False Alarm Control for Change Point Detection: Beyond Average Run Length

J. Kuhn [•,⋆], M. Mandjes [•], T. Taimre [⋆]

December 2, 2015

[•] Korteweg-de Vries Institute for Mathematics, University of Amsterdam, the Netherlands.
[⋆] School of Mathematics and Physics, The University of Queensland, Australia.
E-mail: j.kuhn@uva.nl

Abstract

A popular method for detecting changes in the probability distribution of a sequence of observations is CUSUM, which proceeds by sequentially evaluating a log-likelihood ratio test statistic and comparing it to a predefined threshold; a change point is detected as soon as the threshold is exceeded. It is desirable to choose the threshold in such a way that the number of false detections is kept to a specified level while on the other hand ensuring a quick detection if a change has occurred. In this paper we analyse the distribution of the CUSUM stopping time when observations may be correlated, with the aim of devising simple yet effective methods for selecting the threshold. In addition to the standard CUSUM procedure we consider window-limited testing where only the n most recent observations are considered at each time point. Traditionally, the number of false alarms is measured by the average run length – the expected time until the first false alarm. However, this is a reasonable criterion only when the expectation is finite. We thus propose an alternative criterion that ensures a large average run length and is more generally applicable. We prove that CUSUM is asymptotically optimal under this criterion, and investigate methods for selecting the threshold such that it is approximately achieved. Apart from the above, we note that the average run length criterion does not allow one to restrict the variability of false alarms, which we argue can be crucial. Therefore, we make a case for a stronger false alarm criterion, and show how it is related to the average run length. To illustrate the procedures and evaluate their performance, we provide numerical examples featuring a multidimensional state space model.

1 Introduction

False alarm control for change point detection procedures is an important problem in many application domains ([19], Section 1.3). Many detection procedures feature a test statistic S_t in form of a random walk in discrete time, with possibly dependent increments. A prominent example of such a procedure is the method of cumulative sums (CUSUM) due to [14], in which case S_t is the log-likelihood ratio (LLR) of the observations up to time t. The test statistic is computed sequentially as new observations arrive, and compared to a threshold b. A change point is detected as soon as $S_t > b$, which defines a stopping time T. One then seeks to choose a threshold that ensures that the number of false alarms is kept low with respect to some appropriate criterion.
In existing literature on change point detection typically the average run length (ARL) – the expected time until the first false alarm is raised – is considered as a performance criterion [2, 19]. Then a threshold is chosen such that at least in some asymptotic regime the ARL equals a desired (large) constant. This criterion is, however, not always informative: in [13], examples are provided where the ARL is infinite even though the detection delay is finite. Apart from this issue, we stress that a large ARL does not ensure that the false alarm probability is low at every particular time

instance: Even if the ARL is controlled to exceed a certain level, the variance of the stopping time may still be large. This can be problematic: For example, in a reliability context, imagine one is monitoring the status of many network elements. Then if the false alarm rate is highly variable, it is not unlikely that false alarms are raised for a large number of elements at the same time, which could cause the capacity of the technicians to attend to all (including true) alarms to be exceeded. More generally, this argumentation applies to scenarios where multiple independent data streams are to be monitored in parallel.

Consequently, more stringent false alarm criteria are desirable. Perhaps the best available candidate is a criterion based on the conditional probability of raising a false alarm that was proposed in [11], and coined maximum local false alarm probability in [19]. However, this criterion is difficult to evaluate: In their recent book [19] Tartakovsky *et al.* note that even an upper bound is lacking. The difficulty arises from the fact that the distribution of the stopping time is hard to evaluate in closed form, even if the distribution of the test statistic is known.

In view of the above, one wishes for further understanding of the distribution of the stopping time as well as simple but effective methods for selecting the threshold such that the probability of raising a false alarm is kept low in a stronger sense than allowed by the ARL criterion. Another issue is that in practice testing the full history of observations may be computationally expensive, and data points will typically not be stationary over a long period of time. This motivates one to consider window-limited change point detection where data is tested in windows of fixed size; for every new observation arriving the oldest observation is dropped.

In this paper, we first propose a new approach to false alarm control that still has the aim to ensure a large ARL but is more generally applicable. This first part thus contributes to the more traditional methodology of false alarm control. In the second part of the paper, we then suggest a different approach: By limiting the number of false alarms at any given time point (which is generally feasible if the threshold is allowed to be adaptive), one can ensure that the variability of the number of false alarms is low and still achieve a large ARL. We therefore argue that the probability of a false alarm at a given time serves better as a false alarm criterion than the ARL.

We now describe our contributions in more detail. In the first part of the paper we show that a large ARL is achieved by restricting the probability of raising a false alarm before a fixed time point. We prove that CUSUM (with or without windows) is still asymptotically optimal under this modified false alarm criterion, and investigate methods for selecting the threshold such that it is satisfied. To do this exactly, one would need closed form expressions for the distribution of the stopping time. Since such expressions are not known in general, we first show how the distribution of the window-limited stopping time can be described in terms of recursive integral equations. For detection procedures without windows, integral equations have been derived based on renewal theory [14, 16, 18]; here we follow a different approach, using results on the maximum of autoregressive processes [21]. We remark that the obtained recursions are not restricted to CUSUM but hold for a broad class of testing procedures including exponentially weighted moving average schemes (EWMA, see [17]).

However, while we thus in principle know the exact distribution of the window-limited stopping time, in general these expressions cannot be solved for the threshold other than numerically, and the latter is only feasible for the left tail. We therefore provide non-asymptotic bounds for the distribution of the CUSUM stopping time when windows are used as well as for testing without windows. These bounds relate the distribution of the CUSUM stopping time to the crossing probability of a random walk, and are therefore easier to evaluate. For example, we can then apply available approximation methods to find a threshold (function) that ensures the proposed false alarm criterion is satisfied. We compare the use of central limit theorem (CLT), large deviations (LD) as well as extreme value (EV) approximations. The latter two methods allow one to obtain a threshold *function* rather than a constant threshold; this increased flexibility can yield an improved delay performance (see Section 4 as well as the example in [5].

We then define our second criterion, the probability of raising an alarm at a given time point, and compare it to the ARL approach. This more stringent criterion has been applied in [3] for the case of independent observations, but seemingly the benefit of this approach with respect to using the ARL had not yet been noticed. We motivate its application and show how the aforementioned

approximations can be used to select the threshold such that this criterion is achieved.
To illustrate the proposed methodologies, we apply the obtained procedures to test for a change in mean in a state space model. The latter is of interest because it is a special type of hidden Markov model; for this class of models it has been found in [13] that the ARL can be infinite. It is moreover not a straightforward example because the observations are multivariate, and because the size of the change is not fixed, but rather a function of $t-k$, where t is the current time and k denotes the change point.
The paper is organized as follows. In Section 2 we define the change point detection problem and the CUSUM method. In Section 3 we discuss our more general approach to ensure a large ARL. We provide asymptotic optimality results for CUSUM in Section 3.1, and analyse the distribution of the stopping time with applications for threshold selection in Sections 3.2 (when testing windows of fixed size) and 3.3 (when testing the full history of observations). The alternative of limiting the false alarm probability at any given time is discussed in Section 4. In Section 5 we show numerical examples featuring a state space model. We conclude in Section 6.

2 Problem and Procedures

We are concerned with testing a stationary sequence of (possibly multidimensional) observations $(V_t) \in \mathbb{R}^{d_v}$ against a change in the underlying probability distribution. At every point in discrete time a new observation arrives and is to be included in the test sample. That is, at time $t \in \mathbb{N}$ we want to test the null hypothesis H_0 of no change before time t against the alternative H_1 of a change point k with $k \in \{1, \ldots, t\}$. Note that the alternative is thus essentially a union of hypotheses $H_1(k)$ that a change occurred at a specific time k. In practice, in view of computational expense it may often be necessary to test data in windows of fixed size n, rather than keeping the whole history of observations. In this case we restrict k to the set $\{t-n+1, \ldots, t\}$. Note that testing the full history of observations can equivalently be regarded as testing with expanding windows: At time t the size of the window to be tested is t. We take this viewpoint in the remainder of the paper as it allows to treat both cases in a more unified manner. Throughout the paper we assume that the observations are identically distributed under H_0.
A testing procedure is optimal if it minimizes the detection delay (we define two commonly used delay criteria in Section 3.1), subject to a condition on the number of false alarms. It is known that the CUSUM method due to [14] is optimal in various settings; for further details see Section 3.1. This motivates that we mostly focus on CUSUM in the current paper. The method is essentially a sequential application of a LLR test. A sequence of LLRs can be regarded as a random walk with random increments $\ell(V_t)$ given by

$$\ell(V_t) := \log\left[q\left(V_t \,|\, V_1^{t-1}\right)/p\left(V_t \,|\, V_1^{t-1}\right)\right] ,$$

where p and q denote the observation densities under H_0 and H_1, respectively, and $V_1^{t-1} := \{V_1, \ldots, V_{t-1}\}$. Note that the sequence of observations may be correlated. We can thus identify the LLRs corresponding to $H_1(1), \ldots, H_1(n)$ with a Markov process

$$\boldsymbol{Y}_m := \left(S_{1:n}(m), \ldots, S_{n:n}(m)\right)', \tag{1}$$

where

$$S_{k:n}(m) := \sum_{i=k+m-1}^{n+m-1} \ell(V_i) , \tag{2}$$

so that $m \geq 1$ corresponds to the number of the first observation within the window that is to be tested. If no windows are to be considered (i.e., in the case of expanding windows), then $m = 1$ is fixed, and the size of the window n increases with time. We then write $S_{k:n} := S_{k:n}(1)$ to simplify the notation. To consider windows of fixed size n, let m increase with time instead.
The standard CUSUM testing procedure with expanding windows features the stopping time

$$\tau = \inf\left\{n \geq 1 : \max_{k \in \{1,\ldots,n\}} S_{k:n} > b\right\} . \tag{3}$$

If window-limited testing (with windows of fixed size) is desired, we define the stopping time to be

$$\omega := \inf\left\{m \geq 1 : \max_{k\in\{1,\ldots,n\}} S_{k:n}(m) > b\right\}, \tag{4}$$

The threshold is to be selected in such a way that the number of false alarms is kept at a desired level. In practice, this is often done by simulation: The threshold is tweaked until the desired level of false alarms is achieved. This approach is, however, only practical if one is aiming for a *constant* threshold. In this paper we also provide methods for selecting the threshold as a *function* $b_n(k)$ of k (corresponding to $H_1(k)$) and n (the latter is only needed in Section 4). This greater flexibility can yield performance improvements as discussed in Section 4.

In change point detection literature, typically the ARL $\mathbb{E}_0 T$ is considered as a false alarm performance criterion, where T denotes the stopping time of the applied testing procedure, and $\mathbb{E}_0$ indicates that the expectation is evaluated with respect to $\mathbb{P}_0$, the measure under H_0. In this paper, we propose a method that ensures that the ARL exceeds the requested level yet is applicable even if the ARL is infinite and thus irrelevant as a performance criterion. In addition, we argue that it may be more desirable to fix the false alarm probability *per window* (fixed or expanding), thus ensuring that the average number of false alarms is stable over time. The methods for achieving both false alarm criteria differ depending on whether window sizes are fixed or increasing. Thus, in summary, we consider the following four different false alarm criteria:

Control of:	Window size:	
	Fixed	Expanding
ARL (Section 3)	$\mathbb{P}_0(\omega \leq N) \leq \alpha$ (Section 3.2)	$\mathbb{P}_0(\tau \leq N) \leq \alpha$ (Section 3.3)
False alarms per window (Section 4)	$\mathbb{P}_0(\omega = n \mid \tau \geq n) \leq \alpha$ for all n	$\mathbb{P}_0(\tau = n \mid \tau \geq n) \leq \alpha$ for all n

3 A New Approach to Traditional False Alarm Control

Traditionally, change point detection methods are designed such that the average time between false alarms is large. Specifically, the ARL criterion requires that

$$\mathbb{E}_0 T \geq \kappa\,, \tag{5}$$

for some given (large) constant κ. The following lemma shows that (5) can be achieved by choosing a threshold that ensures

$$\mathbb{P}_0(T \leq N) \leq \alpha\,, \tag{6}$$

with appropriate N and α, where $T = \omega$ or $T = \tau$, depending on whether or not window sizes are fixed.

Lemma 1. *If (6) holds for a stopping time T, then $\mathbb{E}_0 T \geq \kappa$ is satisfied with $\kappa = N(1-\alpha)$.*

Proof. By assumption, $\mathbb{P}_0\left(T \leq N\right) \leq \alpha$. Thus, for $1 \leq h \leq N$ we have

$$\mathbb{P}_0\big(T > h\big) \geq 1 - \alpha\,,$$

which implies that

$$\mathbb{E}_0 T \geq \sum_{h=1}^{N} (1-\alpha) = N(1-\alpha)\,.$$

□

Because the ARL can be infinite and is thus not always applicable as a false alarm criterion (see [13] for examples), we propose to replace it by (6) in view of the lemma. If (5) is desired, then we can simply express the required κ as $N(1-\alpha)$, which gives additional control when designing the procedure. For example, if the maximum testing period is known to be bounded, then N could represent the length of the period. Otherwise, one could specify κ and α as desired, and choose N accordingly.

3.1 Asymptotic Optimality of CUSUM

It is known that if the CUSUM procedure with stopping time τ satisfies $\mathbb{E}\tau = \kappa$, then it is optimal with respect to certain delay criteria among all procedures that satisfy $\mathbb{E}_0\tau \geq \kappa$. However, in practice it is usually not possible to achieve $\mathbb{E}_0\tau = \kappa$ because $\mathbb{E}_0\tau$ is not known in closed form. Therefore, asymptotic optimality results are of interest which establish optimality of CUSUM with τ satisfying $\mathbb{E}_0\tau \geq \kappa$ asymptotically for large κ. (For details see, for example, [19], Ch. 8.) Similarly, based on Lemma 1 we can prove asymptotic optimality of $T \in \{\tau, \omega\}$ under (6).
Let more generally T be a stopping time with respect to the natural filtration $\mathcal{F}$ associated with the observations. As before we write $\mathbb{P}_i$ and $\mathbb{E}_i$, $i \in \{0,1\}$ for the probability measure and expectation under H_i. Furthermore, we define $\mathbb{P}_1^k$ and $\mathbb{E}_1^k$ to be the probability measure and expectation under $H_1(k)$.
The following delay criteria have been considered in the literature: the *worst-case expected delay* due to [12]

$$\sup_{k\geq 1} \operatorname{ess\,sup} \mathbb{E}_1^k\big[(T-k+1)^+\big]\,,$$

and the less pessimistic delay criterion

$$\sup_{k\geq 1} \mathbb{E}_1^k[T-k \,|\, T \geq k]$$

due to [15]. We now show that CUSUM is optimal among all procedures satisfying (6) with respect to both delay criteria. To this end, we first prove an asymptotic lower bound on the detection delay, similar to [11], Thm. 1. Then we show that this lower bound is attained for small α in combination with large N. The proofs are deferred to the appendix.

Theorem 1. *Suppose that for some positive constant I we have*

$$\lim_{m\to\infty} \mathbb{P}_1^1\left(\max_{1\leq t\leq m} S_{1:t} \geq I(1+\delta)m\right) = 0 \quad \forall \delta > 0\,. \tag{7}$$

Assume that $\alpha := \alpha_N \leq (\log N)/N$. Then

$$\begin{aligned}
&\inf\left\{\sup_{k\geq 1} \operatorname{ess\,sup} \mathbb{E}_1^k\big[(T-k+1)^+ \,|\, \mathcal{F}_{k-1}\big] : \mathbb{P}_0(T\leq N) \leq \alpha\right\} \\
&\quad \geq \left\{\sup_{k\geq 1} \mathbb{E}_1^k[T-k \,|\, T\geq k] : \mathbb{P}_0(T\leq N)\leq \alpha\right\} \\
&\quad \geq \big(I^{-1}+o(1)\big)\log\big(N(1-\alpha)\big)\,.
\end{aligned}$$

as $N \to \infty$.

Proof. Suppose $\mathbb{P}_0(T \leq N) \leq \alpha$. To simplify notation, define $\gamma_N := \log\big(N(1-\alpha)\big)$. We show that for any $\delta > 0$,

$$\mathbb{P}_1^1\left(T-1 \geq (1-\delta)I^{-1}\gamma_N\right) \to 1 \tag{8}$$

as $N \to \infty$. This then implies that

$$\sup_{k\geq 1} \mathbb{E}_1^k[T-k \,|\, T\geq k] \geq \mathbb{E}_1^1[T-1] \geq \big(I^{-1}+o(1)\big)\gamma_N\,.$$

Since $\{T \geq k\} \in \mathcal{F}_{k-1}$, we have

$$\operatorname{ess\,sup} \mathbb{E}_1^k\left[(T-k+1)^+ \,|\, \mathcal{F}_{k-1}\right] \geq \mathbb{E}_1^k[T-k \,|\, T \geq k]\,,$$

and we obtain the statement of the theorem.
To show (8), we consider the sets

$$\begin{aligned} C_\delta &:= \left\{T < (1-\delta)I^{-1}\gamma_N,\, S_{1:T} \leq (1-\delta^2)\gamma_N\right\}\,, \\ \overline{C_\delta} &:= \left\{T < (1-\delta)I^{-1}\gamma_N,\, S_{1:T} > (1-\delta^2)\gamma_N\right\}\,. \end{aligned}$$

(i) Show that $\mathbb{P}_1^1(C_\delta) \to 0$ for every $0 < \delta < 1$. First, we note that

$$\mathbb{P}_1^1(C_\delta) = \int_{C_\delta} \frac{\mathrm{d}\mathbb{P}_1^1}{\mathrm{d}\mathbb{P}_0}\mathrm{d}\mathbb{P}_0 = \int_{\mathrm{C}_\delta} \mathrm{e}^{\mathrm{S}_{1:\mathrm{T}}}\mathrm{d}\mathbb{P}_0 \leq \mathrm{e}^{(1-\delta^2)\gamma_\mathrm{N}}\,\mathbb{P}_0(\mathrm{C}_\delta)\,.$$

Therefore, for N large enough such that $\alpha \leq (\log N)/N \leq I$, we have:

$$\begin{aligned} \mathbb{P}_1^1(C_\delta) &\leq e^{(1-\delta^2)\gamma_N}\,\mathbb{P}_0(T < (1-\delta)I^{-1}\gamma_N) \\ &\leq \big(N(1-\alpha)\big)^{1-\delta^2}\,\mathbb{P}_0\left(T \leq N\right) \\ &\leq (1-\alpha)^{1-\delta^2} N^{-\delta^2}\log N\,, \end{aligned}$$

which tends to zero as $N \to \infty$.

(ii) To prove that $\mathbb{P}_1^1(\overline{C_\delta}) \to 0$, we note that

$$\mathbb{P}_1^1(\overline{C_\delta}) \leq \mathbb{P}_1^1\left(\max_{t \leq (1-\delta)I^{-1}\gamma_N} S_{1:1+t} \geq I(1+\delta)\,(1-\delta)I^{-1}\gamma_N\right)\,;$$

the upper bound tends to zero by (7).

□

Next, we show that the lower bound is attained by ω, similar to [11], Thm. 4. Since $\omega \geq \tau$ almost surely, this implies that the bound is also attained by τ, and thus, both are asymptotically optimal under the conditions of the theorem.

Theorem 2. *Assume that the threshold $b = b_N$ and the window size $n = n_N$ are chosen such that $\mathbb{P}_0(\omega \leq N) \leq \alpha$, where $\alpha = \alpha_N \to 0$ as $N \to \infty$. Further assume that for some positive constant I and $m \in \mathbb{N}$ we have*

$$\liminf_{N\to\infty} n_N I/b_N > 1\,, \tag{9}$$

$$\lim_{R\to\infty} \sup_{k\in\{1,\ldots,m\}} \operatorname{ess\,sup} \mathbb{P}_1^k\left(\frac{1}{R}\sum_{i=m}^{m+R} X_i < I \,\Big|\, \mathcal{F}_{m-1}\right) = 0\,, \tag{10}$$

Then we have:

$$\sup_{k\in\mathbb{N}} \operatorname{ess\,sup} \mathbb{E}_1^k\left[(\omega-k+1)^+ \,|\, \mathcal{F}_{k-1}\right] \leq \big(I^{-1}+o(1)\big)b_N \tag{11}$$

where $o(1) \to 0$ as $N \to \infty$.

Proof. Let $u \in \mathbb{N}$, $k \in \{1,\ldots,n\}$, and define $d_N := \lfloor b_N/I \rfloor$. By (9), we have that for large N

$$\begin{aligned} &\operatorname{ess\,sup} \mathbb{P}_1^k\left(\omega-k+1 > (u-1)\,d_N \,\Big|\, \mathcal{F}_{k-1}\right) \\ &\quad\leq \operatorname{ess\,sup} \mathbb{P}_1^k\left(\max_{m\in\{0,\ldots,(u-1)\,d_N+k-1\}} \max_{l\in\{1,\ldots,d_N\}} S_{l:d_N}(m) \leq b \,\Big|\, \mathcal{F}_{k-1}\right)\,. \end{aligned}$$

The RHS is upper bounded by

$$\operatorname{ess\,sup} \mathbb{P}_1^k \left(\sum_{i=(j-1)d_N+k}^{jd_N+k-1} X_i < b \quad \forall\, 1 \le j \le u \,\middle|\, \mathcal{F}_{k-1} \right)$$
$$= \operatorname{ess\,sup} \prod_{j=1}^{u} \mathbb{P}_1^k \left(\sum_{i=(j-1)d_N+k}^{jd_N+k-1} X_i < b \,\middle|\, \mathcal{F}_{(j-1)d_N+k-1} \right) .$$

By (10) we have that, for $m \in \mathbb{N}$,

$$\sup_{k \in \{1,\dots,m\}} \operatorname{ess\,sup} \mathbb{P}_1^k \left(\sum_{i=m}^{\lfloor b_N/I \rfloor + m - 1} X_i < b_N \,\middle|\, \mathcal{F}_{m-1} \right) \to 0 ,$$

as $N \to \infty$. Hence, for any $\delta > 0$ we can find N sufficiently large such that

$$\mathbb{P}_1^k \left(\sum_{i=(j-1)d_N+k}^{jd_N+k-1} X_i < b \,\middle|\, \mathcal{F}_{(j-1)d_N+k-1} \right) \le \delta .$$

Thus, we conclude that, for large N,

$$\operatorname{ess\,sup} \mathbb{P}_1^k \left(\omega - k + 1 > (u-1)\, d_N \right) \le \delta^u ,$$

in which case we have

$$\sup_{k \ge 1} \operatorname{ess\,sup} \mathbb{E}_1^k \left[(\omega - k + 1)^+ / d_N \,\middle|\, \mathcal{F}_{k-1} \right] \le \sum_{u=0}^{\infty} \delta^u = \frac{1}{1-\delta} .$$

Since we can do this for all $\delta > 0$, this implies that

$$\sup_{k \ge 1} \operatorname{ess\,sup} \mathbb{E}_1^k \left[(\omega - k + 1)^+ \,\middle|\, \mathcal{F}_{k-1} \right] \le \big(I^{-1} + o(1) \big) b_N ,$$

as $N \to \infty$. □

For example, if observations are i.i.d., then the conditions (7) and (10) are satisfied with I the Kullback-Leibler information number, $I = \mathbb{E}_1^1 \ell(V_1)$ [11].
In summary, we have the following corollary.

Corollary 1. *If $b \sim \log\big(N(1-\alpha)\big)$, $\alpha \le \log(N)/N$, and (7), (9) and (10) are satisfied with $I > 0$, then ω is asymptotically optimal as $N \to \infty$ in the sense that it minimizes the detection delay among all stopping times T satisfying $\mathbb{P}_0(T \le N) \le \alpha$.*

In order to select a threshold that ensures (6), we need to be able to evaluate the distribution of the stopping time. We focus on ω in Section 3.2 and turn to τ in Section 3.3. In both sections we first provide results on the distribution of the stopping time, and then show how the threshold function can be selected based on approximations to $\mathbb{P}_0(T \le N)$.

3.2 Window-Limited Testing

First, we show an exact expression for the distribution of the stopping time ω in terms of iterated integrals. Since these are hard to evaluate in practice, we then propose an EV approximation that can be used to select the threshold in order to ensure (6).

3.2.1 Exact Expression in Terms of Iterated Integrals

We show that the test statistic of a large class of change point detection procedures (including the window-limited CUSUM procedure) can be expressed in form of a first order vector autoregressive process (VAR(1)). We can then obtain the distribution of the corresponding stopping time using results on the distribution of the maximum of autoregressive processes [21].
We are interested in finding an expression for

$$\mathbb{P}_0(\omega \leq m) = \mathbb{P}_0\big(\exists j \in \{1, \dots, n\} : \boldsymbol{M}_{m,j} > b(j)\big)\,, \tag{12}$$

where $\boldsymbol{M}_m$ is the n-vector with $j-$th element

$$\begin{aligned} M_{m,j} &:= \max\left\{S_{j:n}(1), \dots, S_{j:n}(m)\right\} \\ &= \max\left\{\boldsymbol{Y}_{1,j}, \dots, \boldsymbol{Y}_{m,j}\right\}. \end{aligned} \tag{13}$$

Note that the process $\boldsymbol{Y}_m$ follows the recursion

$$\boldsymbol{Y}_m = \Psi(\boldsymbol{Y}_{m-1}) + \vartheta \mathbf{1} \ell(V_{m+n})\,, \tag{14}$$

where $\mathbf{1}$ denotes an n-vector of ones. To obtain the window-limited CUSUM procedure, ϑ is set equal to one, and Ψ is defined as $\Psi(\boldsymbol{y}) = C\boldsymbol{y}$, where $C = (c_{i,j})_{i,j=1,\dots,n}$ with $c_{i,i+1} = 1$ for $i = 1, \dots, n-1$ and $c_{i,j} = 0$ otherwise. Interestingly, also other popular change point detection methods can be expressed in this way: For example, to obtain an exponentially weighted moving average (EWMA, see [17]) procedure based on LLRs, define $\Psi(\boldsymbol{y}) = (1-\vartheta)C\boldsymbol{y}$ for $\vartheta \in (0,1)$. Thus, while in this paper we are focussed on the CUSUM procedure, the result in Prop. 1 holds more generally.
Note that (14) is a VAR(1) process, albeit with a degenerate noise process. A paper that gives exact expressions (in terms of iterated Fredholm integrals) for the distribution of $\boldsymbol{M}_m$ for a VAR(1) process is [21]. We adapt their results to our setting, the proof can be found in the appendix.
Let, for fixed $\boldsymbol{x} \in \mathbb{R}^n$,

$$Q_m(\boldsymbol{x}, \boldsymbol{y}) := \mathbb{P}\left(\boldsymbol{M}_m \leq \boldsymbol{x},\, \boldsymbol{Y}_m \leq \boldsymbol{y}\right) \tag{15}$$

for $m \geq 0$. Denote by $\boldsymbol{x}_j$ the $j-$th entry of the vector $\boldsymbol{x}$. Let $\min\{\boldsymbol{x}, \boldsymbol{y}\}$ be the componentwise minimum of $\boldsymbol{x}$ and $\boldsymbol{y}$. Let F be the distribution function of $\ell(V_i)$.

Proposition 1. *We have $Q_m(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{K}\, Q_{m-1}(\boldsymbol{x}, \boldsymbol{y})$ for $m \geq 0$, where*

$$\mathcal{K}\, h(\boldsymbol{y}) = \int_{\mathbb{R}} F\left(\frac{1}{\vartheta} \min_{i=1,\dots,n} \left\{\big(\min\{\boldsymbol{x}, \boldsymbol{y}\} - \Psi(\boldsymbol{z})\big)_i\right\}\right) \mathrm{d}h(\boldsymbol{z})\,. \tag{16}$$

Proof. We adapt the steps in the proof of [21], Thm. 3.1). Note that $\boldsymbol{M}_m \leq \boldsymbol{x}$ implies that $\boldsymbol{Y}_l \leq \boldsymbol{x}$ for $l \leq m$, so $Q_m(\boldsymbol{x}, \boldsymbol{y}) = Q_m\big(\boldsymbol{x}, \min\{\boldsymbol{x}, \boldsymbol{y}\}\big)$. Then for $m \geq 1$:

$$\begin{aligned} &Q_m(\boldsymbol{x}, \boldsymbol{y}) \\ &\quad = \mathbb{P}_0\big(\boldsymbol{M}_{m-1} \leq \boldsymbol{x},\, \boldsymbol{Y}_m \leq \min\{\boldsymbol{x}, \boldsymbol{y}\}\big) \\ &\quad = \mathbb{P}_0\big(\boldsymbol{M}_{m-1} \leq \boldsymbol{x},\, \vartheta \mathbf{1} \ell(V_{m+n}) \leq \min\{\boldsymbol{x}, \boldsymbol{y}\} - \Psi(\boldsymbol{Y}_{m-1})\big) \\ &\quad = \mathbb{E}_F\left[\mathbb{P}_0\left(\boldsymbol{M}_{m-1} \leq \boldsymbol{x},\, \ell(V_{m+n}) \leq \frac{1}{\vartheta} \min_{i=1,\dots,n} \left\{\big(\min\{\boldsymbol{x}, \boldsymbol{y}\} - \Psi(\boldsymbol{Y}_{m-1})\big)_i\right\} \middle| \ell(V_{m+n})\right)\right], \end{aligned}$$

where the latter equation follows by invoking the law of total probability and Bayes' rule, writing $\mathbb{E}_F$ for the expectation under F. With $\mathcal{K}$ as defined in (16) we can write this as $\mathcal{K}\, Q_{m-1}(\boldsymbol{x}, \boldsymbol{y})$ as claimed. □

Thus, for $\mathbb{P}_0\left(\boldsymbol{M}_m \leq \boldsymbol{x}\right) = Q_m(\boldsymbol{x}, \boldsymbol{\infty})$ we obtain

$$\mathcal{K}\, h(\boldsymbol{\infty}) = \int_{\mathbb{R}} F\left(\frac{1}{\vartheta} \min_{i=1,\dots,n} \left\{\big(\boldsymbol{x} - \Psi(\boldsymbol{z})\big)_i\right\}\right) \mathrm{d}h(\boldsymbol{z})\,. \tag{17}$$

In principle, this allows us to compute (12) as

$$\mathbb{P}_0(\omega \leq m) = 1 - \mathbb{P}_0\left(\boldsymbol{M}_{m,i} \leq b(i),\, i = 1, \ldots, n\right) .$$

To evaluate this in practice, at least for small m one can use approximations based on the eigenvalues of the Fredholm kernel $\mathcal{K}$ (see [21]). However, we are aiming for expressions that can be solved for the threshold function $b(\cdot)$. Therefore, even though $\mathbb{P}_0(\omega \leq m)$ is known exactly, we are interested in approximate expressions for the latter that are easier to evaluate.

3.2.2 Approximation for Threshold Selection

When testing is window-limited, we can apply EV theory to approximate the false alarm probability (6). This provides an easily applicable method to select b, which we outline in this section for the example of Gaussian observations. Define

$$\gamma_{i,j}(h) := \mathrm{Cov}(\boldsymbol{Y}_{m,i}, \boldsymbol{Y}_{m+h,j})$$

(note that $\mathrm{Cov}(\boldsymbol{Y}_{m,i}, \boldsymbol{Y}_{m+h,j})$ is indeed independent of m as we assume that observations are i.i.d.). By application of a theorem in [1], we obtain the following corollary (the proof can be found in the appendix).

Corollary 2. *Assume that observations are i.i.d., and that $\sigma > 1$. Then the process $(\boldsymbol{M}_m)$ defined by (13) satisfies*

$$\mathbb{P}_0\left(\frac{\boldsymbol{M}_{m,i} - (n-i+1)\mu}{\sqrt{n-i+1}\sigma} \leq a_m\, \boldsymbol{x}_i + c_m,\, i = 1, \ldots, n\right)$$
$$\to \prod_{i=1}^{n} \exp\left(-\exp(-\boldsymbol{x}_i)\right), \quad \text{as } m \to \infty ,$$

where

$$a_m = (2\log m)^{-1/2} ,$$
$$c_m = (2\log m)^{1/2} - \frac{1}{2}(2\log m)^{-1/2}(\log\log m + \log 4\pi) .$$

Proof. It has been shown in [1] that as $m \to \infty$ the limiting distribution of the process of componentwise maxima of any standard Gaussian process coincides with that of n independent Gumbel variables, provided that the following conditions hold:

$$|\gamma_{i,j}(0)| < r \text{ for } i, j = 1, \ldots, n,\, i \neq j , \tag{18}$$

$$\sum_{h=1}^{\infty} |\gamma_{i,j}(h)|^r < \infty \text{ for all } i, j = 1, \ldots, n . \tag{19}$$

(The former condition was overlooked in [1] as has been noted in [8].)
We apply this theorem to the n-dimensional process $\widetilde{\boldsymbol{M}}_m$ with $i-$th component

$$\frac{\boldsymbol{M}_{m,i} - (n-i+1)\mu}{\sigma\sqrt{n-i+1}} .$$

Note that

$$\frac{S_{k:n}(m) - (n-k+1)\mu}{\sigma\sqrt{n-k+1}}$$

has a standard normal distribution, so that $\widetilde{\boldsymbol{M}}_m$ is indeed the process of componentwise maxima of a standard Gaussian process.

To verify (18), we note that, for $l, k \in \{1, \dots, n\}$ with $l > k$,

$$\mathrm{Cov}\left(\frac{S_{k:n}}{\sigma\sqrt{n-k+1}}, \frac{S_{l:n}}{\sigma\sqrt{n-l+1}}\right) = \frac{1}{\sigma\sqrt{n-k+1}}\,, \tag{20}$$

which is smaller than 1 by assumption.
Finally, (19) is satisfied because for $k, l \in \{1, \dots, n\}$, $h \in \mathbb{N}$, we have

$$\mathrm{Cov}\left(\sum_{i=k}^{n} \ell(V_i), \sum_{j=l+h}^{n+h} \ell(V_j)\right) = \big(n - \max\{k, l+h\} + 1\big)^+\,,$$

which is zero for h large enough. □

Recall that we wish to choose a threshold function that yields

$$\mathbb{P}_0(\omega \le N) = 1 - \mathbb{P}_0(\boldsymbol{M}_{m,i} \le b(i),\, i = 1, \dots, n) \le \alpha\,.$$

Thus, for fixed N and n, Cor. 2 suggests to choose

$$\begin{aligned} b(\beta) = &\left[- a_N \log\left(-\frac{1}{n}\log(1-\alpha)\right) + c_N \right] \\ &\times \sqrt{n(1-\beta)}\,\sigma + n(1-\beta)\,\mu + \delta\,, \end{aligned} \tag{21}$$

where the change point k is written as $n\beta + 1$, $\beta \in \mathcal{B}_n := \{0/n, 1/n, \dots, (n-1)/n\}$ (this notation will turn out to be useful particularly in Section 3.3.2). The parameter δ is a design parameter to be chosen based on simulation. Because $c_N \to \infty$, adding a constant δ (constant with respect to N) is negligible for large N. Numerical experiments suggest that, for small n the choice $\delta = 0$ seems to work well, however, for larger n, a negative δ should be chosen, possibly a function of the other parameters. The precise determination of δ is left for future research. We suggest a choice for δ in Section 3.3.2, for the case of expanding windows.

3.3 Testing With Expanding Windows

We first derive non-asymptotic bounds on the distribution of τ, which can then be used to apply the CLT, LD and EV approximations to select the threshold. The latter two approaches yield a threshold function rather than a fixed threshold, and the achieved false alarm performance is overall closer to the desired level.

3.3.1 Non-asymptotic Bounds

The complication in evaluating $\mathbb{P}_0(\tau \le N)$ arises from the fact that this involves a double maximum of a random walk:

$$\mathbb{P}_0\left(\tau \le N\right) = \mathbb{P}_0\left(\max_{1\le m\le N} \max_{1\le k\le m} S_{k:m} > b\right)\,.$$

In this section we provide bounds that circumvent this problem. The upper bounds we provide below in (22) and (23) turn out to be very tight, particularly if the size of the change is large (see Fig. 1). We use these in Section 3.3.2. We remark that similar bounds could be obtained for $\mathbb{P}_0(\omega \le N)$, however, the adaptation to this case is straightforward and thus we do not provide further details in this paper.

First, note that we have

$$\begin{aligned}\mathbb{P}_0\left(\tau \le N\right) &= \mathbb{P}_0\left(\max_{1\le m\le N}\max_{1\le k\le m} S_{k:m} > b\right)\\ &= \mathbb{P}_0\left(\max_{1\le m\le N}\max_{m\le n\le N} S_{m:n} > b\right)\\ &= \mathbb{P}_0\left(\exists\, m\in\{1,\dots,N\} : \max_{m\le n\le N} S_{m:n} > b\right)\\ &= \mathbb{P}_0\left(\min_{1\le m\le N}\tau_m \le N\right),\end{aligned}$$

where

$$\tau_m =: \inf\{n\ge m : S_{m:n} > b\}\,.$$

Therefore, the CUSUM stopping time can be written as

$$\tau = \min_{m\ge 1}\tau_m\,.$$

Hence, we have

$$\begin{aligned}\mathbb{P}_0\left(\tau \le N\right) &= \mathbb{P}_0\left(\min_{1\le m\le N}\tau_m \le N\right)\\ &= 1-\mathbb{P}_0\left(\min_{1\le m\le N}\tau_m > N\right)\\ &= 1-\mathbb{P}_0(\tau_N > N)\prod_{i=2}^{N}\mathbb{P}_0\left(\tau_{i-1} > N \,|\, \tau_i > N\right),\end{aligned}$$

which yields the bounds

$$1-\mathbb{P}_0(\tau_N > N) \le \mathbb{P}_0\left(\tau\le N\right) \le 1-\prod_{i=1}^{N}\mathbb{P}_0(\tau_i > N)\,. \tag{22}$$

We furthermore note that the RHS is smaller than

$$\begin{aligned}&1-\left(\min_{h\in\{1,\dots,N\}}\mathbb{P}_0(\tau_h > N)\right)^N\\ &\qquad = 1-\mathbb{P}_0\left(\max_{h\in\{1,\dots,N\}} S_{1:h}\le b\right)^N.\end{aligned} \tag{23}$$

Approximations to (23) are available based on which we can devise simple yet effective procedures, see Section 3.3.2.

As Fig. 1 shows, the upper bounds turn out to be very tight. The lower bounds are closer when the size of the change is smaller. To see why this should be true, consider the following heuristic argument. Since the mean μ of the LLR increments is negative, let us suppose that all increments were negative. In this case $S_{i:n} < b$ would imply that $S_{i-1:n} < b$, and hence $\tau_i > N$ would imply that $\tau_{i-1} > N$. Thus, when μ is small compared to σ^2, then $\mathbb{P}_0(\tau \le N) \approx \mathbb{P}_0(\tau_N \le N) = \mathbb{P}_0(X_N > b)$. One would thus expect that an alarm is typically raised at the end of the current window, as is confirmed in numerical experiments (see Fig. 4).

We now discuss how the bound (23) can be used for threshold selection.

3.3.2 Approximations for Threshold Selection

From the upper bound (23) we obtain that a sufficient condition for $\mathbb{P}_0(\tau\le N)\le\alpha$ is

$$1-\mathbb{P}_0\left(\max_{h\in\{1,\dots,N\}} S_{1:h}\le b\right)^N \le \alpha\,,$$

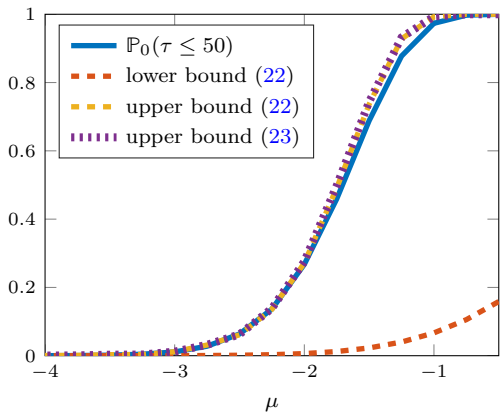

Figure 1: Comparison of $\mathbb{P}_0(\tau \leq N)$ and the bounds provided in (22) and (23), with $N = 50$, $\sigma = 1$ and threshold $b = 0.5$.

or, equivalently,

$$\mathbb{P}_0\left(\max_{h\in\{1,\dots,N\}} S_{h:N} > b\right) \leq 1 - (1-\alpha)^{1/N}\,. \tag{24}$$

Below we discuss different limiting regimes that yield approximations to (24). We focus on the case of independent observations to keep the exposition simple. Note, however, that generalisations to the case of dependent observations of the results we apply in this section are available.

EV Approximation As opposed to the approach in Section 3.2.2, we now consider the univariate process of partial sums. That is, in this case we are interested in the maximum of $S_{k:N}$ over $k \in \{1, \dots, N\}$. Therefore, to achieve (24), the threshold function can be chosen as

$$\begin{aligned} b(\beta) = &\sqrt{N(1-\beta)}\,\sigma\left[-a_N \log\left(-\frac{1}{N}\log(1-\alpha)\right) + c_N\right] \\ &+ N(1-\beta)\,\mu + \delta \end{aligned} \tag{25}$$

where $\beta \in \mathcal{B}_N$. For choosing δ we recall our remark from the previous section that one may expect that – at least for large changes – a change is rather detected at the end of the window, where a single increment is considered. Thus, it seems intuitive to choose δ such that $b\big((N-1)/N\big)$ equals the $1 - \big(1-(1-\alpha)^{1/N}\big)$-quantile of the distribution of the LLR increments. It is confirmed in numerical experiments that this choice indeed yields a good performance of the resulting testing procedure, see the independent data example provided at the end of this section as well as the example in Section 5.

LD Approximation Since we wish the false alarm probability α to be *small*, we may regard this as a rare event scenario; this motivates us to invoke LD theory. Change point detection procedures based on LD approximations have been considered in [3, 5, 9] for i.i.d. and VARMA models, yielding a threshold function $b(\cdot)$ that depends on the assumed position of the change point under the alternative hypothesis. We now explain how to obtain a threshold function from LD approximations. We express the change point k via N, that is, we write $k = N\beta + 1$, where $\beta \in \mathcal{B}_N$. First, note that

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}_0\left(\max_{\beta\in\mathcal{B}_N} \frac{1}{N} S_{N\beta+1:N} > b\right) = \max_{\beta\in\mathcal{B}_N} \lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}_0\left(\frac{1}{N} S_{N\beta+1:N} > b\right)$$

(for details see [5], Section 2). LD theory suggests that for fixed β the false alarm probability can be approximated by

$$\mathbb{P}_0\left(N^{-1}S_{N\beta+1:N} > b(\beta)\right) \approx \exp\Big(-N\mathcal{I}\big(b(\beta)\big)\Big)\,,$$

where $\mathcal{I}$ denotes a function specified below. Recall that we wish the false alarm probability to be kept at a small level α. This suggests to pick the threshold function b such that it satisfies

$$1-(1-\alpha)^{1/N} = \exp\Big(-N\mathcal{I}\big(b(\beta)\big)\Big) \tag{26}$$

for all $\beta \in \mathcal{B}_N$. This choice entails that raising a false alarm is essentially equally likely irrespective of the supposed location of the change point within the window, and it is therefore optimal in terms of type II error performance; see [3], Ch. VI.E.

Now let us make the above more rigorous. The limiting logarithmic moment-generating function $\Lambda(\lambda)$ associated with the distribution of the LLR is defined as

$$\begin{aligned}\Lambda(\lambda) &:= \lim_{N\to\infty}\frac{1}{N(1-\beta)}\log M_{N\beta}(\lambda)\\ &:= \lim_{N\to\infty}\frac{1}{N(1-\beta)}\log \mathbb{E}_0\left(e^{\lambda S_{N\beta+1:N}}\right)\;;\end{aligned} \tag{27}$$

we assume for now that this function exists and is finite for every $\lambda \in \mathbb{R}$. Define $\mathcal{I}$ as the Fenchel–Legendre transform

$$\mathcal{I}\big(b(\beta)\big) := \sup_{\lambda\in\mathbb{R}}\big(\lambda b(\beta) - (1-\beta)\,\Lambda(\lambda)\big)\,.$$

Provided that $\Lambda(\lambda)$ exists for all $\lambda \in \mathbb{R}$, noting that we can rescale as written out in (28), the Gärtner–Ellis theorem [3, 4] yields

$$\lim_{N\to\infty}\frac{1-\beta}{N(1-\beta)}\log\mathbb{P}_0\left(\frac{1}{N(1-\beta)}S_{N\beta+1:N} - \frac{b(\beta)}{1-\beta} > 0\right) = -\mathcal{I}\big(b(\beta)\big)\,. \tag{28}$$

In accordance with the idea expressed in (26), we choose the threshold function $b(\cdot)$ such that it satisfies

$$-\mathcal{I}(b(\beta)) = \lim_{N\to\infty}\frac{1}{N}\log\mathbb{P}_0\left(\frac{1}{N}S_{N\beta+1:N} - b(\beta) > 0\right) = -\gamma \tag{29}$$

for some positive $\gamma = -N^{-1}\log\big(1-(1-\alpha)^{1/N}\big)$, across all $\beta \in \mathcal{B}_N$. Then asymptotically for large N we have that (24) is satisfied.

CLT Approximation As an alternative, we consider the approximation of the false alarm probability based on CLT arguments. Motivated by Donsker's theorem, we can approximate the probability in (24) by [18], Eq. (3.15),

$$\mathbb{P}_0\left(\max_{t\in[0,N]}\sigma B_t + \mu\, t > b\right) = 1-\Phi\left(\frac{b-\mu\,N}{\sigma\sqrt{N}}\right) + e^{\frac{2b\mu}{\sigma^2}}\,\Phi\left(\frac{-b-\mu\,N}{\sigma\sqrt{N}}\right)\,, \tag{30}$$

where B_t is a standard Brownian motion (Wiener process). Then a fixed threshold b (rather than a function as before) can be obtained numerically from setting (30) equal to $1-(1-\alpha)^{1/N}$.

Independent Data Example For illustration we provide an example with independent data, see Figs. 2–3. Note that when testing an independent sequence of $\mathcal{N}(0,\nu)$ observations against a shift in mean of size θ, then the LLR $S_{k:n}(m)$ corresponding to testing against $H_1(k)$ is given by

$$S_{k:n}(m) = \sum_{i=k+m-1}^{n+m-1}\ell(V_i) = \sum_{i=k+m-1}^{n+m-1}\frac{\theta}{\nu^2}\left(V_i - \frac{\theta}{2}\right)\,.$$

Thus, under H_0 the LLR increments are normally distributed with mean $\mu = -(\theta/\nu)^2/2$ and variance $\sigma^2 = (\theta/\nu)^2$.
Application of the EV and CLT approximations is then straightforward. To apply the LD approximation, we need to compute the limiting log-moment-generating function $\Lambda(\lambda)$ in more explicit terms (this way we also check that it indeed exists and is finite for all λ). Because the sequence of observations is independent, with $k = N\beta + 1$, we can write the associated moment-generating function as

$$\begin{aligned} M_{N\beta}(\lambda) &= \mathbb{E}_0 \left[\exp\left(\lambda \sum_{t=k}^{N} \log \frac{q(V_t)}{p(V_t)} \right) \right] \\ &= \prod_{t=k}^{N} \exp\left[\frac{\lambda}{2}(\lambda - 1) \left(\frac{\theta}{\nu} \right)^2 \right] . \end{aligned}$$

With this expression we can compute a threshold function $b(\beta)$ from

$$\gamma = \sup_{\lambda} \left\{ \lambda b(\beta) + (1-\beta)\frac{\lambda}{2}(1-\lambda) \left(\frac{\theta}{\nu} \right)^2 \right\} = \mathcal{I}\big(b(\beta)\big) . \tag{31}$$

The optimizing λ is $1/2 + b(\beta)/\big[(1-\beta)\ (\theta/\nu)^2\big]$, so that from (31) we obtain the desired closed-form expression for $b(\cdot)$:

$$b(\beta) = -\frac{1-\beta}{2} \left(\frac{\theta}{\nu} \right)^2 + \sqrt{2\gamma\,(1-\beta)}\,\frac{\theta}{\nu} . \tag{32}$$

It is interesting to compare this to the EV threshold function (25): We note that in both cases (up to scaling by N because the LD test statistic is divided by N) the threshold function is of the form

$$b(\beta) = N(1-\beta)\mu + \sqrt{N(1-\beta)}\,\sigma\,\zeta(\cdot),$$

where $\zeta(\cdot)$ is some function of the parameters. This form is intuitively appealing: It makes sense to select a threshold that exceeds the expected value of $S_{N\beta+1:N}$ by some function of the standard deviation.
Using the three different thresholds, we can evaluate $\mathbb{P}_0(\tau \leq N)$ by Monte Carlo simulation. Fig. 2 shows that the performance in terms of false alarms is conservative, as was to be expected because we approximate the upper bound (24) rather than $\mathbb{P}_0(\tau \leq N)$ itself. Nevertheless, the false alarm rates are close to the desired level α when the EV approximation is applied. The LD approximation is more conservative, and the CLT approximation does not seem to adjust enough for different α. The latter may be related to the fact that we have to solve for b numerically in this case while in absolute terms $1 - (1-\alpha)^{1/N}$ in (24) does not change much with α. Moreover, it has also been found in Ch. III of [18] that the CLT approximation typically underestimates the probability of interest. An explanation for this is that in (30) it is assumed that the maximum is taken over a continuous (and thus larger) interval.
Fig. 3 displays the obtained delay values for various values of α. We evaluated the delay as the average of the difference of the first detection time and the true change point. Note the trade-off between the false alarm probability and the resulting delay for the LD and CLT approximation. Interestingly, the EV approximation yields a higher delay even though the false alarm probability is higher, suggesting that the shape of the threshold function does not match the shape of the LLRs $S_{N\beta+1:N}$. (Note that this is not generally the case: for the state space model example discussed in Section 5 the EV approximation achieves the better delay performance.)
To further investigate this issue, we plot a graph of the threshold function as well as the LLRs, both as a function of $\beta \in \mathcal{B}_N$, see Fig. 4. Indeed, the distance between the EV threshold and the LLR is not uniform across β. The shape of the LD threshold, however, matches the LLRs very well. The figure also suggests that particularly when using the EV threshold, false alarms usually occur at the end of the window. One may thus wonder whether setting the threshold equal to the $1 - \big(1 - (1-\alpha)^{1/N}\big)$-quantile of the distribution of the LLR increments. This choice, however,

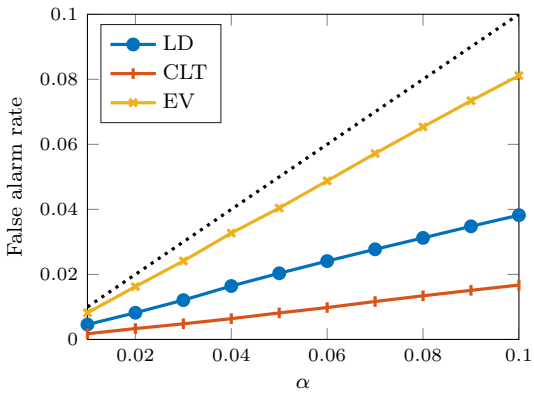

Figure 2: False alarm rates under criterion $\mathbb{P}_0(\tau \leq N) \leq \alpha$, with $\theta = 2$, $N = 150$, $\nu = 1$. Comparison for various α (indicated by the dotted line).

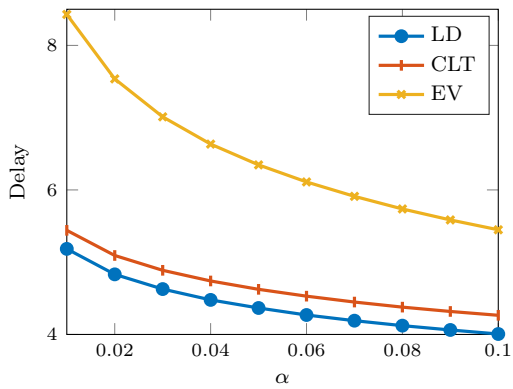

Figure 3: Delay values under criterion $\mathbb{P}_0(\tau \leq N) \leq \alpha$, with $\theta = 2$, $N = 150$, $\nu = 1$.

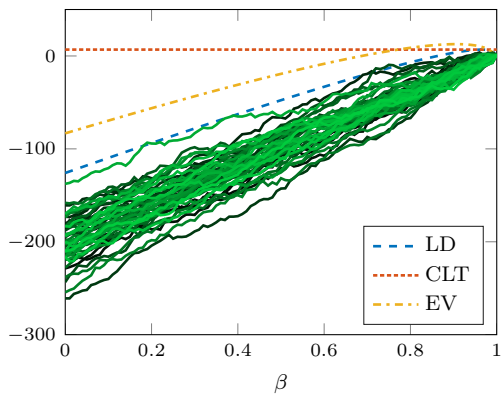

Figure 4: Comparison of LLRs $S_{N\beta+1:N}$ and the threshold functions devised in Section 3.3.2 (the LD threshold is multiplied by N for comparability).

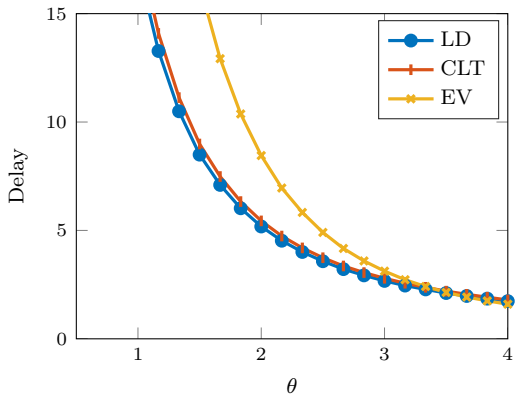

Figure 5: Delay values under criterion $\mathbb{P}_0(\tau \leq N) \leq \alpha$, with $\alpha = 0.01$, $N = 150$, $\nu = 1$. Comparison for various values of the shift size θ.

does not work well, the obtained false alarm rate is usually considerably higher than the desired level (in this example it is close to 1). The figure shows clearly why a threshold function is to be preferred with respect to a constant threshold: the CLT threshold is far away from the actual LLRs, except when β is close to 1. Choosing a function is favourable particularly in view of the detection delay, provided that it closely mimics the behaviour of the LLRs.
Fig. 5 shows a comparison of the delay for various choices of the shift size θ. As expected, the delay performance improves as the shift size increases. We remark that, reassuringly, for different choices of θ the resulting false alarm performance is highly similar to Fig. 2.

4 Better Control over False Alarms

As mentioned in the introduction, the ARL (or the false alarm criterion (6)) may not always be restrictive enough, as is illustrated in Fig. 6. This figure shows the alarm ratio obtained when testing a sequence of independent Gaussian observations with expanding windows. The position of the change point is indicated by the vertical line. The threshold is chosen such that (6) is

achieved. It can be seen that at the beginning of the period, where only a small number of data points are tested, the false alarm ratio is too high but because it then decreases below the desired level, the criterion is still satisfied. This also confirms once more that one should choose b to be a function, rather than a constant threshold as is often assumed. With a constant threshold, as the example shows, $\mathbb{P}_0(\tau \leq N) \leq \alpha$ can only be achieved if $\mathbb{P}_0(\tau = 1) \leq \alpha$. This is true more generally: For the case of independent observations, it has been shown in [20] that the distribution of τ is approximately geometric when the threshold b is large but constant.

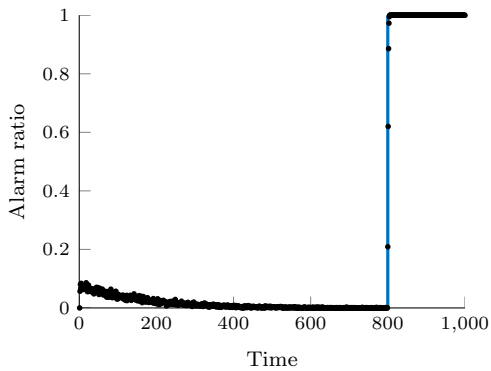

Figure 6: Alarm ratios obtained when testing an i.i.d. Gaussian sequence without windows. A constant threshold is chosen such that $\mathbb{E}_0\tau \geq N(1-\alpha)$, where $N = 1,000$ and $\alpha = 0.01$.

In view of the above, we propose to choose a threshold function that limits the false alarm rate for the current window to be $\widetilde{\alpha}$ (which can be related to α from before as outlined below). That is, we require

$$\mathbb{P}_0(T = n \,|\, T > n-1) = \mathbb{P}_0\left(\max_{k\in\{1,\ldots,n\}} S_{k:n} > b_n\right) = \widetilde{\alpha} \tag{33}$$

to hold, uniformly across all n, where $T \in \{\tau, \omega\}$. If $T = \omega$, the above can be simplified because

$$\mathbb{P}_0(\omega = 1) = \mathbb{P}_0(\omega = n \,|\, \omega > n-1)$$

for any n. Note that

$$\mathbb{P}_0(T \leq N) = \sum_{n=1}^{N} \mathbb{P}_0(T = n)\,, \tag{34}$$

and

$$\begin{aligned}\mathbb{P}_0(T = n) &= \mathbb{P}_0(T = n \,|\, T > n-1)\,\mathbb{P}_0(T > n-1)\\ &= \mathbb{P}_0(T = n \,|\, T > n-1)\left(1 - \sum_{t=1}^{n-1}\mathbb{P}_0(T = t)\right).\end{aligned}$$

Using this recursive equation, it is possible to express each $\mathbb{P}_0(T = n)$ in (34) in terms of conditional probabilities of the form $\mathbb{P}_0(T = n \,|\, T > n-1)$. One obtains that $\mathbb{P}_0(T = n)$ can be written as

$$\mathbb{P}_0(T = n \,|\, T > n-1)\prod_{t=1}^{n-1}\big(1 - \mathbb{P}_0(T = t \,|\, T > t-1)\big)\,.$$

Thus, in principle one can allow for $\widetilde{\alpha}$ to depend on the current window size n as well, and choose a sequence of $\widetilde{\alpha}_n$ such that $\mathbb{P}_0(\tau \leq N) \leq \alpha$ is achieved. For example, we can set

$$\widetilde{\alpha}_1 = \frac{\alpha}{N}, \quad \widetilde{\alpha}_n = \frac{\alpha}{N}\left[\prod_{t=1}^{n-1}(1 - \widetilde{\alpha}_t)\right]^{-1}. \tag{35}$$

Therefore, the condition (33) indeed allows a better control over the false alarm performance as desired.
With respect to the maximum local false alarm probability suggested in [11] (cf. [19], Ch. 8), the criterion (33) has the advantage of simplicity. Indeed, it is not known how the maximum local false alarm probability can be evaluated, even approximately. In contrast, approximations for (33) are readily available. For example, we can apply EV, LD and CLT approximations as in Section 3.3.2 with N replaced by n, and $1-(1-\alpha)^{1/N}$ replaced by $\widetilde{\alpha}$ (we give more details in Section 5.1 for the state space model). In order to ensure (33), we now need the threshold function to depend on the current window size n. Thus, if the window size is fixed, the threshold function is the same for every window. If windows are expanding, we obtain an adaptive threshold function. In the latter case, it is all the more important that evaluation of the threshold function is simple so that this can be carried out on-line as a new observation arrives.

Independent Data Example For illustration we consider again the independent data example from Section 3.2.2, yet now false alarm rates are evaluated according to (33). See Fig. 7 for an example with stopping time ω, which displays the probability $\mathbb{P}_0(\omega = 1)$ that is achieved on average, for various choices of $\widetilde{\alpha}$ (for different shift sizes the false alarm behaviour remains very stable). In comparison to the example in Section 4 we note that the LD and EV approximations are closer but slightly above the desired false alarm rate. This may be explained by the fact that in Section 3.3.2 we approximated an upper bound to $\mathbb{P}_0(\tau \leq N)$ rather than the probability itself. When windows are expanding (and the stopping time is τ and thresholds are adaptive), a very similar false alarm performance is obtained.
Fig. 8 shows that with the sequence $\widetilde{\alpha}_n$ defined by (35) we indeed obtain a false alarm performance similar to Fig. 2, where the threshold was chosen directly to achieve $\mathbb{P}_0(\tau \leq N) \leq \alpha$. The performance in terms of delay is comparable to in Fig. 3, which is not surprising since the probability $\mathbb{P}_0(\tau \leq N)$ is the similar.
In summary, the two figures together confirm that (33) is a stronger false alarm criterion that allows better control over the false alarms at any given time point.

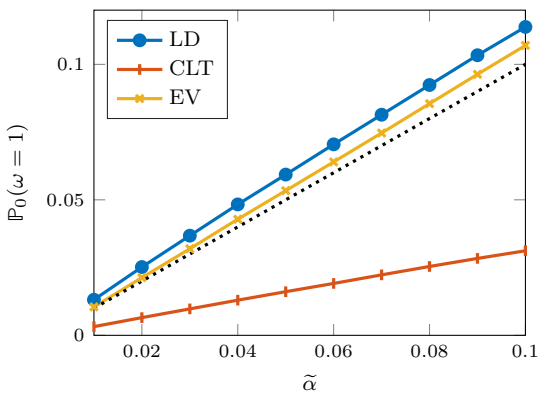

Figure 7: Comparison of probabilities $\mathbb{P}_0(\omega = n|\omega > n-1)$ obtained with adaptive thresholds chosen such that (33) is achieved with $n = 50$, $\theta = 1$, $\nu = 1$, for various $\widetilde{\alpha}$ (indicated by the dotted line).

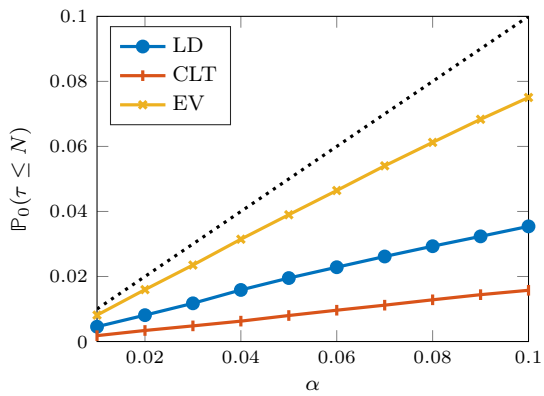

Figure 8: Comparison of probabilities $\mathbb{P}_0(\tau \leq N)$ obtained with adaptive thresholds chosen such that (33) is achieved, where the sequence of $\widetilde{\alpha}_n$ is chosen according to (35) such that $\mathbb{P}_0(\tau \leq N) \leq \alpha$ holds as in (6), with $N = 150$, $\theta = 1$, $\nu = 1$, for various α (indicated by the dotted line).

We provide a more involved example in Section 5, which shows that for Gaussian observations the procedures we proposed can also be applied when observations are not independent.

5 State Space Model Example

We consider an example from [10] that features the following state space model of a sequence of observations (V_t), time t being discrete, with a shift in mean at the *change point* k:

$$X_{t+1} = AX_t + Y_t + \Gamma\, \mathbb{1}_{\{t \geq k\}}\,, \quad V_t = BX_t + Z_t + \Upsilon\, \mathbb{1}_{\{t \geq k\}}\,.$$

The d_x-dimensional process (X_t) represents the unobserved state of the system, with state transition matrix $A \in \mathbb{R}^{d_x \times d_x}$ that has eigenvalues within the unit circle, in which case the system is stable [7]. The vectors Γ and Υ model the shift in mean. We assume that A, B, Q, R, Γ, and Υ, are known, and that the Gaussian white noise processes $Y_t \sim \mathcal{N}(0, Q)$ and $Z_t \sim \mathcal{N}(0, R)$ are independent.
Denote $V_s^t := \{V_s, \ldots, V_t\}$ for s, $t \in \mathbb{N}$. The minimum variance estimator $\widehat{X}_t = \mathbb{E}_0\left[X_t \,|\, V_1^{t-1}\right]$ for the hidden state X_t can be computed efficiently using the well-known Kalman filter [for details see e.g. [6]] as

$$\widehat{X}_t = A\widehat{X}_{t-1} + K_{t-1}(V_{t-1} - B\widehat{X}_{t-1})\,, \quad \widehat{X}_0 = x_0\,.$$

where $K_t := A\Sigma_t B'(B\Sigma_t B' + R)^{-1}$ is the *Kalman gain*, and $\Sigma_t = A\Sigma_{t-1}A' + Q - K_{t-1}(B\Sigma_{t-1}B' + R)K'_{t-1}$ is the *state error covariance matrix*. As a by-product the sequence of *innovations* is obtained,

$$\varepsilon_t := V_t - B\,\widehat{X}_t\,.$$

These represent the new information which is not contained in V_1^{t-1}. They are independent Gaussian zero-mean vectors with covariance

$$\Omega_t := \mathrm{Cov}(\varepsilon_t) = B\Sigma_t B' + R\,.$$

The persistent change in mean in X_t and V_t results in a dynamic change in the innovations; namely, the shift in mean on ε_t is (see [2], Eq. (7.2.110))

$$\rho(t,k) = B\left[\psi(t,k) - A\,\zeta(t-1,k)\right] + \Upsilon,$$

where $\psi(t,k) = A\,\psi(t-1,k) + M$, $\zeta(t,k) = A\,\zeta(t-1,k) + K_t\,\rho(t,k)$, with initial conditions $\psi(k,k) = 0$, $\zeta(k-1,k) = 0$. Thus, the objective is to test whether there is a change point at *some* $k \in \{1, \ldots, n\}$:

$$H_0 : \varepsilon_t \sim \mathcal{N}(0,\, \Omega_{t|t-1}) \quad \text{versus} \quad H_1 : \bigcup_{k=1}^{n} \left[H_1(k) : \varepsilon_t \sim \mathcal{N}\big(\rho(t,k),\, \Omega_{t|t-1}\big)\right]$$

with $t \geq k$. That is, we have to test whether any of the hypotheses $H_1(k)$ holds.
Note that the signature $\rho(t,k)$ of the change on the innovation depends upon both k and t during the transient phase of the Kalman filter. Provided that Σ_t – the estimated covariance matrix of X_t – converges to some matrix Σ as t grows large, it can be seen that the Kalman gain K_t converges to $K = \Sigma B'(B\Sigma B' + R)^{-1}$ ([2], Section 3.2.3.2). For conditions under which this holds see [6], Section 7.3.1.2. The limit Σ (if it exists) can be obtained as the solution of the algebraic Riccati equation

$$\Sigma - A\Sigma A' + A\Sigma B'(B\Sigma B' + R)^{-1}B\Sigma A' - Q = 0\,.$$

In this case as in ([2], Eq. (7.2.112)) we have that asymptotically

$$\begin{aligned}\rho(t,k) \to\; & B\big(I - A(I - KB)\big)^{-1}\Gamma \\ & + (I - B\big(I - A(I - KB)\big)^{-1}AK)\Upsilon =: \boldsymbol{\rho}\,.\end{aligned}$$

Then it also holds that $\Omega_t \to B\Sigma B' + R =: \Omega$.
These limiting expressions are useful for obtaining approximations to the false alarm probability as outlined in Sections 4. Moreover, they yield an approximation to the LLR test statistic that can be computed in a recursive manner – in Section 5.3 we numerically evaluate the test performance when the approximate LLR is used rather than the actual LLR.

5.1 Design of the Testing Procedure

To illustrate the methodology proposed in Section 4, we now define the procedure for testing the state space model against a shift in mean more explicitly. We first evaluate the LLR test statistic $S_{k:n}$, for $k \in \{1, \dots, n\}$, $n \in \{1, \dots, N\}$.
The joint likelihood of V_k^n is given by

$$\begin{aligned} p\big(V_k^n\big) &= \prod_{t=k}^{n} p\big(V_t \,|\, V_1^{t-1}\big) \\ &= \prod_{t=k}^{n} \frac{\exp\left[-\frac{1}{2}\big(V_t - B\widehat{X}_t\big)\Omega_t^{-1}\big(V_t - B\widehat{X}_t\big)'\right]}{\sqrt{(2\pi)^{d_v}\,|\Omega_t|}} \,. \end{aligned}$$

Thus we have that $p(V_k^n) = \prod_{t=k}^{n} p(\varepsilon_t)$, where (abusing notation) $p(\cdot)$ denotes the density function corresponding to its argument. Hence, we can write the LLR as

$$S_{k:n} = \sum_{t=k}^{n} \rho(t,k)'\Omega_t^{-1}\varepsilon_t - \frac{1}{2}\rho(t,k)'\Omega_t^{-1}\rho(t,k)\,. \tag{36}$$

Note that this is not a backward recursion over k because the recursive computation of $\rho(t,k)$ proceeds forward. However, for large $n-k$ we have

$$S_{k:n} \approx \sum_{t=k}^{n} \ell(\varepsilon_t) := \sum_{t=k}^{n} \boldsymbol{\rho}'\Omega^{-1}\varepsilon_t - \frac{1}{2}\boldsymbol{\rho}'\Omega^{-1}\boldsymbol{\rho}\,. \tag{37}$$

We show numerically in Section 5.3 that the test performance remains good if the LLR (36) is replaced by the approximate LLR (37). The mean and variance of the asymptotic likelihood increments (under H_0) are

$$\mu = \mathbb{E}\big[\ell(\varepsilon_t)\big] = -\frac{1}{2}\boldsymbol{\rho}'\Omega^{-1}\boldsymbol{\rho}\,, \quad \sigma^2 = \mathrm{Var}\left(\ell(\varepsilon_t)\right) = \boldsymbol{\rho}'\Omega^{-1}\boldsymbol{\rho}\,. \tag{38}$$

Then the threshold function b can then be chosen as outlined in Section 4.

5.2 Threshold Selection

Using the explicit expressions obtained for μ and σ^2, defining the threshold function according to (25) is straightforward. We obtain the EV threshold

$$b(\beta) = \left[-a_n \log\left(-\frac{1}{n}\log(1-\alpha)\right) + c_n\right]\sqrt{n(1-\beta)\boldsymbol{\rho}'\Omega^{-1}\boldsymbol{\rho}} - n(1-\beta)\frac{1}{2}\boldsymbol{\rho}'\Omega^{-1}\boldsymbol{\rho} + \delta\,, \tag{39}$$

for $\beta \in \mathcal{B}_n$, with a_n and c_n as defined in Cor. 2, and δ chosen to be the $(1-\alpha)$-quantile of the $\mathcal{N}(\mu, \sigma^2)$-distribution.
For the CLT approximation we apply (30), replacing N by n and again using μ and σ as defined in (38).
To obtain the LD based threshold, we can again proceed as in Section 3.3.2. Because the sequence of innovations is independent, with $k = n\beta + 1$, we can write the associated moment-generating function as

$$M_{n\beta}(\lambda) = \mathbb{E}_0\left[\exp\left(\lambda \sum_{t=k}^{n} \log \frac{q\big(V_t \,|\, V_k^{t-1}\big)}{p\big(V_t \,|\, V_k^{t-1}\big)}\right)\right] = \prod_{t=k}^{n} \mathbb{E}_{0,t}\left[\left(\frac{q(\varepsilon_t)}{p(\varepsilon_t)}\right)^{\lambda}\right]\,,$$

where, abusing notation, p and q refer to the distribution of their argument under H_0 and $H_1(k)$ respectively, and $\mathbb{E}_{0,t}$ indicates that the expectation is taken with respect to $p(\varepsilon_t)$. As in [5], Section 3, we can evaluate this as

$$\prod_{t=k}^{n} \exp\left[\frac{\lambda}{2}(\lambda-1)\rho(t,k)'\Omega_{t|t-1}^{-1}\rho(t,k)\right].$$

Combining the above, we may take $\Lambda(\lambda) \approx \frac{\lambda}{2}(\lambda-1)\boldsymbol{\rho}'\Omega^{-1}\boldsymbol{\rho}$ as an approximation for $\Lambda(\cdot)$. This can be used to compute a threshold function $b(\beta)$ as

$$b(\beta) = -\frac{1-\beta}{2}\,\boldsymbol{\rho}'\,\Omega^{-1}\boldsymbol{\rho} + \sqrt{2(1-\beta)\boldsymbol{\rho}'\,\Omega^{-1}\boldsymbol{\rho}\,\gamma}\,, \tag{40}$$

where $\gamma = -n^{-1}\log\widetilde{\alpha}$.

5.3 Numerical Results

We now investigate the performance of the procedures defined in Section 5.1. In order to gain insight regarding the impact of cross-correlation, we fix the diagonal entries of A to be $A_{11} = A_{22} = 0.5$, and vary the off-diagonal entries (both are taken to be equal, $A_{12} = A_{21}$). For various shift sizes, we provide the achieved false alarm and detection rates when using thresholds obtained based on EV, CLT or LD approximations. Further, we fix $B = 0.5\,I_2$, $Q = R = I_2$, and $\widetilde{\alpha}$ and put either Γ or Υ equal to $\mathbf{0}$. The resulting shift sizes are depicted in Fig. 9. The values plotted in Figs. 10–11 were obtained by averaging the relative frequencies of false and true alarms obtained over $10,000$ runs. The significance level $\widetilde{\alpha}$ is indicated by the horizontal dotted black line.

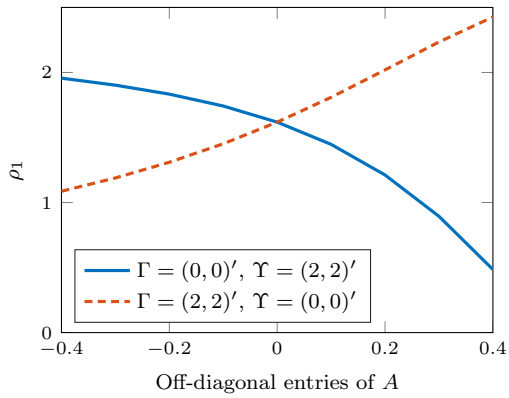

Figure 9: Values for the shift size $\boldsymbol{\rho}$ (here $\rho_1 = \rho_2$).

The LD threshold yields false alarm rates that are consistently close to but slightly above the specified level α, while the CLT threshold is conservative overall. The best false alarm performance is achieved by the EV threshold. The delay values depend on the size of $\boldsymbol{\rho}$: a larger change is easier to detect (compare to Fig. 9). The accuracy of the CLT approximations seems to improve when $\boldsymbol{\rho}$ is small. In this case $\rho(t,k)$ is closer to $\boldsymbol{\rho}$, even when t is small; this may explain why the Brownian approximation works better in this case. Interestingly, the EV approximation works better than the LD approximation in this example: both the false alarm rates as well as the delay values are better.

6 Conclusion

In this paper we have proposed an alternative methodology to achieving the traditional ARL criterion as well as a novel false alarm criterion. The latter allows to restrict the number of false

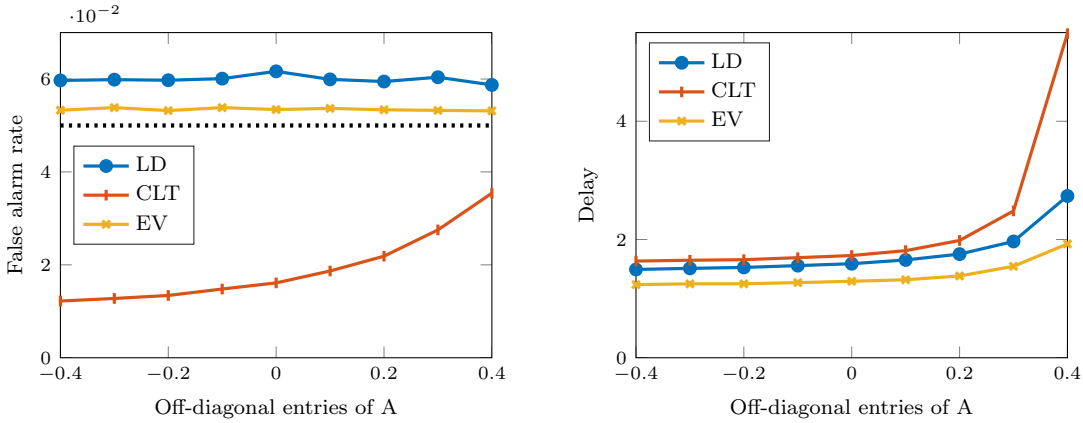

Figure 10: False alarm rate per window and delay values with $\Gamma = (0, 0)'$, $\Upsilon = (2, 2)'$ and $\widetilde{\alpha} = 0.05$ (dotted line).

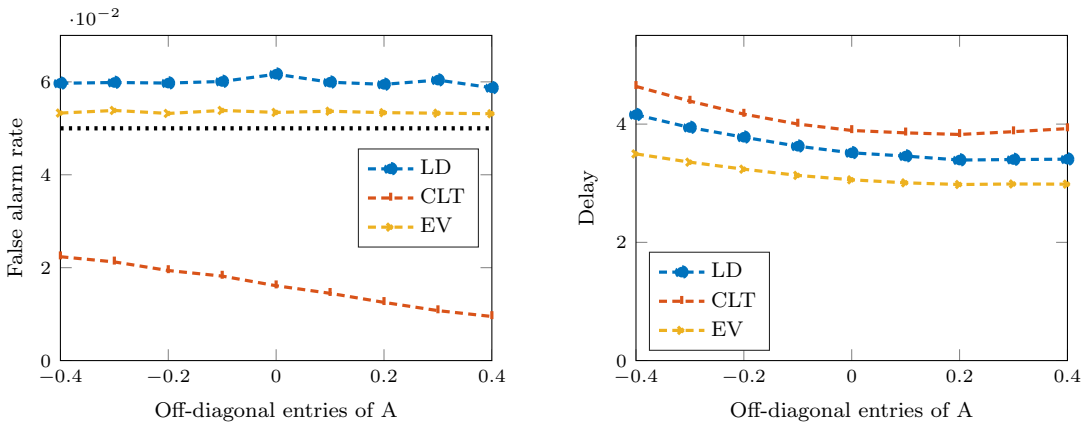

Figure 11: False alarm rate per window and delay values with $\Gamma = (2, 2)'$, $\Upsilon = (0, 0)'$ and $\widetilde{\alpha} = 0.05$ (dotted line).

alarms at every given time point and was shown to be stronger than the ARL. Our main conclusion is that the ARL criterion should be replaced by this alternative false alarm criterion, as the latter allows a better control of the false alarm probabilities.

We moreover provided methods for the selection of the threshold such that the false alarm criteria under consideration hold at least approximately. With respect to numerical methods for threshold selection these are more easily applicable, and moreover allow the selection of a threshold *function* rather than a constant threshold. We investigated the performance of the resulting detection procedures in numerical examples. In terms of false alarm performance, the EV approximation was usually closest to the desired level. However, the LD threshold function typically mimicked the shape of the LLRs more closely, and thus yielded the best trade-off between false alarm and delay performance. We also saw that a threshold function generally is to be preferred in comparison to a constant threshold (and accordingly the EV and the LD threshold functions outperformed the constant CLT threshold).

A topic for future research is the improvement of the EV approximation: We saw that a shift of the resulting threshold function yields a good false alarm performance; however, it should be determined what the optimal size of that shift is, depending on the parameters. Furthermore, the LD approximation requires the evaluation of the limiting logarithmic moment generating function of the LLR. In this paper, we only provided these computations for the case of Gaussian

observations. Similarly, for the EV approximation we assumed Gaussian observations. In future research other distributions should also be considered in more detail.

References

[1] F. Amram. Multivariate extreme value distributions for stationary Gaussian sequences. *Journal of Multivariate Analysis*, 16(2):237–240, 1985.

[2] M. Basseville and I. Nikiforov. *Detection of Abrupt Changes: Theory and Application.* Englewood Cliffs, Prentice Hall, N.J., 1993.

[3] J. Bucklew. *Large Deviation Techniques in Decision, Simulation, and Estimation.* Wiley, New York, 1985.

[4] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications.* Springer-Verlag, New York, 2 edition, 1998.

[5] W. Ellens, J. Kuhn, M. Mandjes, and P. Żuraniewski. Changepoint detection for dependent Gaussian sequences. *Submitted,* `arXiv:1307.0938`, 2013.

[6] G. C. Goodwin and K. S. Sin. *Adaptive Filtering, Prediction and Control.* Information and System Sciences Series. Prentice Hall, Englewood Cliffs, N.J., 1984.

[7] J. D. Hamilton. State-space models. In R. F. Engle and D. L. McFadden, editors, *Handbook of Econometrics*, volume 4, chapter 50, pages 3039 – 3080. Elsevier, 1994.

[8] T. Hsing. Extreme value theory for multivariate stationary sequences. *Journal of Multivariate Analysis*, 29(2):274–291, 1989.

[9] J. Kuhn, W. Ellens, and M. Mandjes. Detecting changes in the scale of dependent Gaussian processes: A large deviations approach. In B. Sericola, M. Telek, and G. Horváth, editors, *Analytical and Stochastic Modeling Techniques and Applications*, volume 8499 of *Lecture Notes in Computer Science*, pages 170–184. Springer International Publishing, 2014.

[10] J. Kuhn, M. Mandjes, and T. Taimre. Mean shift detection for state space models. *To appear in the Proceedings of the 21st International Congress on Modelling and Simulation (MODSIM2015)*, 2015.

[11] T. L. Lai. Information bounds and quick detection of parameter changes in stochastic systems. *IEEE Transactions on Information Theory*, 44(7):2917–2929, 1998.

[12] G. Lorden. Procedures for reacting to a change in distribution. *Annals of Mathematical Statistics*, 42:1897–1908, 1971.

[13] Y. Mei. Is average run length to false alarm always an informative criterion? *Sequential Analysis*, 27:354–376, 2008.

[14] E. Page. Continuous inspection scheme. *Biometrika*, 41:100–115, 1954.

[15] M. Pollak. Optimal detection of a change in distribution. *Annals of Statistics*, 13:206–227, 1985.

[16] M. Pollak and A. G. Tartakovsky. Asymptotic exponentiality of the distribution of first exit times for a class of Markov processes with applications to quickest change detection. *Theory of Probability & Its Applications*, 53(3):430–442, 2009.

[17] S. W. Roberts. Control chart tests based on geometric moving averages. *Technometrics*, 1(3):pp. 239–250, 1959.

[18] D. Siegmund. *Sequential Analysis*. Springer-Verlag New York, 1985.

[19] A. Tartakovsky, I. Nikiforov, and M. Basseville. *Sequential Analysis: Hypothesis Testing and Changepoint Detection*. Monographs on Statistics & Applied Probability 136. Chapman & Hall/CRC, Boca Raton, FL, 2014.

[20] A. G. Tartakovsky. Asymptotic performance of a multichart CUSUM test under false alarm probability constraint. In *44th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC'05)*, pages 320–325. IEEE, 2005.

[21] C. S. Withers and S. Nadarajah. The distribution of the maximum of the multivariate AR(p) and multivariate MA(p) processes. *Statistics and Probability Letters*, 95:48–56, 2014.