *Introduction to Time Series and Forecasting,* 2nd edition. Springer, New York.

[7] B. BRODSKY and B. DARKHOVSKY (1993). *Nonparametric Methods in Change-point Problems.* Kluwer, Dordrecht.

[8] J. BUCKLEW (1985). *Large Deviation Techniques in Decision, Simulation, and Estimation.* Wiley, New York.

[9] C. CALLEGARI, A. COLUCCIA, A. DALCONZO, W. ELLENS, S. GIORDANO, M. MANDJES, M. PAGANO, T. PEPE, F. RICCIATO and P. ŻURANIEWSKI (2013). A methodological overview on anomaly detection. In: *Data Traffic Monitoring and Analysis*, pp. 148-183.

[10] J. CHEN and A. GUPTA (1997). Testing and locating variance change points with application to stock prices. *J. Am. Statist. Assoc.* **92**, pp. 739–747.

[11] M. CZÖRGŐ and L. HORVÁTH (1997). *Limit Theorems in Changepoint Analysis*. Wiley, Chichester.

[12] A. DEMBO and O. ZEITOUNI (1998). *Large Deviations Techniques and Applications,* 2nd edition. Springer, New York.

[13] J. DESHAYES and D. PICARD (1986). Off-line statistical analysis of change-point models using non parametric and likelihood methods. In: *Detection of abrupt changes in signals and dynamical systems. Lecture Notes in Control and Information Sciences*, Vol. 77, pp. 103–168.

[14] J. R. ENGLISH, S.-C. LEE, T. W. MARTIN and C. TILMON (2000). Detecting changes in autoregressive processes with $\bar{X}$ and EWMA charts. *IIE Transactions*, **32**, pp. 1103–1113.

[15] A. GANESH, N. O'CONNELL, and D. WISCHIK (2004). Big Queues. *Lecture Notes in Mathematics*, Volume 1838. Springer, Berlin.

[16] D. GUSTAFSON, A. WILLSKY, J. WANG, M. LANCASTER, and J. TRIEBWASSER (1978). ECG/VCG rhythm diagnosis using statistical signal analysis — I. Identification of persistent rhythms. *IEEE Trans. Biomed. Eng.* **25**, pp. 344-353.

[17] D. GUSTAFSON, A. WILLSKY, J. WANG, M. LANCASTER, and J. TRIEBWASSER (1978). ECG/VCG rhythm diagnosis using statistical signal analysis —II. Identification of transient rhythms. *IEEE Trans. Biomed. Eng.* **25**, pp. 353-361.

[18] R. JOHNSON and M. BAGSHAW (1974). The effect of serial correlation on the performance of CUSUM tests. *Technometrics* **16**, pp. 103–112.

[19] T. L. LAI (1998). Information bounds and quick detection of parameter changes in stochastic systems. *IEEE Transactions on Information Theory*, **44**, pp. 2917–2929.

[20] G. LORDEN (1971). Procedures for reacting to a change in distribution. *Ann. Math. Statist.* **42**, pp. 1897–1908.

[21] J. KUHN, W. ELLENS and M. MANDJES (2014). Detecting changes in the scale of dependent Gaussian processes: a large deviations approach. In: B. Sericola, M. Telek and G. Horváth (Eds.), *Analytical and Stochastic Modeling Techniques and Applications*, Volume 8499 of *Lecture Notes in Computer Science*, pp. 170–184, Springer International Publishing.

[22] M. MANDJES (2007). *Large Deviations for Gaussian Queues*. Wiley, Chichester.

[23] M. MANDJES and P. ŻURANIEWSKI (2011). M/G/∞ transience, and its applications to overload detection. *Perf. Eval.* **68**, pp. 507-527.

[24] I. MCLEOD (1975). Derivation of the theoretical autocovariance function of autoregressive-moving average time series. *Appl. Statist.* **24**, pp. 255–256.

[25] Y. MEI (2008). Is average run length to false alarm always an informative criterion? *Sequential Analysis*, **27**, pp. 354–376.

[26] E. PAGE (1954). Continuous inspection scheme. *Biometrika* **41**, pp. 100–115.

[27] M. POLLAK (1985). Optimal detection of a change in distribution. *Ann. Statist.* **13**, 206–227.

[28] H. POOR and O. HADJILIADIS (2009). *Quickest Detection*. Cambridge University Press, Cambridge, UK.

[29] M. ROBBINS, C. GALLAGHER, R. LUND, and A. AUE (2011). Mean shift testing in correlated data. *J. Time Ser. Anal.* **32**, pp. 498–511.

[30] P. SHAMAN (1976). Approximations for stationary covariance matrices and their inverses with application to ARMA models. *Ann. Statist.*, **4**, pp. 292–301.

[31] A. SHIRYAEV (1963). On optimum methods in quickest detection problems. *Theory Probab. Appl.* **8**, pp. 22–46.

[32] A. SHIRYAEV (1964). On Markov sufficient statistics in non-additive Bayes problems of sequential analysis. *Theory Probab. Appl.* **9**, pp. 604–618.

[33] D. SIEGMUND (1985). *Sequential Analysis*. Springer, New York.

[34] A. SPEROTTO, M. MANDJES, R. SADRE, P.T. DE BOER, and A. PRAS (2012). Autonomic parameter tuning of anomaly-based IDSs: an SSH case study. *IEEE Trans. Netw. Serv. Man.* **9**.

[35] Z. G. STOUMBOS, M. R. REYNOLDS, T. P. RYAN and W. H. WOODALL (2000). The state of statistical process control as we proceed into the 21st century. *J. Am. Stat. Assoc.*, **95**,pp. 992–998.

[36] A. TARTAKOVSKY, I.V. NIKIFOROV and M. BASSEVILLE (2014). *Sequential Analysis: Hypothesis Testing and Changepoint Detection*. Monographs on Statistics & Applied Probability 136, Chapman & Hall/CRC, Boca Raton, FL.

[37] A.G. TARTAKOVSKY, B.L. ROZOVSKII, R.B. BLAŽEK and H. KIM (2006). Detection of intrusions in information systems by sequential change-point methods. *Statistical Methodology*, **3**, pp. 252–293.

[38] M. THOTTAN and J. CHUANI (2003). Anomaly detection in IP networks. *IEEE Transactions on Signal Processing* **8**, pp. 2191–2204.

[39] M. WILSON (2006). A historical view of network traffic models. Unpublished survey paper. See `http://www.cse.wustl.edu/~jain/cse567-06/traffic_models2.htm`

[40] Y. ZHAO, F. TSUNG and Z. WANG (2005). Dual CUSUM control schemes for detecting a range of mean shifts. *IIE Transactions*, **37**, pp. 1047–1057.