Changepoint detection for dependent Gaussian sequences

Citation for published version (APA):
Ellens, W., Kuhn, J., Mandjes, M. R. H., & Zuraniewski, P. (2013). Changepoint detection for dependent Gaussian sequences. (arXiv.org; Vol. 1307.0938 [math.PR]). s.n.

Document status and date:
Published: 01/01/2013

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:
• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher’s website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.
ABSTRACT. In this paper techniques are devised for detecting changepoints in Gaussian sequences, with the distinguishing feature that we do not impose the assumption that the series’ terms be independent. For the specific case that the Gaussian sequence has a creation structure of ARMA type, we develop CUSUM-like procedures; we do so by relying on a large-deviations based approach. In the networking context, these tests can for instance be used to detect a change in traffic parameters, such as the mean, variance or correlation structure. The procedures are extensively validated by means of a broad set of simulation experiments, and generally perform well.

Korteweg-de Vries Institute for Mathematics, University of Amsterdam, the Netherlands.
⋆ TNO Performance of Networks and Systems, Delft, the Netherlands.
M. Mandjes is also with EURANDOM (Eindhoven University of Technology, the Netherlands) and CWI (Amsterdam, the Netherlands).
Email: wendy.ellens@tno.nl, kuhnjulia@outlook.com, m.r.h.mandjes@uva.nl, piotr.zuraniewski@tno.nl

1. INTRODUCTION

The ability to detect changepoints in data sequences (a change in the underlying probability distribution) is of great practical importance. In numerous application domains one is faced with problems of this nature [24]. To mention but a few examples, changepoint techniques have been used in finance [1,9], electrocardiogram analysis [14,15], and climate change [4]. Our own motivation stems from problems in communication networks, most notably detecting changes in the network load [19], and intrusion detection systems [29]. In all these applications the goal is to detect a potential changepoint as soon as possible, while at the same time limiting the number of false alarms.

A commonly used technique in changepoint detection is that of Cumulative Sum (CUSUM) [22]. With \( X_i \) \( (i = 1, \ldots, n) \) denoting the observations, and \( h(\cdot) \) a specific function, the test statistic is based on the sequence of cumulative sums \( \sum_{i=1}^{k} h(X_i) \): if the position of the resulting random walk, relative to the minimum achieved so far, exceeds some predefined threshold, then an alarm is issued. For the specific situation of the \( X_i \)’s being independent, 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 type I error (i.e., false alarms) performance [28]. Apart from the CLT regime, explicit results for the type I error have been derived under a large deviations scaling as well, see e.g. [12] and [8, Ch. VI.E].

Date: July 22, 2013.

Key words and phrases. Changepoint detection, CUSUM, multivariate Normal distribution, ARMA processes, large deviations theory, likelihood ratio, sequential analysis.
The analysis complicates significantly, however, if the \(X_i\) s do not correspond to independent observations. 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 \cite{30} 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 timescales.

An important class of stochastic processes that include dependence is that of the so-called autoregressive moving-average (short: ARMA) processes \cite{5, 6}; assuming that the error terms are i.i.d. samples from a Normal distribution (which is usually done), we thus obtain a versatile class of stationary Gaussian processes that incorporate a broad variety of possible correlation structures. For these ARMA processes Johnson and Bagshaw \cite{16} established the convergence to Brownian motion, thus enabling the type I error analysis of a CUSUM-type procedure. Alternative tests under the CLT scaling were described extensively by Cz"org"o and Horv"ath \cite{10} Ch. IV, with a focus on a Brownian-bridge based test statistic (see also \cite{3}); it is noted that the case of weakly dependent data (think of, e.g., ARMA processes) is to be distinguished from the case of strongly dependent data (think of e.g. fractional Brownian motion). Robbins et al. \cite{25} developed a more explicit procedure to detect a change in the mean of an ARMA process.

While previous work on CUSUM for dependent data has primarily focused on the CLT regime, the main contribution of the present paper is that we consider a large-deviations (short: LD) setting. More specifically, we construct LD-based CUSUM-type changepoint detection tests for dependent, multivariate Normal data, covering the class of (Gaussian) ARMA processes. As LD theory \cite{8, 11} focuses on the rare-event setting, this framework is particularly suitable in situations that the probability of missing an alarm should be kept low — which is generally the case.

An additional attractive feature of applying LD here, is that it nicely facilitates the analysis of multiple hypotheses tests. This is due to the fact that in the LD regime the probability of 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 \cite{13}. The changepoint detection problem under consideration corresponds to a multiple hypotheses setting: it tests if there is change in a parameter value at some point in the dataset.

We now describe the contributions and organization of the paper in greater detail. In Section 2 we provide preliminaries on CUSUM, with a focus on 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. A number of particularly relevant cases are treated in 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 \(f^2\)). 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 of relevant scenarios.

2. Cumulative Sum: preliminaries

Consider a sequence 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 such 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 following technique, known as Cumulative Sum (CUSUM), has been proposed [22] to identify changepoints. We roughly follow the setup presented in [28, Ch. II.6]. Consider first the common likelihood test for $H_0$ versus $H_1(k)$. Evidently, the statistic to be considered is

$$S_k := \frac{\prod_{i=k}^{n} g(X_i)}{\prod_{i=k}^{n} f(X_i)};$$

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 $H_1(k)$, 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

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

with $T_k$ denoting the cumulative sum $\sum_{i=1}^{k} \log g(X_i)/f(X_i)$; 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., [28]. 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 [26, 27], whereas [17, 23] 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 \cite{8,11,18}. 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 \cite[Ch. V.IE]{8}. Under $H_0$, due to reversibility arguments,

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

so that the probability of our interest can be rewritten as

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

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

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

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 \cite{13}).

Relying on Cramér’s theorem \cite[Ch. II.A]{8}, we can rewrite the above decay rate to

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

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

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

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

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

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

In the generic setup focused on in this section, we consider the situation that under $H_0$ the data $X = (X_1, \ldots, X_n)$ has an $n$-dimensional multivariate Normal distribution with mean $\mu_n \equiv \mu$ and covariance matrix $\Sigma_n \equiv \Sigma$, denoted by $\mathcal{N}(\mu, \Sigma)$, while under $H_1$ they stem from $\mathcal{N}(\nu, T)$. It is immediately seen that, without loss of generality, we can pick $\mu = 0$ (by subtracting $\mu$ from $\nu, X_1, \ldots, X_n$). We let $f_n(\cdot)$ and $g_n(\cdot)$ be the corresponding $n$-dimensional densities, that is,

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

and

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

Observe that $\mu$ and $\nu \in \mathbb{R}^n$, while $\Sigma$ 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. Note that the tests in this section test for a single alternative hypothesis. The results of this section will be used in Section 4 to develop a procedure to find a change somewhere in the sequence.

In hypothesis testing, a key concept is that of the likelihood ratio test [28]. It features the test statistic

$$\mathcal{L}_n(X) = \log \frac{g_n(X)}{f_n(X)},$$

which can be evaluated as

$$\mathcal{L}_n(X) = \frac{1}{2} \log |\Sigma| - \frac{1}{2} \log |T| + \frac{1}{2} x^T \Sigma^{-1} x - \frac{1}{2} (X - \nu)^T T^{-1} (X - \nu).$$

At this point the following remark should be made. Note that in principle it is also possible to write the likelihood as the product of the densities of the i.i.d. innovations. These are obtained by first subtracting the conditional means

$$\varepsilon_i := X_i - \mathbb{E}(X_i | X_{i-1}, \ldots, X_1), \quad i = 1, \ldots, n,$$

to obtain independence, and then performing a normalization to obtain identically distributed random variables. Then we are back in the i.i.d. setting, so this reasoning may suggest that we can apply the techniques described in Section 2 and that there is no need for more sophisticated procedures. While such an approach is very well possible in case we wish to detect additive changes (changes in the mean, with the covariance matrix fixed), it turns out that this approach typically cannot be pursued in case we are aiming at changepoint detection of non-additive changes [2] (changes in the correlation structure). Furthermore, for non-invertible ARMA processes (meaning that the MA polynomial has roots on the unit disc) extraction of the innovations from the observed series is at best problematic.

To determine the critical value $nb$ above which the null hypothesis is rejected, we wish to evaluate the type I error probability $\mathbb{P}_0(\mathcal{L}_n(X) \geq nb)$, where $b > \mathbb{E}_0 \mathcal{L}_n(X)/n$. 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, 11], the following equation holds for the decay rate

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

where $\mathcal{I}(b)$ is the rate function.

where $\mathcal{F}(b)$ denotes the associated Legendre transform
\begin{equation}
\mathcal{F}(b) := \sup_\theta \left( \theta b - \lim_{n \to \infty} \frac{1}{n} \log \mathbb{E}_0 \exp(\theta \mathcal{L}_n(X)) \right),
\end{equation}
given that the limiting log-moment generating function exists. This leads to the approximation
\begin{equation}
P_0(\mathcal{L}_n(X) \geq nb) \approx e^{-n\mathcal{F}(b)}.
\end{equation}
To use this approximation, we first compute the moment generating function $\mathbb{E}_0 \exp(\theta \mathcal{L}_n(X))$ in more explicit terms. It is clear that
\begin{equation}
\mathbb{E}_0 \exp(\theta \mathcal{L}_n(X)) = (2\pi)^{-n/2} |\Sigma|^{-1/2} \int_{-\infty}^\infty \cdots \int_{-\infty}^\infty \exp(\theta \mathcal{L}_n(x)) \exp \left( -\frac{1}{2} x^T \Sigma^{-1} x \right) dx_1 \cdots dx_n.
\end{equation}
Then notice that
\begin{equation}
\mathcal{L}_n(x) - \frac{1}{2} x^T \Sigma^{-1} x = \frac{\theta}{2} \log |\Sigma| - \frac{1}{2} x^T (\theta T^{-1} + (1 - \theta) \Sigma^{-1}) x + \theta \nu^T T^{-1} x - \frac{\theta}{2} \nu^T T^{-1} \nu.
\end{equation}
Now realize that $\theta T^{-1} + (1 - \theta) \Sigma^{-1}$ is positive-definite; let $B^T B$ be the corresponding Cholesky decomposition. As a next step, we perform the substitution $y = B x$, so that
\begin{equation}
dx_1 \cdots dx_n = |B^{-1}| dy_1 \cdots dy_n = \frac{1}{|\theta T^{-1} + (1 - \theta) \Sigma^{-1}|^{1/2}} dy_1 \cdots dy_n.
\end{equation}
Then Expression (4) can be rewritten as
\begin{equation}
\frac{\theta}{2} \log |\Sigma| - \frac{1}{2} y^T y + \theta \nu^T T^{-1} B^{-1} y - \frac{1}{2} \nu^T T^{-1} \nu,
\end{equation}
which equals
\begin{align}
\frac{\theta}{2} \log |\Sigma| - \frac{1}{2} (y - \theta (B^{-1})^T \nu)^T (y - \theta (B^{-1})^T \nu) \\
&- \frac{\theta}{2} \nu^T T^{-1} \nu + \frac{\theta^2}{2} \nu^T T^{-1} (\theta T^{-1} + (1 - \theta) \Sigma^{-1})^{-1} T^{-1} \nu.
\end{align}
Recognizing a multivariate Normal density, we conclude that the moment generating function $\mathbb{E}_0 \exp(\theta \mathcal{L}_n(X))$ equals, with $I_n$ denoting an $n \times n$ identity matrix,
\begin{align}
\mathbb{E}_0 \exp(\theta \mathcal{L}_n(X)) & = \left( |\Sigma| \right)^{\theta/2} \left( |\Sigma|^{-1/2} \right) \left( \theta T^{-1} + (1 - \theta) \Sigma^{-1} \right)^{1/2} \\
& \times \exp \left( \frac{\theta}{2} \nu^T T^{-1} \nu + \frac{\theta^2}{2} \nu^T T^{-1} \left( \theta T^{-1} + (1 - \theta) \Sigma^{-1} \right)^{-1} T^{-1} \nu \right) \\
& = \left( \frac{\Sigma}{T} \right)^{\theta/2} \left( \theta T^{-1} \Sigma + (1 - \theta) I_n \right)^{1/2} \\
& \times \exp \left( -\frac{\theta}{2} \nu^T T^{-1} \nu + \frac{\theta^2}{2} \nu^T T^{-1} \left( \theta T^{-1} + (1 - \theta) \Sigma^{-1} \right)^{-1} T^{-1} \nu \right).
\end{align}
The above analysis gives, in principle, a technique to calculate $\mathcal{F}(b)$, 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 3.2, we work out the moment generating function (5) and find the Legendre transform (5) 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

\[
\mathbb{E}_0 \exp(\theta \mathcal{L}_n(X)) = \exp \left( -\frac{\theta}{2} \nu^T T^{-1} \nu + \frac{\theta^2}{2} \nu^T T^{-1} \nu \right).
\]

As a consequence — defining \( \mathcal{F}_n(b) := n \mathcal{F}(b) \) — we have

\[
\mathcal{F}_n(b) = \sup_{\theta} \left( n \theta b + \frac{\theta}{2} \nu^T T^{-1} \nu - \frac{\theta^2}{2} \nu^T T^{-1} \nu \right).
\]

The supremum can be determined explicitly, leading to

\[
\mathcal{F}_n(b) = \frac{(nb + \frac{1}{2} \nu^T T^{-1} \nu)^2}{2 \nu^T T^{-1} \nu}.
\]

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 \( \Sigma \) and \( T \) correspond to independent random variables. In this setting \( \Sigma \) is the diagonal matrix with the vector \( \sigma^2 \) on the diagonal (to be denoted by \( \text{diag}(\sigma^2) \)), while \( T = \text{diag}(\tau^2) \). It is a matter of elementary calculus to verify that

\[
\mathbb{E}_0 \exp(\theta \mathcal{L}_n(X)) = \prod_{i=1}^n \left( \frac{\sigma_i}{\tau_i} \right)^{\theta} \times \prod_{i=1}^n \left( \theta \sigma_i^2 + (1 - \theta) \tau_i^2 \right)^{-1/2} \times \exp \left( -\frac{\theta}{2} \sum_{i=1}^n \frac{\nu_i^2}{\tau_i^2} + \frac{\theta^2}{2} \sum_{i=1}^n \frac{\nu_i^2 \sigma_i^2}{\tau_i^2} + (1 - \theta) \frac{\tau_i^2}{\sigma_i^2} \right).
\]

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, Eqn. (43)] we reject \( H_0 \) if

\[
\max_{\beta \in [0, 1)} \left( \frac{1}{n} \mathcal{L}_{n, \beta}(X) - b(\beta) \right) := \max_{\beta \in [0, 1)} \left( \frac{1}{n} \log g_{n, \beta}(X) - b(\beta) \right) > 0,
\]

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

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

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

\[
- \mathcal{F}(b(\beta)) = \lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}_0 \left( \mathcal{L}_{n, \beta}(X) - b(\beta) > 0 \right) = -\gamma
\]
for a uniform positive $\gamma$, across all $\beta \in [0, 1)$; this enables us to determine $b(\beta)$.

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 $\nu$ 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

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

As an example we could consider $X$ 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 $\varepsilon_i$'s are i.i.d. samples from a zero-mean Normal distribution with variance $\sigma^2$ (where we assume $|\varrho| < 1$). It is known that

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

It is elementary to verify that

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

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

Compared to the function $b(\beta)$ that was derived for the unit-variance i.i.d. case [8, Ch. VI.E, p. 113], $\nu$ needs to be replaced by $\nu(1-\varrho)/\sigma$, 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

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

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 $\varepsilon_i$ are i.i.d. samples from a zero-mean Normal distribution with variance $\sigma^2$.

The following lemma implies that the limiting value of $t_{n,\beta}/(n(1-\beta))$ does not depend on $\beta$, or, put differently, that $t_{n,\beta}$ grows essentially linear in $n(1-\beta)$; cf. [25, Eqn. (9)].

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

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

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

$$-\gamma = \lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}_0 \left( \frac{1}{n} \log f_{n,\beta}(X) \geq b(\beta) \right) = -\frac{1}{2} \left( b(\beta) + \frac{1}{2} \nu^2 \mathcal{F}(1-\beta) \right)^2,$$

and

$$b(\beta) = \nu \sqrt{2\gamma(1-\beta)} - \frac{1}{2} \nu^2 \mathcal{F}(1-\beta).$$

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

$$b(\beta) = \nu \left( \frac{1}{\sigma(1 + \vartheta)} \right) \sqrt{2\gamma(1-\beta)} - \frac{1}{2} \nu^2 \left( \frac{1}{\sigma(1 + \vartheta)} \right)^2 (1 - \beta).$$
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 $\sigma^2$ while under $H_1$ the variance changes from $\sigma^2$ to $\tau^2$ at some specific moment. We set $\nu = 0$, $\Sigma = \sigma^2 I_n$, and $T$ is an $n \times n$ diagonal matrix with $\sigma^2$ at the first $m = \beta n$ diagonal positions ($\beta \in [0, 1]$), and $\tau^2$ at the other diagonal positions. Note that this corresponds to a change in variance at time $\beta n + 1$. Filling out (7), we get
\[
\Lambda_\beta(\theta) := \frac{1}{n} \log \mathbb{E}_0 \exp(\theta \mathcal{L}_n(X)) \\
= \theta(1 - \beta) \log \frac{\sigma}{\tau} + \frac{1}{2} (1 - \beta) \log \tau^2 - \frac{1}{2} (1 - \beta) \log (\theta \sigma^2 + (1 - \theta) \tau^2)
\]
Now let us compute $\mathcal{I}(b(\beta)) = \sup_{\theta} (\theta b(\beta) - \Lambda_\beta(\theta))$. Writing $A_1 + A_2 \theta = \theta \sigma^2 + (1 - \theta) \tau^2$, the optimizing $\theta$ satisfies
\[
b(\beta) = (1 - \beta) \left( \log \frac{\sigma}{\tau} - \frac{1}{2} \frac{A_2}{A_1 + A_2 \theta} \right),
\]
which can be solved, giving
\[
\theta = -\frac{1}{2}(1 - \beta) \frac{b(\beta) - (1 - \beta) \log(\sigma/\tau)}{b(\beta) - (1 - \beta) \log(\sigma/\tau) - \frac{\tau^2}{\sigma^2 - \tau^2}},
\]
so that,
\[
\gamma = \mathcal{I}(b(\beta)) = (1 - \beta) \left( \frac{1}{2} + \frac{\tau^2}{\sigma^2 - \tau^2} \left( \frac{b(\beta)}{1 - \beta} - \log \frac{\sigma}{\tau} \right) - \frac{1}{2} \log \left( \frac{-2\tau^2}{\sigma^2 - \tau^2} \left( \frac{b(\beta)}{1 - \beta} - \log \frac{\sigma}{\tau} \right) \right) \right).
\]
The threshold $b(\beta)$ can be evaluated numerically.

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 $f$. 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 $\nu$ equal 0 and the last $n(1 - \beta)$ equal $\bar{\nu} = f\bar{\mu} - \bar{\mu}$. We suppose that $X$ 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 $\mathcal{L}_{n, \beta}(X)$ of [9] becomes $\mathcal{L}_{n, \beta}(X) = \log g_{n, \beta}(X)/f_n(X)$, where $X = (X_{n}\beta+1, \ldots, X_n)$. This, using the notation of Section 3, reduces to
\[
\frac{1}{2} \log |\Sigma_{n(1-\beta)}| - \frac{1}{2} \log f^{2n(1-\beta)} |\Sigma_{n(1-\beta)}| + \frac{1}{2} X^T \Sigma_{n(1-\beta)}^{-1} X \\
- \frac{1}{2f^2}(X - \nu_{n(1-\beta)})^T \Sigma_{n(1-\beta)}^{-1} (X - \nu_{n(1-\beta)})
\]
\[
= -n(1 - \beta) \log f + \frac{1}{2} X^T \Sigma_{n(1-\beta)}^{-1} X - \frac{1}{2f^2}(X - \nu_{n(1-\beta)})^T \Sigma_{n(1-\beta)}^{-1} (X - \nu_{n(1-\beta)}).
\]
Using (3), it is not hard to verify that the moment generating function $\mathbb{E}_0 \exp(\theta \mathcal{L}_{n, \beta}(X))$ of our test statistic equals
\[
f^{-(1-\beta)n} \left( \sqrt{\theta/f^2 + (1-\theta)} \right)^{-n(1-\beta)} \times \exp \left( -\frac{\theta s_{n, \beta}^2}{2f^2} \theta^2 + \frac{\theta^2 s_{n, \beta}^2}{2(\theta f^2 + (1-\theta)f^4)} \right),
\]
with
\[ s_{n,\beta} := \sum_{i=n+1}^{n+\beta} \sum_{i=n+1}^{n} (\Sigma^{-1})_{i,j}, \]
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
\[ \gamma = \mathcal{I}(b(\beta)) = \sup_{\theta} \left( \theta b(\beta) + \theta (1 - \beta) \log f + \frac{1 - \beta}{2} \log \left( \frac{\theta}{f^{2}} + (1 - \theta) \right) \right). \]
The optimizing \( \theta \) is
\[ -\left( \frac{1}{2} (1 - \beta) \right) \left( b(\beta) + (1 - \beta) \log f + \frac{1}{1/f^{2} - 1} \right), \]
so that
\[ \gamma = \mathcal{I}(b(\beta)) = (1 - \beta) \left( \frac{1}{2} - \frac{1}{1/f^{2} - 1} \left( b(\beta) + \log f \right) - \frac{1}{2} \log \left( \frac{-2}{1/f^{2} - 1} \left( b(\beta) + \log f \right) \right) \right). \]
The threshold \( b(\beta) \) can again be evaluated numerically.

Note that the last equation of Section 4.2 follows directly from the above equation when \( f \) is replaced by \( \tau/\sigma \).

5. NUMERICAL EVALUATION

In Section 4 we have developed changepoint detection tests for dependent sequences. In this section we apply such tests in a set of numerical experiments. The tests of Section 4 take as an input a sequence of size \( n \). In practice the tests are used sequentially, this means that the observations arrive one by one and at every new observation \( X_{m} \) the changepoint detection test is 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 \( \mathcal{L}_{n,\beta}(X) \) 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 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. 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 \( \sigma \) of the \( \varepsilon_{i} \) 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 $L_{50,\beta}(X) = \log \left[ g_{50,\beta}(X) / f_{50}(X) \right]$ is computed according to (2). Here $\nu_i$ is 0 for $i < 100$ and $\nu_i$ is 3 for $i \geq 100$, the covariance matrix $\Sigma = T$ of an ARMA process is computed using the algorithm developed in [21] and $X$ 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 $\alpha$ is put to 0.01, so that $\gamma$ in these equations can be found from $e^{-50 \gamma} = 0.01$. Third, we calculate $1_{50}L_{50,\beta}(X) - b(\beta)$ for $\beta = i_{50}, i_{1}, \ldots, i_{49}$. If the maximum of this difference (taken over $\beta$) is bigger than zero, we raise an alarm. Otherwise we conclude that there is no change-point in the current window. We repeat this step for all windows. All the steps above are repeated 300 times.

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 occurred. For instance, if the change-point 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.

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.

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
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. This interval covers most practically relevant cases – see for instance [20] for a specific example featuring Voice-over-IP, as well as the survey [30]. 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.
We now provide an intuitive explanation as to why our testing procedure tend 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 $\beta$ 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) 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.

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 $\alpha$ (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 [19, 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 $\beta$. Therefore, we can reduce the false alarm rate substantially by ignoring changepoints that correspond to $\beta$ 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.

---

2To 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.
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 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 defined 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. The figure confirms that the changepoint is detected almost immediately for most coefficients. Also note that we detect the changepoint earlier for coefficients that showed a higher false alarm ratio in the previous section.

A notable exception is the case of an AR(1) process with a large positive coefficient where both the false alarm ratio (recall Fig. 9) 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, higher 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 differs 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.

5.3. Sensitivity analysis. In the above experiments, we chose a specific mean and assessed the test’s performance for this mean. In the current section we analyze how this performance (in terms of false alarms and detection delay) is affected by the specific value of the mean. We will see that the delay decreases when the change in mean is larger, which allows 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 the mean change.

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 mean. 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 \( T \) 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 \( T \) 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.

**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 \( \geq -0.3 \), 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.

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 data sequences that obey a multivariate Normal law. These consist of a log-likelihood test statistic, and the corresponding threshold (where large-deviations theory is used to approximate the type I error 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 (dependent) multivariate 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 (dependent) multivariate Normal data.

We have demonstrated our new changepoint detection test by means of a set of examples in which we wish to find a change in the mean of a typical dependent sequence like an AR(1) or an MA(1) process. These simulations have shown that the test performs well (in terms of false alarm ratio and detection delay) for all practically relevant scenarios, i.e., for AR(1) and MA(1) coefficients.
between $-0.3$ and $0.6$, as long as the change in mean is considerably 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 — reducing the number of false alarms with minor negative influence on the detection delay. Moreover, the test performance is to a large extent resilient to changes that are smaller than the expected change (and that are used in the test setup).

Various next steps could be thought of. A first step would be to modify the test such that it can detect a change in the correlation structure within a data sequence. Adapting this type of tests for the analysis of multi-dimensional processes (i.e., detecting a changepoint in a set of data streams, where there may be dependence within each data stream as well as between the various individual data streams) would be a second step. Also, the extension to regression models or state-space models could be investigated.

**Appendix A. Proof of Lemma**

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

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

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

$$\frac{v(n)}{n} \to \left( \sigma\left(1 + \sum_{j=1}^{q} \vartheta_j \right) \right)^2,$$

this identity can alternatively be deduced relying on the spectral density formula for ARMA processes [25].

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

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

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’ [18, Section 4.2], recalling that $T \equiv T_n$ is the covariance matrix of the $X_i$,

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

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

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

because if this holds, then the claim of the lemma is an immediate consequence of the fact that $s^{-2} = \mathcal{F}_0$. Equation (15) can be proved in three steps.

- We first observe that, due to ‘Schilder’,

$$\lim_{n \to \infty} \frac{1}{n} \log \pi_n = \lim_{n \to \infty} \frac{1}{n} \left( -\inf_{x \in \mathcal{F}_0} \frac{1}{2} xT^{-1}_n x \right),$$

where $\mathcal{F}_0$ is the covariance matrix of the $X_i$. Equation (15) can be proved in three steps.

- We first observe that, due to ‘Schilder’,
with \( \mathcal{A}_n := \{ x \mid \sum_{j=1}^n x_i \geq n \} \). It is known [18 Section 6.1] that the optimizing \( x \), say \( x^* \), is such that

\[
\sum_{j=1}^i x_j^* = \frac{\sum_{j=1}^i x_j^*}{n} = \frac{\sum_{j=1}^i x_j^*}{n} = \frac{v(n) + v(i) - v(n-i)}{2v(n)} \cdot n.
\]

It now follows from [14] that

\[
\lim_{n \to \infty} \sum_{j=1}^i x_j^* = \lim_{n \to \infty} \frac{n s^2 + i s^2 - (n-i) s^2}{2 n s^2} \cdot n = i.
\]

Therefore \( \sum_{j=1}^n x_j^* (n)/n \to \beta \) (as \( n \to \infty \)) for \( \beta \in [0, 1) \).

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

\[
\lim_{n \to \infty} \frac{1}{n} \log \pi_n (\varepsilon) = \lim_{n \to \infty} \frac{1}{n} \left( -\inf_{x \in \mathcal{A}_n} \frac{1}{2} x^T T_n^{-1} x \right),
\]

with, for \( \varepsilon > 0 \),

\[
\mathcal{B}_n (\varepsilon) := \left\{ x \mid \forall i \in \{ 1, \ldots, n \} : \sum_{j=1}^i x_j \in (i(1-\varepsilon), i(1+\varepsilon)), \sum_{j=1}^n x_j \geq n \right\}.
\]

- Obviously, we have that \( \mathcal{B}_n (\varepsilon) \subseteq \mathcal{A}_n \) for all \( \varepsilon > 0 \). By construction \( x^* \) lies in \( \mathcal{A}_n \), but, due to the fact that \( \sum_{j=1}^n x_j^* (n)/n \to \beta \), we also have that \( x^* \) lies in \( \mathcal{B}_n (\varepsilon) \) (as \( n \to \infty \)). As a consequence, Expressions (16) and (17) coincide.

Now let \( \varepsilon \downarrow 0 \), and conclude that \( s^{-2} = \mathcal{B}_0 \), as claimed.

\[\square\]

REFERENCES

[1] R. AGGARWAL, C. INCLÁN, and R. LEAL (1999). Volatility in emerging stock markets. J. Fin. Quant. Anal., 34, pp. 33-55.
[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. DARKEVITSKY (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] 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.
[10] M. CZERGÓ and L. HORBÁTH (1997). Limit Theorems in Change-point Analysis. Wiley, Chichester.
[11] A. DEMBO and O. ZEITOUNI (1998). Large Deviations Techniques and Applications, 2nd edition. Springer, New York.
[12] 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.
[13] A. GANEŠ, N. O’CONNELL, and D. WISCHIK (2004). Big Queues. Lecture Notes in Mathematics, Volume 1838. Springer, Berlin.
[14] 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.
[15] 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.
[16] R. Johnson and M. Bagshaw (1974). The effect of serial correlation on the performance of CUSUM tests. Technometrics, 16, pp. 103–112.

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

[18] M. Mandjes (2007). Large Deviations for Gaussian Queues. Wiley, Chichester.

[19] M. Mandjes and P. Zuraniewski (2011). M/G/∞ transience, and its applications to overload detection. Perf. Eval. 68, pp. 507-527.

[20] F. Mata, P. Zuraniewski, M. Mandjes, and M. Mellia (2012). Anomaly detection in VoIP traffic with trends. Proc. ITC 24, Kraków, Poland.

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

[22] E. Page (1954). Continuous inspection scheme. Biometrika 41, pp. 100–115.

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

[24] H. Poor and O. Hadjiliadis (2009). Quickest Detection. Cambridge University Press, Cambridge, UK.

[25] M. Robbins, C. Gallagher, R. Lund, and A. Aue (2011). Mean shift testing in correlated data. J. Time Ser. Anal. 32, pp. 498–511.

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

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

[28] D. Siegmund (1985). Sequential Analysis. Springer, New York.

[29] 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.

[30] M. Wilson (2006). A historical view of network traffic models. Unpublished survey paper. See http://www.arl.wustl.edu/~mlw2/classpubs/traffic_models/