Is Average Run Length to False Alarm Always an Informative Criterion?

Yajun Mei
Georgia Institute of Technology, Atlanta, Georgia, USA

Abstract: Apart from Bayesian approaches, the average run length (ARL) to false alarm has always been seen as the natural performance criterion for quantifying the propensity of a detection scheme to make false alarms, and no researchers seem to have questioned this on grounds that it does not always apply. In this article, we show that in the change-point problem with mixture prechange models, detection schemes with finite detection delays can have infinite ARLs to false alarm. We also discuss the implication of our results on the change-point problem with either exchangeable prechange models or hidden Markov models. Alternative minimax formulations with different false alarm criteria are proposed.

Keywords: Average run length; CUSUM; Expected false alarm rate; Quantile run length; Statistical process control; Surveillance.

Subject Classifications: 62L10; 62L15; 60G40.

1. INTRODUCTION

In sequential change-point detection problems, one seeks a detection scheme to raise an alarm as soon as unusual or undesired events happen at (unknown) time \( \nu \) so that appropriate action can be taken. Construction of the detection scheme is based on a sequence of (possibly vector) observations \( X_1, X_2, \ldots \) that are observed sequentially, i.e., one at a time, and it is assumed that the distribution of the \( X \)'s will change if the undesired events occur. The decision whether to raise an alarm at time \( n \) will only depend on the first \( n \) observations. That is, our current decision only depends on our current and past observations, but not on future observations.
Such change-point problems are ubiquitous and, as a consequence, have many important applications, including statistical process control (SPC), industrial quality control, target or signal detection, and epidemiology; see, for example, Basseville and Nikiforov (1993) and Lai (1995, 2001). There has recently been a surge of interest in application to information assurance, network security, health-care and public-health surveillance, particularly a relatively new area called syndromic surveillance. For recent reviews on new applications, see Fienberg and Shmueli (2005), Tartakovsky et al. (2006), Woodall (2006), and the references therein.

A tremendous variety of statistical methods and models have been developed for change-point detection. A partial list includes cumulative sum (CUSUM), Shewhart's control chart, exponentially weighted moving average (EWMA) charts, Shiryayev-Roberts procedures, window-limited control charts, and scan statistics. See, for example, Shewhart (1931), Page (1954), Roberts (1966), Shiryayev (1963, 1978), Lorden (1971), Moustakides (1986), Pollak (1985, 1987), Ritov (1990), Lai (1995) and Kulldorff (2001). New methods and new variants on existing methods are being developed all the time, but the discussion on formulating the right problem and assessing detection schemes under appropriate performance measures is rather limited.

There are two standard mathematical formulations for change-point problems. The first one is a Bayesian formulation, due to Shiryayev (1963), in which the change-point is assumed to have a known prior distribution. The second is a minimax formulation, proposed by Lorden (1971), in which the change-point is assumed to be unknown (possibly $\infty$) but nonrandom. In the literature, both formulations have been extended to dependent observations by simply using new probability measures under which the observations may be dependent. See, for example, Bansal and Papantoni-Kazakos (1986), Basseville and Nikiforov (1993), Brodsky and Darkhovsky (1993, 2000), Yakir (1994), Beibel (1997), Lai (1995, 1998, 2001), Fuh (2003, 2004), Mei (2006b), and Baron and Tartakovsky (2006). However, there seems to be controversies on the optimality properties of widely used Page's CUSUM or Shiryayev–Roberts procedures when observations are not independent, especially under the minimax formulation, as these procedures are (asymptotically) optimal in some situations but can be suboptimal in other situations.

In this article, instead of studying the optimality properties of CUSUM or other detection schemes, we take a further step to look at the appropriateness of the standard minimax formulation when observations are not independent. In the literature, the performance of a detection scheme is typically evaluated by two types of criteria, one being a measure of the detection delay after a change occurs, and the other being a measure of a frequency of false alarms. The importance of the appropriate definition of detection delay has gained a lot of attention in the literature, and several rigorous definitions of detection delay have been proposed under minimax formulations, e.g., the “worst-case” detection delay in Lorden (1971), the “average” detection delay in Shiryayev (1963) and Pollak (1985), and the “exponential penalty” detection delay in Poor (1998). On the other hand, for the false alarm criterion, it is historically standard to use the “average run length” (ARL) to false alarm, which is the expected number of samples to be taken before a false alarm is signaled. Despite some concerns about the ARL to false alarm criterion and some alternative criteria proposed in the literature, see, for example, Barnard (1959), Brodsky and Darkhovsky (1993, 2000), Kenett and Zacks (1998),...
Lai (1998, 2001), and Tartakovsky et al. (2006), most researchers are still using the ARL to false alarm to evaluate the detection schemes, partly because the ARL, a simple function of the distribution of run length to false alarm, seems to be always well-defined. To the best of our knowledge, no researchers have questioned the appropriateness of the ARL to false alarm criterion on grounds that it does not always apply.

The primary goal of this article is to show that a detection scheme with finite detection delay can have infinite ARL to false alarm when the observations are not independent. Moreover, under the standard minimax formulation with the ARL to false alarm criterion, even if well-defined, we are in danger of finding a detection scheme that focuses on detecting larger changes instead of smaller changes when observations are not independent. We illustrate this through a specific example with the “mixture” prechange distribution, and also discuss the implication on the change-point problem either with the “exchangeable” prechange distribution or in the hidden Markov models. Although our example is only a theoretical example to show the invalidity of standard minimax formulation, particularly the ARL to false alarm criterion, we hope researchers pay more attention on the appropriateness of the performance criteria of detection schemes for dependent observations, especially for exchangeable prechange models.

A closely related goal of this article is to propose two new minimax formulations for our example, because the standard minimax formulation is inappropriate. For that purpose, we introduce two new definitions: (1) “asymptotic efficiency,” defined as the divergence rate of the logarithm of the ARL to false alarm; and (2) the “expected false alarm rate” (EFAR), which seems to be closely related to the quantiles of the distribution of the run length to false alarm. We acknowledge that our new formulations may not be applicable directly to the real-world problems. Neither will we attempt to develop a general theory in a general setting. Rather, by presenting new performance criteria for detection schemes in a specific example, we hope it will open new directions to study change-point problems and their applications with dependent observations, particularly with exchangeable prechange models. In view of the history of sequential change-point detection problems for independent observations, which were studied by Shewhart (1931) and Page (1954) but not rigorously formulated and solved until between 1963 and 1990 (Shiryayev, 1963; Lorden, 1971; Moustakides, 1986; Pollak, 1985, 1987; Ritov, 1990), it should be anticipated that significant time and effort will be required for appropriate general theory of change-point problems with exchangeable prechange distributions, or more generally, with any kinds of dependent observations.

The remainder of this article is organized as follows. Section 2 provides a motivation of our example where the prechange distributions is a so-called “mixture” distribution, and presents the standard minimax formulation of the problem. Section 3 defines two families of detection schemes and studies their properties under the standard minimax formulation. Section 4 proposes alternative minimax formulations with different false alarm criteria for our example, and proves the asymptotic optimality property of a class of detection schemes. Section 5 considers a further example with an exchangeable prechange distribution and also discusses the implicit of our results on the change-point problems in Hidden Markov Models (HMMs). Section 6 includes some final remarks, and the Appendices include the proofs of Lemma 3.1 and Theorem 4.1.
2. MOTIVATION AND STANDARD FORMULATION

Before we state our example, let us first consider a classical change-point problem with independent observations. Suppose that $X_1, X_2, \ldots$ are independent normally distributed random variables with variance 1, and assume we want to detect a change from $\theta$ to $\lambda$ in the mean of the $X$'s, where $\theta \leq 0 < \lambda$ and the postchange parameter $\lambda$ is completely specified.

If the prechange parameter $\theta$ is also completely specified, say $\theta = \theta_0$, then a classical detection scheme is Page’s CUSUM procedure, which would declare a change has occurred at the time

$$T_{CM}(\theta_0, b) = \text{first } n \geq 1 \text{ such that } \max_{1 \leq k \leq n} \sum_{i=k}^{n} \log \frac{\phi_\lambda(X_i)}{\phi_\theta(X_i)} \geq b,$$

$$= \text{first } n \geq 1 \text{ such that } \max_{1 \leq k \leq n} \left[ X_i - \frac{\lambda + \theta_0}{2} \right] \geq \frac{b}{\lambda - \theta_0}, \quad (2.1)$$

where the threshold $b > 0$ is prespecified, and $\phi_\mu(x) = (1/\sqrt{2\pi}) \exp(-x^2/2)$ is the probability density function of a normal random variable with mean $\mu$ and variance 1.

Now let us assume that we know the prechange parameter $\theta \leq 0$ but we do not know the exact value of $\theta$. This is one of key features arising from new applications such as syndromic surveillance, where the baseline model when there is no disease outbreak is not completely specified. Even in well-established applications such as quality control, although the assumption of known prechange parameters seems to be reasonable as quality of products can be predefined, the issue of partially specified prechange parameters has recently been recognized, see, for example, Jensen et al. (2006).

Several parametric approaches have been proposed to tackle this problem in the literature. The first is to specify the nominal value $\theta_0$ of the prechange parameter. The choice of $\theta_0$ can be made directly by considering a (prechange) parameter that is close to the postchange parameters because it is always more difficult to detect a smaller change, e.g., $\theta_0 = 0$ or $\lambda/2$ for this example. Alternatively, $\theta_0$ can be estimated from a training sample. However, it is well-known that the performances of such procedures can be rather poor if the true prechange parameter $\theta$ is not $\theta_0$, see, for example, Stoumbos et al. (2000).

The second approach, proposed in Mei (2006a), is to specify a required detection delay at a given $\lambda > 0$ while trying to maximize the ARLs to false alarm for all possible values of prechange parameter $\theta \leq 0$. Mei (2006a) also introduced the idea of “optimizer” to provide a general theory for change-point problems when both the prechange parameter $\theta$ and the postchange parameter $\lambda$ are only partially specified.

The third approach is to eliminate the nuisance, prechange parameter $\theta$, see, for example, Pollak and Siegmund (1991), Yakir et al. (1999), and Krieger et al. (2003). All assume the availability of a training sample and eliminate the nuisance parameter via invariance. Besides invariance, another widely used method to eliminate the nuisance parameter is to integrate the nuisance parameter with respect to weight functions (or priors), see, for example, Wald (1947) and Kiefer and Sacks (1963) for the application in hypothesis testing problems, and Pollak (1987).
and Lai (1998) in the change-point problems when the nuisance parameters are present in the postchange distribution. However, to the best of our knowledge, this method has not been applied to the change-point problems when the nuisance parameters are present in the prechange distribution. This motivates us to consider the approach of eliminating the nuisance prechange parameter via weight-functions, giving us a change-point problem with dependent observations.

Now let us state our example rigorously. Denote by \( P_\theta \) and \( P_{\lambda} \) the probability measures when \( X_1, X_2, \ldots \) are independent and identically distributed (i.i.d.) normal random variables with means \( \theta \) and \( \lambda \) and variance 1, respectively. Assume \( \lambda > 0 \) is completely specified and \( \theta \leq 0 \) has a prior half-cauchy distribution with density

\[
\pi(\theta) = \frac{2}{\pi(1 + \theta^2)} \quad \text{for} \quad \theta \leq 0. \tag{2.2}
\]

Different choices of \( \pi(\theta) \) and their implications will be explained in Section 5. Define a “mixture” probability measure \( P_f = \int_{-\infty}^{0} P_\theta \pi(\theta) d\theta \). That is, under \( P_f \), \( X_1, \ldots, X_n \) have a mixture joint density

\[
f(x_1, \ldots, x_n) = \int_{-\infty}^{0} \left[ \prod_{i=1}^{n} \phi_\theta(x_i) \right] \pi(\theta) d\theta. \tag{2.3}
\]

To emphasize the postchange distribution, we denote by \( P_g \) the probability measure when \( X_1, X_2, \ldots \) are i.i.d. normal with mean \( \lambda \) and variance 1, i.e., \( g = \phi_\lambda \) and \( P_g = P_{\lambda} \).

The problem we are interested in is to detect a change in distribution from the mixture distribution \( f \) in (2.3) to \( g = \phi_\lambda \). Mathematically, for some unknown change-point \( v \) (possibly \( \infty \)), \( X_1, \ldots, X_{v-1} \) are distributed according to the joint mixture density \( f \), whereas \( X_v, X_{v+1}, \ldots \) are independently distributed according to a common density \( g \). Moreover, the postchange observations \( X_v, X_{v+1}, \ldots \) are independent of the prechange observations \( X_1, \ldots, X_{v-1} \). For \( 1 \leq v < \infty \), denote by \( P^{(v)} \) and \( E^{(v)} \) the probability measure and expectation when a change occurs at time \( v \). We shall also use \( P_f \) and \( E_f \) to denote the probability measure and expectation when there is no change, i.e., \( v = \infty \), in which \( X_1, X_2, \ldots \) are distributed with the mixture joint density \( f \).

A detection scheme for detecting that a change has occurred is defined as a stopping time \( T \) with respect to \( \{X_n\}_{n \geq 1} \). The interpretation of \( T \) is that, when \( T = n \), we will raise an alarm at time \( n \) and declare that a change has occurred somewhere in the first \( n \) observations. We want to find a detection scheme that will raise an alarm as soon as possible after a change occurs, but will take observations as many as possible if no change occurs.

For a detection scheme \( T \), the detection delay can be defined by the following “worst case” detection delay defined in Lorden (1971),

\[
\bar{E}_f(T) = \sup_{1 \leq v < \infty} \{\text{ess sup} E^{(v)}[(T - v + 1)^+ | X_1, \ldots, X_{v-1}]\}.
\]

It is worth pointing out that the definition of \( \bar{E}_f(T) \) does not depend on the prechange distribution \( f \) by virtue of the essential supremum, which takes the worst possible \( X \)’s before the change. In our results we can also use the “average” detection delay, proposed by Shiryaev (1963) and Pollak (1985), \( \sup_{1 \leq v < \infty} E^{(v)}(T - v | T \geq v) \), which is asymptotically equivalent to \( \bar{E}_f(T) \).
It is important to mention the relationship between the detection delay \( \mathbb{E}_g(T) \) with \( \mathbb{E}_f(T) \). On the one hand, for many widely used detection schemes, \( \mathbb{E}_g(T) = \mathbb{E}_f(T) \), i.e., the worst case detection delay often occurs when the change-point \( \nu = 1 \), because it is often more difficult to detect when a change occurs at early stages than at latter stages. On the other hand, \( \mathbb{E}_g(T) \) is a more rigorous measurement of detection delay in theory because \( \mathbb{E}_g(T) \) takes into account probability measures \( P^{(i)} \) that are included in the change-point problems.

The desire to have small detection delay \( \mathbb{E}_g(T) \) must, of course, be balanced against the need to have a controlled false alarm rate. When there is no change, \( T \) should be as large as possible, hopefully infinite. However, Lorden (1971) showed that for independent observations, if \( \mathbb{E}_g(T) \) is finite, then \( P_f(T < \infty) = 0 \), i.e., the probability of ever raising a false alarm is 1. This means that we cannot use the probability of ever raising a false alarm as a false alarm criterion. Moreover, Lorden (1971) also showed that, for independent observations, an appropriate measurement of false alarms is \( \mathbb{E}_f(T) \), the ARL to false alarm.

A good detection scheme \( T \) should have large values of the ARL to false alarm \( \mathbb{E}_f(T) \) while keeping the detection delay \( \mathbb{E}_g(T) \) small. To balance the trade-off between these two quantities, the standard minimax formulation of change-point problems for independent observations is then to seek a detection scheme \( T \) that minimizes the detection delay \( \mathbb{E}_g(T) \) subject to \( \mathbb{E}_f(T) \geq \gamma \), where \( \gamma \) is a given constant. In practice, due to the close relationship between \( \mathbb{E}_g(T) \) and \( \mathbb{E}_f(T) \), it often (but not always) suffices to study \( \mathbb{E}_f(T) \) and \( \mathbb{E}_g(T) \).

Much research has been done in the literature to extend the standard minimax formulation to dependent observations by simply replacing the probability densities with the corresponding conditional densities, see, for example, Lai (1998). In particular, in our example where the prechange distribution is the mixture distribution \( f \) in (2.3) and the postchange distribution is \( g = \phi_\lambda \), the standard minimax formulation will evaluate the performance of a detection scheme \( T \) by the detection delay \( \mathbb{E}_g(T) \) and the ARL to false alarm \( \mathbb{E}_f(T) \).

3. INFINITE ARL TO FALSE ALARM

In this section, we illustrate that finite detection delay may be achieved even with infinite ARL to false alarm in the change-point problem with the mixture prechange distribution \( f \) in (2.3). This, of course, is a severe criticism of the standard minimax formulation with the ARL to false alarm as an operating characteristic of a detection scheme, at least in the change-point problem with mixture prechange distributions.

As mentioned in Section 2, the problem of detecting a change in distribution from the mixture distribution \( f \) in (2.3) to \( g = \phi_\lambda \) is motivated from that of detecting a change in the mean of independent normal observations from \( \theta \leq 0 \) to \( \lambda \). Hence detection schemes in the latter problem can be applied to the former problem, although its efficiency or optimality properties could be different. In the following we will consider two families of detection schemes, which correspond to the first two approaches of the latter problem mentioned in Section 2.

Let us first consider Page’s CUSUM procedures \( T_{CM}(\theta_0, b) \) in (2.1) for a given \( \theta_0 \). That is, we will choose a nominal value \( \theta_0 \) for the prechange parameter, and then declare a change from the mixture distribution \( f \) in (2.3) to \( g = \phi_\lambda \) occurs...
if and only if $T_{CM}(\theta_0, b)$ stops. Of course $T_{CM}(\theta_0, b)$ is designed to detect a change in the mean from $\theta_0$ to $\lambda$, and thus it may or may not be efficient to detect a change in distribution from $f$ to $g$. Nevertheless, we can apply it to this problem and study its corresponding properties. For the sake of generality, we assume $0 \leq \theta_0 < \lambda$.

The following lemma, whose proof is in Appendix A, establishes the asymptotic performance of $T_{CM}(\theta_0, b)$.

**Lemma 3.1.** For any $b > 0$ and any $\theta \leq (\theta_0 + \lambda)/2$,

$$
E_{\theta}(T_{CM}(\theta_0, b)) \geq \exp \left( \frac{\lambda + \theta_0 - 2\theta}{\lambda - \theta_0} b \right),
$$

(3.1)

where $E_{\theta}$ denotes the expectation when $X_1, X_2, \ldots$ are i.i.d. normal with mean $\theta$ and variance $1$, and as $b \to \infty$

$$
\bar{E}_x(T_{CM}(\theta_0, b)) = E_x(T_{CM}(\theta_0, b)) = \frac{b}{I(\lambda, \theta_0)} + O(1),
$$

(3.2)

where

$$
I(\lambda, \theta) = E_x \log(\phi_1(X)/\phi_\theta(X)) = (\lambda - \theta)^2/2
$$

(3.3)

is the Kullback–Leibler information number.

The following theorem establishes the performance of $T_{CM}(\theta_0, b)$ in the standard minimax formulation of the problem of detecting a change from the mixture distribution $f$ in (2.3) to $g = \phi_2$.

**Theorem 3.1.** Assume $0 \leq \theta_0 < \lambda$ and $b > 0$. Then the detection scheme $T_{CM}(\theta_0, b)$ has a finite detection delay $\bar{E}_x(T_{CM}(\theta_0, b))$, but has an infinite ARL to false alarm $E_f(T_{CM}(\theta_0, b))$.

**Proof.** The finiteness of the detection delay $\bar{E}_x(T_{CM}(\theta_0, b))$ follows at once from Lemma 3.1 and the fact that the worst-case detection delay of $T_{CM}(\theta_0, b)$ always occurs at $\nu = 1$ regardless of prechange distributions. To derive the ARL to false alarm, note that by the definition of the mixture prechange $f$ in (2.3), for any stopping time $T$,

$$
E_f(T) = \int_{-\infty}^{0} E_{\theta}(T)\pi(\theta)d\theta,
$$

(3.4)

where $\pi(\theta)$ is defined in (2.2). Thus, by Lemma 1,

$$
E_f(T_{CM}(\theta_0, b)) \geq \int_{-\infty}^{0} \exp \left( \frac{\lambda + \theta_0 - 2\theta}{\lambda - \theta_0} b \right) \pi(\theta)d\theta
$$

$$
= \int_{-\infty}^{0} \exp \left( \frac{\lambda + \theta_0 - 2\theta}{\lambda - \theta_0} b \right) \left[ \frac{2}{\pi(1 + \theta^2)} \right] d\theta
$$

which diverges for any $b > 0$. That is, $E_f(T_{CM}(\theta_0, b)) = \infty$ for any $b > 0$. □
To illustrate that \( \{T_{CM}(\theta_0, b)\} \) in (2.1) is not the only family of detection schemes with finite detection delay and infinite ARLs to false alarm, the second family of detection schemes we considered is those proposed in Mei (2006a), which is defined by

\[
T^*(a) = \text{first } n \geq a \text{ such that } \max_{1 \leq k \leq n-a+1} \sum_{i=k}^{n} \left[ X_i - \frac{\lambda}{2} \right] \geq \frac{\lambda}{2} a, \tag{3.5}
\]

for \( a > 0 \). As shown in Mei (2006a), \( \{T^*(a)\} \) are asymptotically optimal solutions in the problem of maximizing the ARL to false alarm, \( \mathbb{E}_g T \), for all possible values \( \theta \leq 0 \) subject to the constraint on the detection delay \( \mathbb{E}_g T \leq \gamma \). The following lemma establishes the asymptotic performance of \( T^*(a) \). Because this lemma is a special case of Theorem 2.3 of Mei (2006a), we state it here without proof.

**Lemma 3.2.** For any \( a > 0 \) and any \( \theta \leq 0 \),

\[
\mathbb{E}_\theta(T^*(a)) \geq \exp(I(\lambda, \theta)a), \tag{3.6}
\]

where \( I(\lambda, \theta) \) is defined in (3.3) and as \( a \to \infty \)

\[
\mathbb{E}_g(T^*(a)) \leq a + (C + o(1))\sqrt{a}, \tag{3.7}
\]

where \( C = (1/\lambda)\sqrt{2/\pi} \).

From Lemma 3.2 and relation (3.4), it is straightforward to show that for any \( a > 0 \), the detection scheme \( T^*(a) \) has a finite detection delay \( \mathbb{E}_g(T^*(a)) \), but has an infinite ARL to false alarm \( \mathbb{E}_g(T^*(a)) \).

It is worth pointing out several implications of our example of detecting a change in distribution from the mixture distribution \( f \) in (2.3) to \( g = \phi_\lambda \). First of all, besides \( T_{CM}(\theta_0, b) \) in (2.1) and \( T^*(a) \) in (3.5), it is easy to construct many other families of detection schemes with similar properties. With suitably chosen boundary values, these detection schemes can have the same detection delays, and have infinite ARL to false alarm. Hence the performance of these detection schemes are indistinguishable under the standard minimax formulation or the ARL to false alarm criterion.

Second, for dependent observations, the fact that a detection scheme \( T \) has infinite ARL to false alarm does not necessarily imply a small probability of ever raising a false alarm. In fact, in our example, any detection scheme \( T \) with finite detection delay \( \mathbb{E}_g(T) \) will raise a false alarm with probability 1. To see this, fix a detection scheme \( T \) with finite detection delay \( \mathbb{E}_g(T) \). Applying the results of Lorden (1971) to the problem of detecting a change in the mean from \( \theta \) to \( \lambda \), we have \( \mathbb{P}_\theta(T < \infty) = 1 \). By the definition of \( f \) in (2.3), \( \mathbb{P}_f(T < \infty) = \int_{-\infty}^{\theta} \mathbb{P}_\theta(T < \infty) \pi(\theta)d\theta = 1 \), implying that the probability of \( T \) ever raising a false alarm is 1. In particular, for the detection scheme \( T_{CM}(\theta_0, b) \) in (2.1) or \( T^*(a) \) in (3.5), although their ARLs to false alarm are infinite, they raise a false alarm with probability 1 even if there is no change from the mixture distribution \( f \) in (2.3).

Third, under the standard minimax formulation with the ARL to false alarm criterion, detecting larger changes will play a more important role in the change-point with mixture prechange distributions, which may be undesirable. To see this,
for a given detection scheme \( T \) in our example, \( E_g(T) \) is generally an exponential decreasing function of \( |\theta| \) as \( \theta \leq 0 \), see, for example, relations (3.1) and (3.6) for the detection schemes \( T_{CM}(\theta_0, b) \) and \( T^*(a) \), respectively. Hence, by (3.4), with suitably chosen \( \pi(\theta) \), the main contribution to \( E_g(T) \) will come from \( E_g(T) \) with negatively large \( \theta \) values (far away from the postchange distribution \( g = \phi_j \)). This means that detecting larger changes will play a more important role under the standard minimax formulation. However, larger changes should be easily detected by any reasonable detection schemes, and one may be more interested in developing sophisticated schemes to detect a smaller change.

Finally, in the standard minimax formulation of change-point problems with dependent observations, an influential result is that of Lai (1998), which showed that

\[
E_g(T) \geq (1 + o(1)) \frac{\log E_f(T)}{I}
\]

for any stopping time \( T \), if the following sufficient condition holds: there exists some constant \( I > 0 \) such that for any \( \delta > 0 \),

\[
\lim_{n \to \infty} \sup \sup_{\nu \geq 1} \sup_{t \leq n} \left\{ \max_{i \leq t} Z_i \geq I(1 + \delta)n | X_1, \ldots, X_{t-1} \right\} = 0,
\]

where \( Z_i = \log(g(X_i | X_1, \ldots, X_{i-1})/f(X_i | X_1, \ldots, X_{i-1})) \) is the conditional log-likelihood ratio. That is, if Lai’s sufficient condition in (3.8) holds, then any detection schemes with finite detection delay \( E_g(T) \) will have finite ARL to false alarm \( E_f(T) \). Combining this with Theorem 3.1 yields that in our example with the mixture distribution \( f \) defined in (2.3), there is no constant \( I > 0 \) satisfying Lai’s sufficient condition in (3.8). Therefore, although Lai’s sufficient condition can be useful in some situations, e.g., it will be used in Appendix B to prove Theorem 4.1 of the present article, it may not be applicable in general.

4. ALTERNATIVE MINIMAX FORMULATIONS

The purpose of this section is to present two alternative minimax formulations for the change-point problem when \( f \) is the mixture distribution defined in (2.3) and \( g = \phi_j \), as the standard minimax formulation with the ARL to false alarm criterion is inappropriate here.

4.1. Relative Divergence Rate

From the mathematical point of view, if some integrals or series sums go to \( \infty \), it is natural to consider their divergence rates. It turns out that this strategy can be adopted to our example: suppose the problem of detecting a change in distribution from \( f \) to \( g \) can be treated as a limit of the problems of detecting changes from \( f_\xi \) to \( g \) with \( f_\xi \) converging to \( f \) as \( \xi \to -\infty \), then for a given detection scheme \( T \), \( E_{f_\xi}(T) \) is the limit of \( E_{f_\xi}(T) \) as \( \xi \to -\infty \). If \( E_{f_\xi}(T) = \infty \), it is possible that \( E_{f_\xi}(T) \) is finite and well-defined for each \( \xi \), but \( E_{f_\xi}(T) \) \( \to \infty \) as \( \xi \to -\infty \). In that case, it is natural to consider the divergence rate of \( E_{f_\xi}(T) \). However, we need to take into account the
trade-off between the detection delays and false alarms, which will likely provide a bound on the divergence rate of $E_{f_x}(T)$. Thus, it will be more informative to treat that bound as a baseline rate and consider the relative divergence rate with respect to the baseline rate.

To present our ideas rigorously, first note that for independent observations, the asymptotic efficiency of a family of detection schemes $\{T(a)\}$ in the problem of detecting a change in distribution from $f$ to $g$ can be defined as

$$e(f, g) = \liminf_{a \to \infty} \frac{\log E_f(T(a))}{I(g, f)E_g(T(a))}$$

(4.1)

where $I(g, f) = E_g(\log(g(X)/f(X)))$ is the Kullback–Leiber information number, and the family $\{T(a)\}$ is required to satisfy $\bar{E}_g(T(a)) \to \infty$ as $a \to \infty$. It is well-known (Lorden, 1971) that for independent observations, $e(f, g) \leq 1$ for all families, and the equality can be achieved by Page’s CUSUM procedures for detecting a change in distribution from $f$ to $g$. This suggests to define a family of detection schemes $\{T(a)\}$ is asymptotically efficient at $(f, g)$ if $e(f, g) = 1$. It follows that Page’s CUSUM procedure for detecting a change in distribution from $f$ to $g$ is asymptotically efficient at $(f, g)$ when observations are independent.

Now in our context when $f$ is the mixture distribution defined in (2.3) and $g = \phi_\mu$, the definition of the asymptotic efficiency $e(f, g)$ in (4.1) cannot be applied directly for the following two reasons: (a) $E_f(T)$ can be $\infty$ as shown in the previous section, and (b) it is not clear how to define the information number $I(g, f)$.

Fortunately, this concept can be salvaged if we replace $f$ by a sequence of distributions $\{f_\xi\}$ with $f_\xi$ converging to $f$. Define “new mixture” distributions

$$f_\xi(x_1, \ldots, x_n) = \int_\xi^0 \left[ \prod_{i=1}^n \phi_\mu(x_i) \right] \pi_\xi(\theta) d\theta,$$

(4.2)

where

$$\pi_\xi(\theta) = \frac{\pi(\theta)}{\int_\xi^0 \pi(u) du} \quad \text{for } \theta \leq 0.$$

Note that as $\xi \to -\infty$, $\pi_\xi(\theta) \to \pi(\theta)$ for all $\theta \leq 0$, and $f_\xi(x_1, \ldots, x_n) \to f(x_1, \ldots, x_n)$ for all $n \geq 1$. That is, the problem of detecting a change in distribution from $f$ to $g$ can be thought of as a limit of the problems of detecting changes from $f_\xi$ to $g$ as $\xi \to -\infty$.

Denote by $P_{f_\xi}$ and $E_{f_\xi}$ the probability measure and expectation, respectively, when $X_1, X_2, \ldots$ are distributed according to the mixture distribution $f_\xi$. The following theorem, whose proof is highly nontrivial and is presented in Appendix B, establishes the information bounds in the problem of detecting a change in distribution from $f_\xi$ in (4.2) to $g = \phi_\mu$.

**Theorem 4.1.** For each pair $(f_\xi, g)$,

$$\bar{E}_{f_\xi}(T) \geq (1 + o(1)) \frac{\log E_{f_\xi}(T)}{I(\lambda, \xi)}$$

(4.3)

for any stopping time $T$ as $E_{f_\xi}(T)$ or $E_g(T)$ goes to $\infty$, where $I(\lambda, \xi) = (\lambda - \xi)^2/2$ is the Kullback–Leibler information number.
We are now in a position to present the new minimax formulation in the problem of detecting a change in distribution from the mixture distribution \( f \) in (2.3) to \( g = \phi_\zeta \). To accomplish this, define the asymptotic efficiency of a family of detection schemes \( \{T(a)\} \) at \( (f, g) \) as

\[
e^*(f, g) = \lim_{\xi \to -\infty} \inf e^*(f_\xi, g),
\]

where

\[
e^*(f_\xi, g) = \lim_{a \to \infty} \inf \frac{\log E\{f_\xi(T(a))\}}{I(\lambda, \zeta) - I(\lambda, \xi)}
\]

is the asymptotic efficiency in the problem of detecting a change in distribution from \( f_\xi \) to \( g \). Then Theorem 4.1 implies that \( e^*(f_\xi, g) \leq 1 \) and \( e^*(f, g) \leq 1 \) for all families of detection schemes, so we can define

**Definition 4.1.** \( \{T(a)\} \) is asymptotically efficient at \( (f, g) \) if \( e^*(f, g) = 1 \).

It is useful to note that our new definition of asymptotic efficiency, \( e^*(f, g) \) in (4.4), can be thought of as the relative divergence rate of \( \log E\{f_\xi(T(a))\} \) with respect to the baseline rate \( I(\lambda, \zeta) \), because we have \( \log E\{f_\xi T(a)\} \to \log E\{f T(a)\} = \infty \) and \( I(\lambda, \zeta) = (\lambda - \zeta)^2/2 \to \infty \) as \( \zeta \to -\infty \). Alternatively, the definition of \( e^*(f, g) \) can also be thought of as taking into account the difficulties of detecting different changes. This viewpoint can be very useful in theory and practice because it may allow one to detect both large and small changes efficiently. Of course the efficiencies will have different meanings for large and small changes.

It is not clear so far whether 1 is the sharp upper bound of \( e^*(f, g) \) over all families of detection schemes. In other words, can we find a family of detection schemes that is asymptotically efficient at \( (f, g) \) under this definition? It turns out that the answer is “yes,” and such asymptotically efficient detection schemes exist. Before offering such detection schemes, let us point out it is nontrivial to find detection schemes that are asymptotically efficient under our new definition.

For instance, the family of detection schemes \( T_{CM}(\theta, b) \) in (2.1) is not asymptotically efficient at \( (f, g) \). In fact, the asymptotic efficiency of \( T_{CM}(\theta, b) \) is \( e^*(f, g) = 0 \). An outline of the proof is as follows. It is easy to show that \( E\{T_{CM}(\theta, b)\} \) is a decreasing function of \( \theta \in (-\infty, \theta_0) \), because, intuitively, it will take longer to make a false alarm for a larger change. Then by relation (4.6) below, we have

\[
\log E\{f_\xi(T_{CM}(\theta_0, b))\} \leq \log E\{f_\theta(T_{CM}(\theta_0, b))\} = (1 + o(1))\left(\frac{\lambda + \theta_0 - 2\zeta}{\lambda - \theta_0}b\right),
\]

where the last equation follows from the classical results on Page’s CUSUM procedures for independent observations. By (3.2) and the definition of \( e^*(f, g) \) in (4.5), we have

\[
e^*(f_\xi, g) \leq \frac{(\lambda - \theta_0)(\lambda + \theta_0 - 2\zeta)}{(\lambda - \zeta)^2}.
\]
Because the right-hand side of the above inequality converges to 0 as $\xi \to -\infty$, the asymptotic efficiency of the family of detection schemes $T_{CM}(\theta_0, b)$ in (2.1) is $e^*(f, g) = 0$ by the definition of $e^*(f, g)$ in (4.4).

Now let us present asymptotic efficient detection under our new definition. It turns out that one family of such detection schemes is the detection scheme $T^*(a)$ defined in (3.5). The following theorem establishes the asymptotic optimality properties of $T^*(a)$ under our new definition in the problem of detecting a change in distribution from the mixture distribution $f$ in (2.3) to $g = \phi_2$.

**Theorem 4.2.** The family of the detection schemes $\{T^*(a)\}$ defined in (3.5) is asymptotically efficient at $(f, g)$.

**Proof.** It suffices to show that for any arbitrary $\xi \leq 0$, the family $\{T^*(a)\}$ satisfies $e^*(f, g) \geq 1$. Now fix $\xi \leq 0$. By the definition of $f^*_\xi$ in (4.2), for any detection scheme $T$,

$$E_{f^*_\xi}(T) = \int_{\xi}^{0} E_{\theta}(T) \pi_\xi(\theta) d\theta, \tag{4.6}$$

Combining this with Lemma 3.2 yields

$$E_{f^*_\xi}(T^*(a)) \geq \int_{\xi}^{0} e^{I(\lambda, \theta)a} \pi_\xi(\theta) d\theta.$$

For any arbitrary $\eta > 0$, by the continuity of $\pi(\theta)$ in (2.2) and $I(\lambda, \theta)$ in (3.3) as well as the relation between $\pi(\theta)$ and $\pi_\xi(\theta)$, there exists a positive number $\delta$ such that $\delta + \xi \leq 0$ and for all $\theta \in (\xi, \xi + \delta)$,

$$I(\lambda, \theta) \geq (1 - \eta)I(\lambda, \xi) \quad \text{and} \quad \pi_\xi(\theta) \geq (1 - \eta)\pi_\xi(\xi).$$

The value of $\delta$ may depend on $\eta$ and $\xi$, but it does not depend on $a$ or the stopping time $T^*(a)$. Hence,

$$E_{f^*_\xi}(T^*(a)) \geq \int_{\xi}^{\xi + \delta} e^{I(\lambda, \theta)a} \pi_\xi(\theta) d\theta \geq e^{I(\lambda, \xi)(1 - \eta)a} (1 - \eta) \pi_\xi(\xi) \geq e^{I(\lambda, \xi)(1 - \eta)a} \delta (1 - \eta) \pi_\xi(\xi).$$

By Lemma 3.2, for sufficiently large value $a$,

$$\frac{\log E_{f^*_\xi}(T^*(a))}{I(\lambda, \xi) E_{\phi_2}(T^*(a))} \geq \frac{I(\lambda, \xi)(1 - \eta)a + \log(\delta(1 - \eta)\pi_\xi(\xi))}{I(\lambda, \xi)(a + (C + o(1))\sqrt{a})},$$

which converges to $1 - \eta$ as $a \to \infty$ because $\xi$, $\eta$, $\delta$ and $C$ does not depend on $a$. Thus, for the detection schemes $\{T^*(a)\}$, we have

$$e^*(f, g) = \lim_{a \to \infty} \frac{\log E_{f^*_\xi}(T^*(a))}{I(\lambda, \xi) E_{\phi_2}(T^*(a))} \geq 1 - \eta.$$

Because $\eta > 0$ is arbitrary, we have $e^*(f, g) \geq 1$. The theorem follows at once from the fact that $\xi \leq 0$ is arbitrary. $\square$
4.2. Expected False Alarm Rate

Although the proposed asymptotic efficiency $e^*(f, g)$ in Section 4.1 has theoretical meaning and seems to be good for evaluation of the overall quality of a detection scheme, it can be hardly considered as an appropriate characteristic for the false alarm rate only, and it perhaps will not satisfy practitioners who would like to see plots of the detection delay versus some reasonable false alarm rate measure in a particular problem. In addition, it is not clear how to run numerical simulations efficiently to calculate the asymptotic efficiency. For this reason, we propose another formulation with a new false alarm criterion: expected false alarm rate (EFAR).

To motivate our criterion, let us first look at the ARL to false alarm criterion when observations are i.i.d. under the prechange distribution. As we mentioned earlier, even for independent observations, there are some concerns on the ARL to false alarm criteria, see Barnard (1959), Kenett and Zacks (1998), Lai (2001), and Tartakovsky et al. (2006). But the general conclusion is that because the distribution of stopping times of widely used procedures such as Page’s CUSUM and Shiryayev–Roberts procedures are asymptotically exponential under the prechange hypothesis, the ARL to false alarm criterion is a reasonable measure of the false alarm rate for independent observations.

In the following, we provide another viewpoint of understanding the ARL to false alarm criterion. This viewpoint allows us to show that for any stopping time $T$ (no matter whether its distribution is asymptotically exponential or not), the reciprocal of the ARL to false alarm, $1/E_f(T)$, can always be thought of as the false alarm rate when the observations are i.i.d. under the prechange probability measure $P_f$.

To see this, imagine repeated applications of a detection scheme (stopping time) $T$ to the observations $X$’s under the probability measure $P_f$. That is, define stopping times $T_k$ inductively as follows: $T_1 = T$, and $T_{k+1} = T$ when applying $T$ to the observations $X'T_1 + T_{k+1}, X'T_2 + T_{k+2}, \ldots$. Because the $X'$s are i.i.d. under $P_f$, the stopping times $T_1, T_2, \ldots$ are defined everywhere and are mutually independent with identical distribution, which is the same as that of the detection scheme $T$. Now under $P_f$, the new stopping times $T_1, \ldots, T_k$ are simply repeated application of the detection scheme $T$, and they raise $k$ false alarms out of a total number $T_1 + \cdots + T_k$ of observations $X$’s. Thus under $P_f$, the false alarm rate of $T$ can be thought of as

$$\frac{k}{T_1 + \cdots + T_k},$$

which converges to $1/E_f(T)$ as $k$ goes to $\infty$ by the strong law of large numbers. Note that this approach has also been used in Blackwell (1946) to provide an alternative proof of Wald’s equation.

If the observations are not independent, then the above argument may still work if $T_1, T_2, \ldots, T_k$ are identically distributed and the strong law of large numbers for dependent observations is applicable to the $T_k$’s. In that case, $1/E_f(T)$ can still be thought of as the false alarm rate, even if the observations are not independent. The stationary autoregressive (AR) models in time series analysis seems to be one of such examples. Unfortunately, the above arguments may not work in general when the observations are not independent, and it is not clear whether it is reasonable to treat $1/E_f(T)$ as the false alarm rate.
Average Run Length to False Alarm

Nevertheless, this above approach shows that for the mixture prechange distribution $f$ defined in (2.3), although $E_r(T)$ may be misleading as a false alarm criterion under the mixture distribution $f$, $E_o(T)$ remains informative as a function of $\theta$, as the observations are i.i.d. under $P_\theta$. This motivates us to define the “expected false alarm rate” in the problem of detecting a change from the mixture prechange distribution $f$ in (2.3) to $g = \phi_2$. Specifically, under $P_\theta$, it is reasonable to define the the false alarm rate as $\text{FAR}_\theta(T) = 1/E_o(T)$. Using the definition of the mixture distribution $f$ in (2.3), it is nature to define the expected false alarm rate under $P_j$ as

$$\text{EFAR}_j(T) = \int_{-\infty}^{0} \text{FAR}_\theta(T) \pi(\theta) d\theta = \int_{-\infty}^{0} \frac{1}{E_o(T)} \pi(\theta) d\theta,$$

and then the problem can be formulated as follows: Minimize the detection delay $E_j(T)$ subject to the expected false alarm rate $\text{EFAR}_j(T) \leq \alpha$ for some $\alpha \in (0, 1)$.

Note that by Jensen’s inequality, we have

$$\frac{1}{E_j(T)} = \frac{1}{\int_{-\infty}^{0} E_o(T) \pi(\theta) d\theta} \leq \int_{-\infty}^{0} \frac{1}{E_o(T)} \pi(\theta) d\theta = \text{EFAR}_j(T).$$

That is, in our example, the reciprocal of the ARL to false alarm $E_j(T)$ is less than the proposed expected false alarm rate $\text{EFAR}_j(T)$. As we see in Section 2, the main contribution to $E_j(T)$ typically come from $E_o(T)$ with negatively large $\theta$ values (far away from the postchange distribution $g = \phi_2$), which may be undesirable. On the other hand, the main contribution to the expected false alarm rate $\text{EFAR}_j(T)$ tends to come from those $\theta$ values which are close to the postchange distribution $g = \phi_2$. Moreover, although there are schemes with infinite ARL to false alarm, there exist no schemes with zero expected false alarm rate $\text{EFAR}_j(T)$.

In statistics, besides mean, another widely used statistic to capture the distribution function of a random variable is median, or more generally, quantiles. It is interesting to point out that in our context, $\text{EFAR}_j(T)$ is closely related to the “quantile run length” of false alarm, which is defined as $\xi^q_j(T)$ as the value $u$ such that $P_j(T \leq u) = 1 - P_j(T > u) = q$ for some $0 < q < 1$. Note the the idea of the quantile run length of false alarm was expressed in Barnard (1959). The close relationship between EFAR and the quantile run length can be seen from the following heuristic argument.

Recall that under $P_\theta$, the observations are independent and thus $T$ is generally asymptotically exponential distributed with mean $E_o(T)$, implying that $P_j(T > u) \approx \exp(-u/E_o(T))$ for $u > 0$. By the definition of the quantile $\xi^q_j(T)$, we have

$$1 - q = P_j(T > \xi^q_j(T)) = \int_{\xi^q_j(T)}^{\infty} P_\theta(T > \xi^q_j(T)) \pi(\theta) d\theta \approx \int_{-\infty}^{0} \exp \left( - \frac{\xi^q_j(T)}{E_o(T)} \right) \pi(\theta) d\theta,$$

if we assume the results on asymptotic exponential distributions hold uniformly for all $\theta$. Now let us assume $\xi^q_j(T)/E_o(T)$ is small, which seems to be reasonable for small $q > 0$. Then using the fact that $\exp(-x) \approx 1 - x$ for small $x$, we have

$$1 - q \approx \int_{-\infty}^{0} \left( 1 - \frac{\xi^q_j(T)}{E_o(T)} \right) \pi(\theta) d\theta.$$ 

Combining this with the fact that $\int_{-\infty}^{0} \pi(\theta) d\theta = 1$ yields

$$\xi^q_j(T) \approx q \int_{-\infty}^{0} \frac{1}{E_o(T)} \pi(\theta) d\theta = \frac{q}{\text{EFAR}_j(T)}.$$
Therefore, EFAR\(_f\)/\(T\) provides useful asymptotic information about the quantile run length to false alarm, \(\xi^q_f(T)\), at least for small \(q > 0\).

It is also worth pointing out that the above arguments suggest us to consider a general false alarm criterion of the form

\[
H_f(T) = \int_{-\infty}^{0} h(E_0(T)) \pi(\theta) d\theta, \tag{4.7}
\]

where the non-negative function \(h(\cdot)\) preferably is decreasing or very slowly increasing so that the value of \(H_f(T)\) depends more on smaller values of \(\theta\), e.g., \(h(u) = 1/u\) or \(\exp(-u)\). Note that the ARL to false alarm, \(E_f(T)\), can also be thought of as a special case with \(h(u) = u\), but unfortunately, our results show that linear increase is too fast in our context.

It will be interesting to develop asymptotically optimal procedure under the expected false alarm rate criterion, or more generally under the criterion \(H_f(T)\) defined in (4.7). One possible candidate is the detection schemes \(T^*(a)\) in (3.5), which asymptotically maximizes \(E_0(T^*(a))\), or equivalently, optimizes \(h(E_0(T^*(a)))\), for every prechange \(\theta \leq 0\), subject to a constraint on the detection delay \(E_g(T)\). The detailed arguments are beyond the scope of this article and will be investigated elsewhere.

5. FURTHER EXAMPLES AND IMPLICATION

Although the specific example we considered is in essence the problem of detecting a change in the mean of normal observations, we want to point out that the ideas and results can be generalized to many other problems. In the following, we discuss the implication of our example in change-point problems with either exchangeable prechange models or hidden Markov models.

5.1. Exchangeable Prechange Models

It is straightforward to extend our ideas to the change-point problem in which the observations \(X_n\)'s are exchangeable under the prechange distribution \(f\). Here the \(X\)'s are exchangeable under a probability measure \(P_f\) if for each \(n \geq 1\), the \(n!\) permutations \((X_{k_1}, \ldots, X_{k_n})\) have the same \(n\)-dimensional joint probability distribution. To see this, by deFinetti’s representation theorem for exchangeable distribution and its extension in Hewitt and Savage (1955), in most cases of interest, an exchangeable distribution for a data sequence is a mixture of i.i.d. distributions. That is, for an exchange distribution \(f\), in general, we have \(f(x_1, \ldots, x_n) = \int_{0}^{1} \prod_{i=1}^{n} p_\theta(X_i) d\Pi(\theta)\), and thus our results for the mixture distribution can be easily extended to the problem with exchangeable prechange distributions. Note that many widely used models or methods lead to exchangeable distributions, see, for example, linear random-effects models as well as invariance reduction and weight-function approaches we mentioned in Section 2. These include the sequential \(\chi^2\), \(F\), and \(T^2\)-tests, as pointed out in Berk (1970).

In the following we discuss another concrete example to illustrate that the ARL to false alarm criterion may not be applicable for exchangeable prechange models.
Assume we are interested in detecting a change in distribution in a sequence of observations \( X_1, X_2, \ldots \), in which, for some unknown time \( v \),

\[
X_i = \begin{cases} 
Y_i - (a + b|Y|), & \text{if } i < v; \\
Y_i + (a + b|Y|), & \text{if } i \geq v.
\end{cases}
\]

where \( a, b > 0 \) are two known constants, \( Y_1, Y_2, \ldots \) are unobservable, independent random variables, \( Y_i \sim N(0, 1) \), and \( Y \) has a know density \( h(\cdot) \).

In this problem, both the prechange distribution \( f \) and the postchange distribution \( g \) are exchangeable. Let \( \mu_0 = a + bE|Y| \) and \( \sigma_0^2 = b^2\text{Var}(|Y|) \). Then it is straightforward to show that the \( X_i \)'s have mean \( -\mu_0 \) and variance \( 1 + \sigma_0^2 \) under the prechange distribution \( f \), and have mean \( \mu_0 \) and variance \( 1 + \sigma_0^2 \) under the postchange distribution \( g \). Moreover, conditional on \( v \), \( \text{cov}(X_i, X_j) = -1 - \sigma_0^2 \) if \( i < v < j \), i.e., the prechange and postchange observations are (negatively) correlated.

Thus, this problem can be thought of as detecting a change in the mean from \( -\mu_0 \) to \( \mu_0 \) (in which the observations are dependent). Motivated by Page's CUSUM procedure for independent observations, let us consider the detection scheme

\[
T(\gamma) = \inf \left\{ n : \max_{1 \leq k \leq n} \sum_{i=k}^{n} X_i \geq \gamma \right\}.
\]

Then for \( \gamma > 0 \), \( T(\gamma) \) will have finite detection delay \( E_x(T(\gamma)) \) but \( E_f(T(\gamma)) \) can be infinite.

To see this, first note that the detection delay \( E_x(T(\gamma)) = E_x(T(\gamma)) = E(h(|Y|)) \), where \( h(z) = E(N_{(1)}^{(v)}) \), and \( N_{(1)}^{(v)} \) is a new stopping time defined by

\[
N_{(1)}^{(v)} = \inf \left\{ n : \max_{1 \leq k \leq n} \sum_{i=k}^{n} (Y_i + a + bz) \geq \gamma \right\}.
\]

Because the \( Y_i \)'s are i.i.d. \( N(0, 1) \), the property of \( N_{(1)}^{(v)} \) follows immediately from the classical results of Page's CUSUM procedure with independent observations. In particular, for \( z > 0 \) we have \( h(z) \leq h(0) = E(N_{(1)}^{(v)}) = \gamma/a + O(1) \), implying that \( E_x(T(\gamma)) = E_x(T(\gamma)) = E(h(|Y|)) \leq h(0) \) is finite when \( a, b > 0 \).

Similarly, the ARL to false alarm \( E_x(T(\gamma)) = E(k(|Y|)) \), where \( k(z) = E(N_{(0)}^{(v)}) \), and \( N_{(0)}^{(v)} \) is a new stopping time defined by

\[
N_{(0)}^{(v)} = \inf \left\{ n : \max_{1 \leq k \leq n} \sum_{i=k}^{n} (Y_i - (a + bz)) \geq \gamma \right\}.
\]

Because the \( Y_i \)'s are i.i.d. \( N(0, 1) \), for \( z > 0 \), we have \( k(z) \geq \exp(2\gamma(a + bz)) \) by the classical result for Page's CUSUM procedure with independent observations. Thus

\[ E_f(T(\gamma)) = E(k(|Y|)) \geq E(\exp(2\gamma(a + b|Y|))). \]

Therefore, if the distribution of \( Y \) satisfies that \( E(\exp(2\gamma(a + b|Y|))) = \infty \) for some \( a, b > 0 \), then \( E_f(T(\gamma)) = \infty \) but \( E_x(T(\gamma)) \) is finite.

Note that our proposed alternative false alarm criteria in Section 4 can be easily extended to this example. For instance, for the detection scheme \( T(\gamma) \) in (5.1), if \( E_f(T(\gamma)) = \infty \), then the asymptotic efficiency (or the relative divergence rate) can
be defined by considering a sequence of problems in which \( Y \) is replaced by the truncated random variables \( Y_i = Y_i I(Y_i \leq \xi) \) as \( \xi \to \infty \). Here \( I(A) \) is the indicator function. Meanwhile, the expected false alarm rate can be defined by conditioning on \( Y \). Specifically, using the function \( k(\cdot) \) defined in \( E_T(T(\gamma)) \), we can define

\[
\text{EFAR}_T(T(\gamma)) = E \left( \frac{1}{k(\gamma)} \right) \leq E(\exp(-2\gamma(a + b|Y|))).
\]

It will be interesting to investigate whether \( T(\gamma) \) in (5.1) is asymptotically optimal under these two alternative formulations.

5.2. Hidden Markov Models (HMMs)

Surprisingly, our example is also closely related to the problem of detecting changes in hidden Markov models (HMMs) or state-space models, which have many important applications, including speech recognition and edge detection. Some general asymptotic theorems have been developed under the standard ARL to false alarm criterion, see, for example, Fuh (2003, 2004) and Lai (1995, 1998).

Let us first introduce hidden Markov models. Assume that \( U_1, U_2, \ldots \) is an unobserved Markov chain with states \( \{1, 2, \ldots, K\} \), transition probability matrix \( M = \{\pi(i, j)\}_{i,j=1,\ldots,K} \), and initial probability \( \pi = (\pi_1, \ldots, \pi_K) \). Given \( U_1, \ldots, U_n \), the observations \( X_i \)’s are conditionally independent, and given \( U_i \), \( X_i \) is independent of \( U_j \) for \( j \neq i \). Moreover, the conditional distribution of \( X_n \) given \( U_n \) does not depend on \( n \). We also assume the conditional distributions of \( X_n \) given \( U_n = i \) are dominated by a \( \sigma \)-finite measure, and denote by \( h_i(\cdot) \) the corresponding conditional density.

For our purpose, here we will focus on a special scenario of change-point problems in the hidden Markov model. Assume that initially the observations \( X_1, X_2, \ldots, X_{n-1} \) are from the hidden Markov model with initial probability \( \pi_0 = (\pi_1^0, \ldots, \pi_K^0, 0) \) with \( \sum_{k=1}^{K-1} \pi_k^0 = 1 \). At some unknown time \( \nu \), the observations \( X_\nu, X_{\nu+1}, \ldots \) are from the hidden Markov model with another initial probability \( \pi_1 = (0, \ldots, 0, 1) \). Here it is useful to think the state \( U_n = K \) as an absorbing state. The transition probability matrices and conditional densities \( h_i(\cdot) \) are the same before and after the change. The problem is to detect the true change as soon as possible, while raising as few false alarms as possible.

Note that this setting is more general than it appears to be at first sight. For example, if \( K = 2 \) and the transition probability matrix \( M = \) identity matrix, then the problem becomes the classical setting of detecting a change in distribution for independent observations, where the prechange distribution is \( f(x) = h_1(x) \) and the postchange distribution is \( g(x) = h_2(x) \). As another example, if \( K = 3 \) and the transition probability matrix

\[
M = \begin{pmatrix}
\rho & 1 - \rho & 0 \\
\rho & 1 - \rho & 0 \\
0 & 0 & 1
\end{pmatrix}
\]

for some \( 0 < \rho < 1 \), then using the standard probability calculation method in Markov chain, the problem again becomes the classical setting of detecting a change in distribution for independent observations. The only difference is that the prechange distribution \( f(x) = ph_1(x) + (1 - p)h_2(x) \) and the postchange distribution \( g(x) = h_3(x) \).
So far we have replicated the classical problem with independent observations, and thus Page’s CUSUM or Shiryayev–Roberts procedures are (asymptotically) optimal under the standard minimax formulation for these two special cases. Now let us look at another special case where $K = 3$ and the transition probability matrix $M = \text{identity matrix}$. Then the prechange distribution is the mixture distribution $f(X_1, \ldots, X_n) = \pi_0^1 \prod_{i=1}^n h_1(X_i) + \pi_0^2 \prod_{i=1}^n h_2(X_i)$ with $\pi_0^0 + \pi_0^1 = 1$, whereas the observations are i.i.d. with density $h_3(x)$ after a change occurs. In this case, Mei (2006a) showed that if both $\pi_0^1$ and $\pi_0^2$ are positive, Page’s CUSUM or Shiryayev–Roberts procedures are asymptotically suboptimal under the standard minimax formulation.

More generally, if the transition probability matrix $M$ is identity matrix, and the number of state $K$ is greater than 2 (possibly $\infty$), then the prechange distribution becomes a mixture (or exchangeable) distribution. This sheds light on the practical situation when the transition probability matrix $M$ is close to the identity matrix, particularly when the total sample size of the observations is only moderately large. In these situations, under the ARL to false alarm criterion, detecting larger changes may play more important role in the problem of detecting changes in the hidden Markov models.

6. DISCUSSIONS

A different choice of a prior distribution $\pi(\theta)$ in (2.2) will have an impact on whether or not a particular detection scheme $T$ has infinite ARL to false alarm $E_f(T)$. By (3.4), a detection scheme $T$ has an infinite ARL to false alarm, $E_f(T)$, if and only if $E_f(T) = \int_{-\infty}^0 (E_o(T))\pi(\theta)d\theta$ diverges. Thus, if $\pi(\theta)$ converges to 0 very fast as $\theta \to -\infty$, then $E_f(T)$ can be finite.

For instance, in our example, if we redefine $\pi(\theta)$ in (2.2) as $\pi(\theta) = \sqrt{2/\pi} \exp(-\theta^2/2)$ for $\theta \leq 0$. Then for any $b > 0$, Page’s CUSUM procedure $T_{CM}(\theta_0, b)$ defined in (2.1) will have a finite ARL to false alarm. However, $T^*(a)$ in (3.5) still has an infinite ARL to false alarm under this choice of $\pi(\theta)$ if $a \geq 1$.

Meanwhile, if we choose $\pi(\theta) = C_r \exp(-|\theta|^r)$ with $r > 2$ and the constant $C_r$ chosen so that $\int_{-\infty}^0 \pi(\theta)d\theta = 1$, then any detection scheme $T$ with finite detection delay $E_f(T)$ will also have finite ARL to false alarm, $E_f(T)$. To see this, for any stopping time $T$ satisfying the detection delay $E_f(T) = \lambda$, the maximum value of $E_o(T)$ for each $\theta \leq 0$ is of order $\exp(I(\lambda, \theta)\gamma) = \exp((\theta - \lambda)^2\gamma/2)$. Hence the choice of $\pi(\theta) = C_r \exp(-|\theta|^r)$ with $r > 2$ will guarantee that $E_f(T) = \int_{-\infty}^0 E_o(T)\pi(\theta)d\theta$ converges, implying that $E_f(T)$ is finite.

Indeed, most of our results in this article are not invariant to the choice of the prior distribution $\pi(\theta)$ for the unknown parameter $\theta$ in the distribution function of observations. One reviewer asked what should be done if we do not know the prior distribution $\pi(\theta)$. One possible approach is to use the general theory developed in Mei (2006a), which enables one to construct a family of detection schemes that asymptotically minimizes $E_o(T)$ for every (prechange) $\theta$ value without any knowledge of the prior $\pi(\theta)$ in a general context (not necessarily for normal distributions or other exponential families). For instance, $T^*(a)$ in (3.5) requires no knowledge of the prior $\pi(\theta)$, but it is asymptotically efficient under the divergence rate criterion and also seems to be asymptotically optimal in the sense
that it asymptotically minimizes the expected false alarm rate (EFAR) subject to a constraint on detection delay.

In our contexts, even if the ARL to false alarm $E_f(T)$ is finite, it does not mean that it is an appropriate false alarm criterion, as illustrated in Section 4.2. Specifically, in our example with mixture prechange distribution, the main contribution to $E_f(T)$ comes from $E_{\theta}(T)$ with negatively large $\theta$ values (far away from the postchange distribution $g = \phi_j$), implying that the ARL to false alarm criterion pays more attention to detecting larger changes, which is undesirable. Similar conclusions also hold for other exchangeable prechange distributions, or in the problem of detecting a change in hidden Markov models, especially if the transition matrix is close to identity matrix. Thus we should be cautious to use the ARL to false alarm criterion to assess detection schemes if we are more interested in detecting smaller changes.

Besides the minimax formulation, another widely used formulation in change-point problems is the Bayesian formulation, in which the change-point $v$ is a random variable with known a prior distribution. It is interesting to note that in our example with the mixture prechange distribution $f$ in (2.3), although the standard minimax formulation is inappropriate, the Bayesian formulation can still be applied, and both Page’s CUSUM and Shiryeyev–Roberts procedures are asymptotically optimal under the Bayesian formulation. This is because the false alarm criterion in the Bayesian formulation is $P_f(T \leq v)$, which is asymptotically equivalent to $P_{\theta}(T \leq v)$ with $\theta = 0$. Thus, the Bayesian formulation of detecting a change from the mixture prechange distribution $f$ to $g = \phi_j$ is asymptotically equivalent to that of detecting a change from $\theta = 0$ to $\lambda$ in the mean of normally distributed random variables.

It is possible that the Bayesian formulation, the alternative formulations proposed here, or other formulations such as Quantile Run Length, may not be appropriate to other change-point problems with dependent observations. In fact, a message of this article is that well-known theorems and results in classical change-point problems could no longer hold for dependent observations, at least for exchangeable prechange distributions. Although this article is “incomplete” in the sense that it does not develop general theory in a general setting, we hope it is adventurous enough to give theoreticians and practitioners something serious to contemplate.

**APPENDIX A: PROOF OF LEMMA 3.1**

*Proof of Lemma 3.1.* Relation (3.1) with $\theta = \theta_0$ and relation (3.2) are well-known for Page’s CUSUM procedure $T_{CM}(\theta_0, b)$ defined in (2.1), see, for example, Lorden (1971) or Siegmund (1985). It suffices to show relation (3.1) holds for general $\theta$.

To prove this, note that Page’s CUSUM procedure $T_{CM}(\theta_0, b)$ defined in (2.1) can be rewritten as

$$T_{CM}(\theta_0, b) = \text{first } n \geq 1 \text{ such that } \max_{1 \leq k \leq n} \sum_{i=k}^{n} \left[ X_i - \frac{\lambda + \theta_0}{2} \right] \geq \frac{b}{\lambda - \theta_0},$$

$$= \text{first } n \geq 1 \text{ such that } \max_{1 \leq k \leq n} \log \frac{\phi_{\theta_0}(X_i)}{\phi_{\theta}(X_i)} \geq \frac{\lambda + \theta_0 - 2\theta}{\lambda - \theta_0} b.$$
where \( \hat{\lambda}_1 = \lambda + \theta_0 - \theta \). Hence, \( T_{CM}(\theta_0, b) \) can also be thought of as Page’s CUSUM procedure in the problem of detecting a change in the mean from \( \theta \) to \( \hat{\lambda}_1 = \lambda - \theta_0 - \theta \) with log-likelihood boundary \( b_1 = b(\lambda + \theta_0 - 2\theta)/(\lambda - \theta_0) \). Applying the classical results on Page’s CUSUM procedures again, \( E_\theta(T_{CM}(\theta_0, b)) \geq \exp(b_1) \), completing the proof of (3.1).

\[ \Box \]

**APPENDIX B: PROOF OF THEOREM 4.1**

*Proof of Theorem 4.1.* Denote by \( P^{(\gamma)}_{f, g} \) the probability measure when the distribution of the observations \( X \)'s changes from \( f_\gamma \) to \( g \) at time \( v \). To simplify notations, we also denote by \( f_\gamma(\cdot | X_1, \ldots, X_{i-1}) \) and \( g(\cdot | X_1, \ldots, X_{i-1}) \) the conditional density functions of \( X_i \) given \( X_1, \ldots, X_{i-1} \) under the probability measure \( P_{f_i} \) and \( P_g \), respectively. Note that \( g(\cdot | X_1, \ldots, X_{i-1}) \) because observations are independent under \( P_g \). Then the conditional log-likelihood ratio of \( X_i \) given \( X_1, \ldots, X_{i-1} \) is \( Z_i = \log \left( g(X_i | X_1, \ldots, X_{i-1}) / f_\gamma(X_i | X_1, \ldots, X_{i-1}) \right) \).

To prove (4.3), it suffices to show that the constant \( I = I(\lambda, \xi) \) satisfies Lai’s sufficient condition in (3.8), with \( P^{(\gamma)}_{f_1, g} \) replacing by \( P^{(\gamma)}_{f_1, g} \). To do so, under \( P^{(\gamma)}_{f_1, g} \), define

\[
\pi^{(\gamma)}(\theta) = \frac{\phi_\theta(X_1) \cdots \phi_\theta(X_{i-1}) \pi_\gamma(\theta)}{f_\gamma(X_1) \cdots \phi_\theta(X_{i-1})} d\theta.
\]

then \( f_\gamma \pi^{(\gamma)}(\theta) d\theta = 1 \), and thus

\[
\sum_{i=1}^{\infty} \log \frac{\phi_\lambda(X_i) \cdots \phi_\lambda(X_{i+1})}{\phi_\theta(X_i) \cdots \phi_\theta(X_{i+1})} \pi_\gamma(\theta) d\theta 
\leq \log \frac{\phi_\lambda(X_i) \cdots \phi_\lambda(X_{i+1})}{\min_{\xi \in \theta \in \theta_0} \left[ \phi_\gamma(X_i) \cdots \phi_\lambda(X_{i+1}) \right]}
= \max_{\xi \leq \theta \leq \theta_0} \sum_{i=1}^{\infty} \log \frac{\phi_\lambda(X_i)}{\phi_\theta(X_i)}.
\]

Because under \( P^{(\gamma)}_{f_1, g} \), \( X_1, X_2, \ldots \) are i.i.d. normal with mean \( \lambda \) and variance 1, no matter what we observed \( X_1, \ldots, X_{i-1} \), Lai’s sufficient condition in (3.8) holds if we can show that for all \( \delta > 0 \),

\[
\lim_{n \to \infty} P_{g} \left\{ \max_{\xi \leq \theta \leq \theta_0} \sum_{i=1}^{t} \log \frac{\phi_\lambda(X_i)}{\phi_\theta(X_i)} \geq I(\lambda, \xi) (1 + \delta) n \right\} = 0. \tag{B.1}
\]

Now it suffices to show that (B.1) holds.

To prove (B.1), let \( \bar{X} = \sum_{i=1}^{t} X_i/t \), then it is easy to see that

\[
\sum_{i=1}^{t} \log \frac{\phi_\lambda(X_i)}{\phi_\theta(X_i)} = t \left( \frac{(\theta - \bar{X})^2}{2} - (\lambda - \bar{X})^2 \right), \tag{B.2}
\]

and the maximum of (B.2) over \( \xi \leq \theta \leq 0 \) is attained at \( \theta = \xi \) if \( \bar{X} > 0 \). Moreover, \( \bar{X} \) will always be positive if \( t \) is sufficiently large. This suggests us splitting
the probability in (B.1) into three parts, depending on whether \( t \leq \sqrt{n} \) or \( \min_{\sqrt{n} \leq t \leq \bar{n}} \overline{X}_t > 0 \). Specifically, define

\[
C1 = \mathbb{P}_g \left\{ \max_{t \leq \sqrt{n}} \max_{i \leq \theta \leq 0} \sum_{i=1}^{t} \log \frac{\phi_g(X_i)}{\phi_{\theta}(X_i)} \geq I(\lambda, \zeta)(1 + \delta)n \right\},
\]

\[
C2 = \mathbb{P}_g \left\{ \max_{\sqrt{n} \leq t \leq \bar{n}} \max_{i \leq \theta \leq 0} \sum_{i=1}^{t} \log \frac{\phi_g(X_i)}{\phi_{\theta}(X_i)} \geq I(\lambda, \zeta)(1 + \delta)n; \ \min_{\sqrt{n} \leq t \leq \bar{n}} \overline{X}_t \geq 0 \right\},
\]

\[
C3 = \mathbb{P}_g \left\{ \max_{\sqrt{n} \leq t \leq \bar{n}} \max_{i \leq \theta \leq 0} \sum_{i=1}^{t} \log \frac{\phi_g(X_i)}{\phi_{\theta}(X_i)} \geq I(\lambda, \zeta)(1 + \delta)n; \ \min_{\sqrt{n} \leq t \leq \bar{n}} \overline{X}_t < 0 \right\}.
\]

Then the probability in the left-hand side of (B.1) is less than or equal to \( C1 + C2 + C3 \). Therefore, it now suffices to show that each of \( C1, C2 \) and \( C3 \) goes to 0 as \( n \rightarrow \infty \).

For the \( C1 \) term, by (B.2), we have

\[
\max_{i \leq \theta \leq 0} \sum_{i=1}^{t} \log \frac{\phi_g(X_i)}{\phi_{\theta}(X_i)} \leq \frac{t}{2} \max_{i \leq \theta \leq 0} (\theta - \overline{X}_t)^2 \leq \frac{t}{2} \left[ (\zeta - \overline{X}_t)^2 + (\overline{X}_t)^2 \right]
\]

because \((\theta - \overline{X}_t)^2\) is a convex function of \( \theta \) and the maximum is attained at either \( \theta = \zeta \) or \( \theta = 0 \). Thus

\[
C1 \leq \mathbb{P}_g \left\{ \max_{t \leq \sqrt{n}} \frac{t}{2} \left[ (\zeta - \overline{X}_t)^2 + (\overline{X}_t)^2 \right] \geq I(\lambda, \zeta)(1 + \delta)n \right\}
\]

\[
\leq \mathbb{P}_g \left\{ \max_{t \leq \sqrt{n}} \left[ (\zeta - \lambda)^2 + (\overline{X}_t)^2 \right] \geq 2I(\lambda, \zeta)(1 + \delta)\sqrt{n} \right\}
\]

using the fact that \( t \leq \sqrt{n} \) in the \( C1 \) term. As \( n \rightarrow \infty \), the term \((\zeta - \overline{X}_t)^2 + (\overline{X}_t)^2\) converges to \((\zeta - \lambda)^2 + \lambda^2\) under \( \mathbb{P}_g \), whereas \( 2I(\lambda, \zeta)(1 + \delta)\sqrt{n} \) goes to \( \infty \). This implies that \( C1 \rightarrow 0 \) as \( n \rightarrow \infty \).

For the \( C2 \) term, we have \( \overline{X}_t \geq 0 \) and the maximum over \( \zeta \leq \theta \leq 0 \) is attained at \( \theta = \zeta \), so

\[
C2 \leq \mathbb{P}_g \left\{ \max_{1 \leq t \leq n} \sum_{i=1}^{t} \log \frac{\phi_g(X_i)}{\phi_{\zeta}(X_i)} \geq I(\lambda, \zeta)(1 + \delta)n \right\}.
\]

The strong law implies that under \( \mathbb{P}_g \),

\[
\frac{1}{n} \max_{1 \leq t \leq n} \sum_{i=1}^{t} \log \frac{\phi_g(X_i)}{\phi_{\zeta}(X_i)} \rightarrow I(\lambda, \zeta)
\]

with probability 1. Thus for any \( \delta > 0 \), the term \( C2 \rightarrow 0 \) as \( n \rightarrow \infty \).

For the \( C3 \) term, we have \( \min_{\sqrt{n} \leq t \leq \bar{n}} \overline{X}_t \rightarrow \lambda > 0 \) with probability 1 under \( \mathbb{P}_g \), and thus the term \( C3 \rightarrow 0 \) as \( n \rightarrow \infty \). This completes the proof of Theorem 4.1. \( \square \)

**ACKNOWLEDGMENTS**

The author would like to thank his advisor, Dr. Gary Lorden, for his support, guidance, and encouragement, and Drs. Moshe Pollak and Yi-Ching Yao for
stimulating discussions. The author would also like to thank the reviewers for their thoughtful comments. This work was supported in part by Dr. Sarah Hotle’s National Institutes of Health under grant R01 AI055343.

REFERENCES


Discussion on “Is Average Run Length to False Alarm Always an Informative Criterion?” by Yajun Mei

Boris Brodsky¹ and Boris Darkhovsky²
¹Department of Statistics Econometrics, Central Institute for Mathematics and Economics, Moscow, Russia
²Department of Macrosystems Dynamics, Institute for Systems Analysis RAS, Moscow, Russia

Abstract: Criteria of optimality in change-point detection are discussed.

Keywords: Asymptotic optimality; Change-point detection; Supremal probability of a false decision.

Subject Classifications: 62L10; 62L15; 62L99.

1. INTRODUCTION

The paper written by Dr. Mei is of substantial interest to specialists in sequential analysis. The author demonstrates that the well-known ARL to false alarm criterion can be inappropriate for evaluating one of the main operating characteristics of a change-point detection method. Then the author proposes two alternative criteria: the relative divergence rate and expected false alarm rate—for evaluation of false alarm characteristics of a method. After that Dr. Mei proposes an asymptotically optimal method with respect to a new minimax formulation of sequential change-point detection problem involving the divergence rate criterion. In conclusion the author discusses practical implications of his results for exchangeable prechange models and hidden Markov models.

The widespread opinion “Choice of a criterion is the matter of taste” is not applicable to this situation due to the practical importance of change-point detection problems. In these problems we should understand the real practical sense of our optimality criterion.

Received March 31, 2008, Revised June 6, 2008, Accepted June 15, 2008
Recommended by Alex Tartakovsky
Address correspondence to Boris Darkhovsky, Department of Macrosystems Dynamics, Institute for Systems Analysis RAS, 9, prospekt 60-letya Oktyabria, Moscow 117312, Russia; E-mail: darbor@isa.ru
This paper is instructive and important for the following reasons. The author’s message is clear and fully justified: “In change-point problems with mixture prechange models detection methods with finite detection delays can have infinite ARLs to false alarms.” The author considers the case of dependent observations (in a specific case of mixture models before a change-point), which is quite rarely analyzed in the literature.

In our opinion, the relative divergence rate criterion is really good for evaluation of the overall quality of a change-point detection method but can be hardly considered as an appropriate characteristic for estimation of the false alarm rate only.

2. ALTERNATIVE CRITERIA

Alternative criteria of false alarm rate were introduced much earlier (see Brodsky and Darkhovsky, 1990, 1993, 2000). We mean the supremal probability of a false decision at each step of decision making.

Let us discuss this issue in more details. Consider the following problem: to detect an instant of a change from the density function (d.f.) $f_0$ to $f_1$ in a sequence of independent random variable’s (r.v.) $x_1, x_2, \ldots$ Let us introduce a decision function $d(n)$ taking values 1 and 0, e.g., for CUSUM method, $d^{CS}(n) = \mathbb{I}(\max_{1 \leq k \leq n} \sum_{i=1}^{n} \ln \frac{f_1(x_i)}{f_0(x_i)} > C)$ ($\mathbb{I}(A)$ is the indicator of the set $A$).

The stopping time $\tau$ is defined as follows $\tau = \min \{n : d(n) = 1\}$. For $d^{CS}(n)$, we obtain the stopping time of the CUSUM method.

Clearly, we have the strict inclusion $(\tau = n) \subset (d(n) = 1)$.

Define the following supremal probability of false decision

$$
\sup_n P_\infty \{d(n) = 1\} \overset{\text{def}}{=} \gamma
$$

This value is closely connected with the false alarm probability. As Dr. Mei remarks in his paper, the traditional “false alarm probability” equals 1 if the average delay time in change-point detection is finite. However, this is not the case for the supremal probability of false decision.

For an integer-valued r.v. $\zeta$, given the relationship $\sup_n P\{\zeta = n\} \leq a$, it can be shown that $E\zeta \geq (2a)^{-1}(1 + o(1))$ as $a \to 0$ and this inequality is strict.

Hence, we obtain $E\zeta \geq (2\gamma)^{-1}(1 + o(1))$ as $\gamma \to 0$, where $\gamma$ is the supremal probability of false decision.

Therefore, given the value of the supremal probability of false decision, we can estimate ARL to false alarm from below.

Remark now that all well-known stopping rules depend on a certain “large parameter” $N$ such that the probability of a false decision tends to zero as the large parameter goes to infinity. Very often this large parameter is simply a decision threshold (as in CUSUM test), but it is not necessary.

In 1990 we found that under very general conditions on (dependent) sequence $\{x_n\}$, there exist (under any fixed change-point moment $m$) the limits

$$
\lim_{N \to \infty} \frac{(\tau - m)^{+}}{N} = T \ P_m \ a.s., \quad \lim_{N \to \infty} \frac{[\ln \gamma]}{N} = \Gamma
$$

These limits depend only of a detection method, the size of change, the mixing conditions, and the tails of distributions.
Now we can introduce the following performance criterion of change-point detection:

$$\mathcal{K} = \frac{E_m(\tau - m)^+ + 1}{|\ln \gamma|}$$  \hspace{1cm} (2.2)

Taking into account (2.1), it is easy to see that this criterion is the ratio (in an appropriate scale) of the average delay time to the worst (i.e., the minimal) average time before a false decision.

For this criterion and all sequential methods of change-point detection that satisfy some nonrestrictive assumptions, we can prove the a priori nonasymptotic theoretical lower bound:

$$\mathcal{K} \geq \Gamma^{-1}(1 - L/|\ln \gamma|),$$ \hspace{1cm} (2.3)

where $L$ is a certain constant, $\mathcal{I}$ is the Kullback–Leibler information.

A method of change-point detection we call optimal if the equality sign in this lower bound is attained for it.

This approach is free from all drawbacks mentioned in Dr. Mei’s paper. However, it satisfies all practical requirements to change-point detection methods. Moreover, this approach can be generalized to the case of composite hypotheses both before and after a change-point (see Brodsky and Darkhovsky, 2008, and also an earlier sketch of this approach announced in Brodsky and Darkhovsky, 2005). Here we would like to mention that, from our point of view, it is more practical to consider parametric families of d.f.’s instead of mixtures of d.f.’s as in Dr. Mei’s paper.

It turns out that the CUSUM method for simple hypotheses and the minimax CUSUM method for composite hypotheses are asymptotically (as $N \rightarrow \infty$) optimal according to criterion (2.2) and the above definition. It was quite interesting for us to know that Dr. Mei (Mei, 2006) also obtained some kind of the minimax CUSUM method by another approach.

3. SOME FURTHER DIRECTIONS OF RESEARCH

In conclusion, we would like to give a brief description of some further directions of this approach. In many papers the following delay measure in change-point detection is widely used: $E_m(\tau - m \mid \tau \geq m)$. This measure was introduced by Shiryaev in his first works.

We can generalize inequality (2.3) for this measure. Suppose for each $n$ the following condition is satisfied

$$\sup_{k \geq 1} P_\infty(\tau = n + k \mid \tau \geq n) \leq \gamma$$

Then the following nonasymptotical a priori inequality holds for any method under certain nonrestrictive assumptions:

$$\frac{E_m(\tau - m \mid \tau \geq m)}{|\ln \gamma|} \geq \Gamma^{-1}(1 - M/|\ln \gamma|),$$ \hspace{1cm} (3.1)

where $M$ is a certain constant.
This inequality can be further generalized to the case of composite hypotheses. Inequalities of types (2.3) and (3.1) can be proved also for dependent observations, e.g., for the case when some mixing conditions are satisfied. In this case the Kullback-Leibler information in (2.3) and (3.1) is substituted by an analogous value for multidimensional distributions. For example, in the case of switching Markov sequences we use the following value:

\[
\left( \int \ln \frac{\varphi_1(u \mid x)}{\varphi_0(u \mid x)} \varphi_1(u \mid x) f_1(x) dx \, du \right)^{-1}
\]

where \(\varphi_i(u \mid x), i = 0, 1\) are densities of transfer probabilities, \(f_1(x)\) is the one-dimensional density after a change.

We think these thoughts are paralleled by many findings of Dr. Mei in his last papers. These findings are important and instructive for specialists in modern sequential analysis.

REFERENCES


**Discussion on “Is Average Run Length to False Alarm Always an Informative Criterion?” by Yajun Mei**

**John A. D. Aston,¹ Cheng-Der Fuh,² and Sheng-Feng Luo³**

¹Department of Statistics, University of Warwick, Coventry, UK
²Graduate Institute of Statistics, National Central University, Jhongli, Taiwan, ROC
³Institute of Statistical Science, Academia Sinica, Taipei, Taiwan, ROC

**Abstract:** This paper gives a few remarks and comments on the paper by Yajun Mei. These are primarily concerned with the relationship between the method proposed by the author and the case of Hidden Markov models studied in the paper.

**Keywords:** Change-point detection; Hidden Markov models.

**Subject Classifications:** 60B15; 60F05; 60K15.

1. **INTRODUCTION**

The authors would like to congratulate Yajun Mei on a very interesting and thought provoking piece of work. Indeed it has been long thought that there are issues with the Average Run Length criteria, but these have now been rigourously investigated in this paper, and several deficiencies highlighted. The two new criteria proposed, along with possible related criteria that could also be formulated, will provide an interesting avenue of further research.

Dr. Mei gave some special cases of the research with reference to Hidden Markov Models (HMMs) and it will be to these and possible extensions that we address our particular discussion of the paper.

2. **IRREDUCIBLE HMMs AND STRUCTURAL CHANGE**

The paper highlights a very specific class of HMMs, ones where the HMM is defined to be in a known state at the change-point. It is not completely clear as to whether...
the generality of the definition, even in the case where \( K = 2 \) is completely justified. Indeed, the thought that this state should be considered as an absorbing state is slightly confusing, in so much that it is the communicating classes of the states that are important not their necessity of being an absorbing state. Take, for example, the transition matrix

\[
\begin{pmatrix}
p & 1 - p \\ q & 1 - q
\end{pmatrix}, \quad 0 < p, q < 1, \quad p \neq q.
\] (2.1)

Although this formulation violates the “thought” of having the second state as an absorbing state, it is of interest to consider what happens in this case. Given the restriction from the paper that \( \pi_1^0 = 1 \) here, the Markov chain starts in state 1, and at some point (given a change-point does occur) will be guaranteed to be in state 2. However, the chain will pass through state 2 infinitely often, and due to the irreducible nature of the HMM in this case, both the model when a change-point is present, and that where a change-point is not present have the same limiting distribution. Indeed, even if \( q \) was zero, making state 2 an absorbing state, the same would hold. It would therefore appear that the chain not being irreducible is a fundamental requirement, and moreover that the communicating classes of the chain must be distinct between the two initial distributions, the one before and the one after the change-point.

Given this irreducibility and communicating class restriction, a more intuitive understanding of the HMM case given in the paper is perhaps the following. The two models separated by the change-point are an HMM under \( P_f \) and an i.i.d. model under \( P_g \). Given this definition, it raises the immediate question as to how things would work under the setting that under \( P_g \), the model was also an HMM but one with differing structure, either in the parameters or the transition probabilities, or even in the number of states. This would lead to the more usual definition of a change point in an HMM model, where each HMM is usually considered to be irreducible.

To illustrate this idea, we study a two-state hidden Markov model with transition probability (2.1). We consider the case when \( q \) changes to 0. The parameters we choose are as follows: initial distribution \( \pi' = [0.5, 0.5] \), \( p = 0.9 \), state 1 is \( N(0, 1) \), and state 2 is \( N(0.5, 1) \). Let \( \tau \) be the stopping time of CUSUM procedure (cf. (5.1) of Fuh, 2003), which is implemented by particle filter. For different \( q \), take different boundary \( b \) to match the independent case. Precisely speaking, \( b \) is chosen for each case by simulations such that the ARL to false alarm is about 800. In the independent case, equation (14) of Pollak and Siegmund (1985) can be used to approximate the boundary \( b \). Simulations based on Theorem 6 of Fuh (2004) to get the boundary, and other computational issues will be published in a separate paper. Table 1 was obtained from 1000 runs of 2000 observations having a change point at time one.

Note that the number in parentheses stands for the percentage of runs that truly hit the boundary. This example showed that the ARL is still a reasonable criterion. An example of showing that CUSUM is suboptimal under ARL criterion can be found in Tartakovsky and Blanding (2006) and Blanding et al. (2007).

3. DETERMINING PARAMETER CHANGE IN A HMM

To investigate the problem of determining parameter change, as opposed to structural change, and how the Average Run Length criteria performed in practice
in the setting of an HMM, a small simulation was undertaken. A two-state irreducible HMM was set up and the CUSUM statistic investigated for a constant threshold (this could easily be replaced with the Shiryaev–Roberts criteria if so desired). In order to further extend the case from that of the paper, the independence requirement was also relaxed. Therefore this model could be considered a Markov switching model if the term HMM is used to indicate only those models with the independence assumption. The setup for the model is the following: let $X_t$, $t = 1, 2, \ldots$ be a two-state Markov chain with state space $D = \{1, 2\}$, and transition probability matrix $P = \begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix}$. Assume the initial state $X_0 \sim \pi = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$ is the stationary distribution. We consider the Markov switching model

$$Y_t = \alpha_{X_t} Y_{t-1} + \varepsilon_t,$$

(3.1)

where $Y_0 \sim N(0, 1)$, $\varepsilon_t \sim N(0, 1)$, $\alpha_1 = 0.9$, and $\alpha_2 = 0.7$.

Table 2 was obtained from 1000 runs of 2000 observations having a change point at time one (if one is present).

<table>
<thead>
<tr>
<th>$\mu$</th>
<th>$\tau$ (%)</th>
<th>$b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>799.2 (93.8)</td>
<td>6</td>
</tr>
<tr>
<td>0.25</td>
<td>145.5 (100)</td>
<td>6</td>
</tr>
<tr>
<td>0.5</td>
<td>47.7 (100)</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>15.4 (100)</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>4.5 (100)</td>
<td>6</td>
</tr>
</tbody>
</table>

Varying $\alpha_2$: change, $\alpha_2 \rightarrow 0.9$

<table>
<thead>
<tr>
<th>$\alpha_2$</th>
<th>$\tau$ (%)</th>
<th>$b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7</td>
<td>173.7 (100)</td>
<td>6</td>
</tr>
<tr>
<td>0.5</td>
<td>113.4 (100)</td>
<td>12.6</td>
</tr>
<tr>
<td>0.3</td>
<td>89.6 (100)</td>
<td>17.38</td>
</tr>
<tr>
<td>0.1</td>
<td>79.4 (100)</td>
<td>25.722</td>
</tr>
</tbody>
</table>
As can be seen, the ARL, although large for the no-change case, is bounded away from the end, even taking into account the finite nature of the samples. This suggests that the ARL criteria might be more useful in this case than expected under the assumptions in the paper.

REFERENCES


Discussion on “Is Average Run Length to False Alarm Always an Informative Criterion?” by Yajun Mei

Tze Leung Lai¹ and Hock Peng Chan²
¹Department of Statistics, Stanford University, Stanford, California, USA
²Department of Statistics and Applied Probability, National University of Singapore, Singapore

Abstract: We comment on Mei’s critique of the standard minimax formulation with the ARL to false alarm of a sequential detection rule as its operating characteristic, and provide an alternative Bayesian approach for his use of a mixing distribution to handle pre-change parameters. We also consider computational issues in implementing the ARL and better alternatives for evaluating sequential detection rules in complex stochastic systems.

Keywords: Bayesian change-point; Hidden Markov models; Importance sampling and resampling.

Subject Classifications: 62L15; 60G40; 62F15.

1. INTRODUCTION

This discussion consists of two parts. The first part shows that the probability of false alarm per unit time is a better alternative to the ARL for evaluating the performance of a sequential detection procedure. The second part considers Professor Mei’s critique of the ARL to false alarm as a performance measure in his example in Section 3, and provides an alternative Bayesian approach which we find to be more natural for such problems.

2. PROBABILITY OF FALSE ALARM PER UNIT TIME

Before commenting on Mei’s information-theoretic critique of the ARL constraint on the false alarm of a sequential detection rule, we would like to supplement
it with implementation considerations that are more computational in nature. As noted by Lai (1995, pp. 630–631), the ARL constraint \( E_0 T \geq \gamma \) stipulates a long expected duration to false alarm, where \( E_0 \) denotes expectation under the baseline distribution. However, a large (and even infinite) mean of \( T \) does not necessarily imply that the probability of having a false alarm before some specified time is small, and it is easy to construct positive integer-valued random variables \( T \) having both a large mean and a high probability that \( T = 1 \). Moreover, for efficient detection rules in complex systems, \( E_0 T \) has to be computed by Monte Carlo and it is only an inner loop in a series of computations to determine the design parameters of \( T \) so that \( E_0 T \) is approximately equal to the constraint \( \gamma \). It is prohibitively difficult to simulate \( E_0 T \) when \( \gamma \) is large. Moreover, Mei’s examples in the paper suggest that it may even be impossible to simulate the baseline ARL of certain rules because they can be arbitrarily large.

Instead of imposing a constraint on the ARL, Lai (1995, p. 631) proposed to impose a constraint on the probability of “false alarm per unit time,” defined as \( T \) for sufficiently large (but not too large) \( m \). He noted that for many detection rules in practice, which do not include the examples with \( E_0 T \) considered by Mei,

\[
E_0 T \sim m/P_0[T \leq m]
\]  

uniformly in \( m = o(E_0 T) \) as \( E_0 T \to \infty \). Clearly (2.1) holds for the classical Shewhart chart based on independent observations; in fact, in this case, \( P_0[T = 1] = 1/E_0 T \) because \( T \) has a geometric distribution. Moreover, (2.1) also holds as \( E_0 T \sim \gamma \to \infty \) and \( m/\log \gamma \to \infty \) but \( \log m = o(\log \gamma) \) for the class of window-limited generalized likelihood ratio (GLR) detection rules developed by Lai (1995, 1998), Lai and Shan (1999), and Chan and Lai (2005), not only when the observations are independent but also under Markovian settings for which Monte Carlo simulations are needed to compute either the left-hand side or the right-hand side of (2.1). Whereas Mei uses mixture likelihood ratios to handle the case of unknown nuisance parameters in the baseline distribution, more convenient GLR rules and their window-limited variants can be used; see Siegmund and Venkatraman (1995, pp. 263–264) and Lai (1995, pp. 638–639).

As noted by Hunter (1990) and Lai (1995, p. 620), although the ARL is a reasonable performance measure for the simple Shewhart and CUSUM charts based on independent observations, it is difficult to analyze and conceptually unsatisfactory for more complicated detection rules based on non-independent observations. Lai (1995, 1998) proposed alternative performance criteria that are less simplistic and more tractable. One such criterion is an upper bound \( x \) on \( m^{-1}P_0[T \leq m] \), which for many window-limited detection rules is asymptotically equivalent to \( \sup_{v \geq 1} m^{-1}P_0[v \leq T < v + m] \) as \( x \to 0 \), \( m/|\log x| \to \infty \) and \( \log m = o(\log x) \); see Lai (1995, p. 631). As shown by Lai (1998, p. 2920), under a constraint on \( \sup_{v \geq 1} m^{-1}P_0[v \leq T < v + m] \), an asymptotic lower bound for \( E^{(\gamma)}(T - v)^+ \) holds uniformly in \( v \geq 1 \) under a much weaker condition than Mei’s (3.8), which is not satisfied by the example in Section 3 and which Lai (1998) uses to develop the asymptotic lower bound for \( \sup_{v \geq 1} \text{ess sup} E^{(v)}(T - v + 1)^+[X_1, \ldots, X_{v-1}] \) under the ARL constraint \( E_0 T \geq \gamma \), where \( E^{(\gamma)} \) denotes the probability measure under which the change-point occurs at time \( v \) as in Mei’s paper.
To implement this constraint on the false detection probability per unit time for complex stochastic systems by Monte Carlo evaluation of $P_0\{T \leq m\}$, Lai and Shan (1999) and Chan and Lai (2005, 2007) have introduced efficient importance sampling methods that use optimal importance densities that yield frequent visits to $T \leq m$, even though $P_0\{T \leq m\}$ is too small to be computable by direct Monte Carlo. To address long-standing difficulties with optimal importance densities that are available in theory but are hard to compute or to simulate from, Chan and Lai (2008) have recently developed flexible sequential importance sampling and resampling (SISR) schemes that use resampling to tilt the empirical measure of the sequentially generated sample paths towards the optimal importance measure. Due to the dependence of the SISR samples, the usual standard error estimator cannot be used, and Chan and Lai (2008) have provided a consistent estimate of the standard error of the simulated probability.

3. BAYESIAN MODELING OF CHANGE-POINTS

Returning to Mei’s critique and in particular the “severe criticism of the standard minimax formulation with the ARL to false alarm as an operating characteristic of a detection scheme” in the first paragraph in Section 3, we find it not surprising that “finite detection delay may be achieved with infinite ARL to false alarm.” In fact, by using the approach of Pollak and Siegmund (1975, Section 6), it is even possible to achieve such finite detection delay under probability constraints of the type $P_0\{T < \infty\} \leq x$. A shortcoming of the asymptotic theory of these detection rules is that it is applicable to detecting only large changes, but not smaller ones for which $I(\theta_0, \theta)$ also approaches to 0 as $x \to 0$. Mei has made a similar criticism in his paragraph on the primary goal of the paper in Section 1 and in another paragraph on the third implication of his example in Section 3. By using a range of window sizes in the composite GLR window-limited rule, Lai (1995, Section 3.2) shows that the asymptotic lower bound for the post-change ARL can be attained uniformly over changes of various magnitudes.

Mei’s use of a mixing distribution to handle the unknown pre-change parameter is akin to a Bayesian approach, which he does not use for the post-change parameter. We find it much more natural to adopt a full Bayesian change-point model for the unknown pre- and post-change parameters and the unknown change-time. Lai and Xing (2008) have recently used hidden Markov models (HMMs) for Bayesian modeling of change-points and have developed BCMIX (bounded complexity mixture) approximations to the HMM filters, which they use not only for sequential change-point detection, but also for recursive estimation of the time-varying parameters and adaptive control of stochastic systems whose parameters may undergo occasional changes. The BCMIX filters turn out to be similar to the moving-window GLR scan statistics used by Lai (1995, 1998) and Lai and Shan (1999).

ACKNOWLEDGMENTS

Chan’s research is supported by the National University of Singapore under grant R-155-000-062-112. Lai’s research is supported by the National Science Foundation under grant DMS-0305749.
REFERENCES


Discussion on “Is Average Run Length to False Alarm Always an Informative Criterion?” by Yajun Mei

Moshe Pollak
Department of Statistics, Hebrew University of Jerusalem,
Jerusalem, Israel

Abstract: There are many reasons why the average run length has been the operating characteristic of choice to quantify the propensity of a control scheme to give rise to false alarms. Professor Mei’s paper dents this consensus. An alternative to Professor Mei’s suggested direction is considered.

Keywords: Average run length; Bayesian considerations; Change-point.

Subject Classifications: 62L10; 62L15; 60G40.

1. INTRODUCTION

Every once in a long while something comes up that forces one to stop in one’s tracks and reconsider concepts that theretofore seemed obvious and self-evident. Professor Mei’s article is a good example: the question that he poses and the evidence that he presents are a challenge, and demand revisiting and reassessment of a basic concept.

2. WHY ARL?

For many reasons, the average run length has been the sole practical candidate for quantification of the propensity of a detection policy to raise a false alarm. In a sense, this is natural: often we use expectations when required to characterize a random quantity.

Furthermore, when considering the simplest context of a changepoint problem—when observations are independent, i.i.d. prechange and i.i.d. postchange,
with known pre- and postchange distributions—the average run length to false alarm emerges in a natural context of the following Bayesian formulation. With a loss function having a price tag of one unit for a false alarm and a cost \( c > 0 \) per postchange observation, applying a stopping time \( N \) entails a

\[
\text{risk} = P(N < v) + cE(N - (v - 1) \mid N \geq v),
\]

(2.1)

where \( v \) is the serial number of the first postchange observation. Because one does not know when a change will take place (if it does), a uniform-like prior on \( v \) would be natural. A Geometric(\( p \)) prior with small \( p \) would be a reasonable proxy, and an approximation for the risk can be obtained by taking a limit (as \( p \to 0 \)) of

\[
\frac{1 - \text{risk}}{p} = \frac{P(N \geq v)}{p} [1 - cE(N - (v - 1) \mid N \geq v)].
\]

(2.2)

An easy calculation yields

\[
\lim_{p \to 0} \frac{P(N \geq v)}{p} = E_{v=\infty}N = \text{Average Run Length to False Alarm of } N.
\]

(2.3)

Because Bayesian analysis requires minimization of the risk (eq. (2.1)), which is equivalent to maximization of the expression in eq. (2.2), it follows from eq. (2.3) that solving the asymptotic Bayes problem involves minimizing the Average Delay to Detection \( E(N - (v - 1) \mid N \geq v) \) for some value of the ARL to false alarm. In addition, most classical detection procedures assume a known baseline, and the distribution of their run length is approximately exponential. Estimating the parameter of an exponential distribution is best made by averages, so characterizing the run length to false alarm by its average seems natural.

3. THE BLISS OF IGNORANCE?

Professor Mei’s examples point out that when the premises for the above break down, the average run length may not be the ideal functional to represent the propensity of a method to raise a false alarm. This is an important observation, and Professor Mei’s suggestions for addressing this problem point a way to other characterizations. His article opens whole new avenues of research.

Nonetheless, I have a hard time letting go of the ARL to false alarm (and Professor Mei’s proposals to address the problem indicate that he, too, finds it difficult): it is an easy concept to grasp, and makes sense in many applications, even when observations are dependent. I would like to point out that—at least in the first example in the article—the problem could be regarded as the bliss of ignorance rather than a weakness of the ARL. It’s nice and dandy to put on a Cauchy prior and then believe that the ARL to false alarm is infinite. A similar phenomenon appears in other contexts as well. For instance, consider the following hypothesis testing problem. Let the null hypothesis be that an observation \( X \) has an exponential(1) distribution and the alternative be that \( X \) has an exponential(\( \theta \)) distribution, with \( \theta < 1 \). An \( \alpha \)-level test would reject \( H_0 \) if \( X > -\log(\alpha) \). If \( \theta \) is known, then the power of the test is \( x^\theta \). Now, suppose that \( \theta \) is unknown, and
has a prior with density \((1 - \eta)/\theta^\eta\) for \(\theta \in (0, 1)\), where \(0 < \eta < 1\). The larger \(\eta\) the larger the power; actually, the power converges to 1 as \(\eta \to 1\). So, by “not knowing” the value of \(\theta\), one can blissfully pretend that the power of the test is very large.

The analog is clear: in the paper, it is the half-Cauchy distribution that brings one to believe that the ARL to false alarm is infinite. In fact, as Professor Mei himself notes, for every parameter value it is finite.

### 4. A POSSIBLE ALTERNATIVE

The latter comment notwithstanding, Professor Mei’s point is well taken. In addition to the proposals suggested in Professor Mei’s article, perhaps another way out is to translate the feeling that “a change can occur equally probably at any time” into a requirement that “at any time the response should be stochastically roughly equal in magnitude”; e.g., that the conditional probability of raising an alarm in a given interval (conditional on a false alarm not having been raised earlier) should not depend much on the interval’s location. The functional representing the propensity for raising a false alarm could, for example, be the maximal conditional probability (conditional in the sense as above) of raising a false alarm at a given observation (or any interval of observations).

### 5. ADDITIONAL COMMENTS

To carry things a step further, an additional iconoclastic program would be to dethrone the average run length (from change to detection) as the sole representative of detection delay. For example, consider surveillance of the population for an epidemic of influenza. Let \(\theta\) denote the average number of new cases of influenza per day, and suppose that under normal circumstances observations \(X_i\) taken on a daily basis have a \(\mathcal{N}(\theta, \sigma^2)\) distribution. When being on watch for an epidemic, one is on the lookout for an increase of the mean. The goal ought to be minimization of the expected number of cases from the onset of an epidemic to its detection rather than minimization of the expected number of days (observations). If the postchange value is unknown (as is the case in general), the two do not coincide.

In conclusion, Professor Mei’s paper is an important article. I expect that it will spawn much research.
Discussion on “Is Average Run Length to False Alarm Always an Informative Criterion?” by Yajun Mei

Wolfgang Schmid¹ and Sven Knoth²

¹Department of Statistics, European University Viadrina, Frankfurt (Oder), Germany
²Engineering, Advanced Mask Technology Center, Dresden, Germany

Abstract: The paper of Y. Mei provides an important contribution to the evaluation of control charts. In our discussion of his paper we focus on a new performance measure introduced in his paper, the expected false alarm rate. Several control procedures are compared based on this criterion.

Keywords: Average run length; Control chart; CUSUM; Statistical process control.

Subject Classifications: 62L10; 62L15; 60G40.

1. INTRODUCTION

In statistical process control (SPC) the control limit of a control chart is usually determined by fixing a desired value of the in-control average run length (ARL). The control limit is obtained as the solution of this equation. Based on these control limits, control charts are compared by using, e.g., the out-of-control ARL, the limit, or the maximum of the average delay. Recently, some papers question the performance of the out-of-control ARL (e.g., Frisén and Sonesson, 2006; Morais et al., 2008).

Mei considers the case that the prechange parameter is not uniquely identified. In SPC such a situation occurs for one-sided schemes like, e.g., one-sided variance charts. Then the control limit is calculated such that the smallest in-control ARL does not drop below the specified value. In the present paper Mei deals with another approach. The prechange parameter is considered to be a random variable with known a priori density. It is shown that the unconditional in-control ARL may be equal to infinity if the left tail of the a priori density is large. This is a very
remarkable result because it says that the control limits cannot be determined via the in-control ARL in such a case. In order to solve the problem, Mei proposes another measure for calculating the control limits, the expected false alarm rate (EFAR). The EFAR approximately reflects the quantile run length. A further performance measure is introduced in the paper, the asymptotic efficiency. This quantity seems to be of minor practical relevance because its calculation turns out to be quite difficult.

2. MORE SPECIFIC COMMENTS

In applications the a priori density will not be known. If the parameter space in the in-control state is bounded, most practitioners will make use of a uniform distribution. On the other hand, for reflecting uncertainty of prerun estimates, one would use (truncated) normal distributions. The author shows that problems arise if the a priori distribution has heavy tails and unbounded support. Where can we encounter such a situation in practice? All the examples in Mei (2006) have an intrinsic bounded support for the unknown parameter. Additionally, introducing a prior for the unknown in-control parameter calls instantly for a Bayesian type control chart that makes use of the priori distribution. In order to avoid the problems with the in-control ARL, in the present case another possibility would be to apply the median run length as proposed by Barnard (1959).

In order to get a better understanding of the EFAR, we made a small study. We are working under the same assumptions as in the paper, i.e., the variables are independent and normally distributed. However, the prechange parameter is assumed to follow a uniform distribution on \((-1, 0)\). Hence, EFAR simplifies to

\[
\text{EFAR}_j(T) = \int_{-1}^{0} \frac{1}{\mathbb{E}_0(T)} \, d\theta.
\]

Mei proposed to deploy \(q/\text{EFAR}_j(T)\) as approximation of the quantile \(\xi_j(T)\), that is,

\[
1 - q \approx \int_{-1}^{0} P_{\theta}\left( T > \frac{q}{\text{EFAR}_j(T)} \right) \, d\theta.
\]

For four different values of \(\mathbb{E}_0(L)\), 100, 1000, 10 000, 100 000, we check the validity of this approximation by utilizing a method given in Waldmann (1986). In this paper, the survival function of CUSUM stopping time \(T\) is calculated recursively. Additionally, we exploit the geometric (exponential) tail of the run length distribution. For calculating the above mixture integral, we applied Gauss–Legendre quadrature, which was also used for the survival function sequence and for determining the dominant eigenvalue of the CUSUM transition kernel. The latter resembles the parameter of the geometric tail. The number of quadrature nodes was set to 200. This number ensures sufficient accuracy even for the huge ARL values we get for \(\mathbb{E}_0(L)\), which exceeds \(10^{13}\) for \(\mathbb{E}_0(L) = 100\,000\). All these numerical algorithms are used to get the actual quantile probability for a given nominal level \(q\).

In Figure 1 the nominal level \(q\) versus the actual level \(q\) is plotted. Both levels are shown on a log scale. Additionally, the relative error (nominal – actual)/nominal is calculated, which is shown on its original scale. Figure 1 illustrates that both
Figure 1. Comparison of nominal and actual quantile levels for different minimal in-control ARLs $E_0(T)$—results are calculated numerically.

large $E_0(L)$ and small $q$ are needed to attain usable accuracy of the approximation. However, if $q$ gets further smaller, then the exponential approximation $P_0(T > u) \approx \exp(-u/E_0(T))$ is not longer valid and the accuracy deteriorates. The steps and the oscillating shapes stem from the discreteness of the run length distribution. To sum up, the approximation works only on a small subinterval of nominal
values $q$. For the setup we considered here, uniform prior on $(-1, 0)$ and four different false alarm levels, the approximation ability of $q/E_f(T)$ remains questionable.

REFERENCES

Discussion on “Is Average Run Length to False Alarm Always an Informative Criterion?” by Yajun Mei

Alexander G. Tartakovsky
Department of Mathematics, University of Southern California,
Los Angeles, California, USA

Abstract: Dr. Mei points out an important issue of whether a conventional average run length to false alarm is a proper measure of the false alarm rate in change-point detection problems. In this discussion, we address this issue in detail and show that for non-i.i.d. observations this is indeed a big question.

Keywords: Asymptotic optimality; Average detection delay; Average run length to false alarm; Change-point detection; False alarm rate.

Subject Classifications: 62L15; 60G40; 62F12; 62F15.

1. INTRODUCTION

It is a pleasure to welcome Dr. Yajun Mei’s paper that raises a series of important questions in change-point detection for non-i.i.d. data. Dr. Mei’s paper provides a brief overview of the state of research in the area and addresses an issue of false alarms, which is perhaps the most crucial bottleneck when detecting changes in nonhomogeneous and correlated processes.

Although for i.i.d. data models optimality of several conventional detection schemes (CUSUM, Shiryaev–Roberts, Shiryaev) is well understood both in minimax and Bayesian contexts, optimality properties of these conventional detection procedures for non-i.i.d. models have been established relatively recently by Lai (1998), Tartakovsky (1998), Moustakides (2004), Tartakovsky and Veeravalli (2004, 2005), Fuh (2003, 2004), and Baron and Tartakovsky (2006). The author raises a question as to whether the CUSUM and Shiryaev–Roberts tests are always
asymptotically optimal in the minimax sense under the constraint imposed on the average run length to false alarm (ARL2FA). He argues that Lai’s conditions, in particular the essential supremum condition on the conditional probability for the log-likelihood ratio (LLR), which is needed to obtain a lower bound for the supremum average delay to detection (AD2D), do not necessarily hold in general. More importantly, the author shows that even the ARL2FA may be infinite when considering mixture-based models. It turns out to be the case for the standard CUSUM detection test and, of course, for the Shiryaev–Roberts (S–R) test, too, in the specific problem considered in the paper, which, in essence, is the problem of detecting a change in the mean of the Gaussian sequence when the prechange mean is unknown and is modeled as random with a prior distribution that is taken to be half-Cauchy. The problem then is to detect a change in the mean from the mixture prechange to a postchange fixed value. Previously, the author has demonstrated the same issues for the finite mixture, in which case the ARL2FA of conventional CUSUM and S–R detection schemes is finite, but these schemes are not optimal even asymptotically in the minimax setting.

This kind of discussion is useful for understanding that in certain scenarios there may be a problem in applying conventional approaches. The author proposes an alternative approach to asymptotic optimality.

2. FALSE ALARM RATE: ARL VERSUS PFA

In this section, we argue that the ARL2FA constraint may not be feasible for non-i.i.d. models, but we also show that this issue requires a broader discussion than just the fact that it is infinite for the specific mixture-based example.

Suppose one is able to sequentially observe a sequence \( \{X_n\}_{n \geq 1} \) of i.i.d. random variables having density \( f \). At an unknown point in time \( \nu \geq 1 \) something happens and the observations change their statistical properties, so that thereafter their density is \( g \neq f \). The objective is to detect the change as soon as possible, subject to a FAR (false alarm rate) constraint. Let \( \mathcal{L}_n = \log[g(X_n)/f(X_n)] \) be the corresponding log-likelihood ratio (LLR). The two popular change-point detection procedures that have certain optimality properties are the CUSUM and S–R procedures, which are given by the stopping times:

\[
N_{CS}^A = \min\{n \geq 1 : W_n > \log(1 + A)\} \quad \text{and} \quad N_{SR}^A = \min\{n \geq 1 : R_n > A\},
\]

where the CUSUM statistic \( W_n = \min\{0, W_{n-1} + \mathcal{L}_n\} \) and the S–R statistic \( R_n = (1 + R_{n-1})e^{\mathcal{L}_n} \) (usually with null initial conditions \( W_0 = R_0 = 0 \)).

There are two kinds of FAR performance metrics: global and local. A conventional ARL2FA is an example of the global metric. On the other hand, Tartakovsky (2005) discussed usefulness of such local measures as the local conditional probability of false alarm \( \text{CPFA}_{m}^k(T) = \mathbb{P}_{\mathcal{M}}(k \leq T < k + m | T \geq k) \), where \( T \) stands for a generic stopping time and where \( m \), the size of the time window, may be fixed or go to infinity when the constraint \( \alpha \) imposed on \( \text{CPFA}_{m}^k(T) \) becomes small (or threshold \( A \to \infty \)).

It turns out that for i.i.d. data, the \( \mathbb{P}_{\mathcal{M}} \)-distributions (suitably standardized) of stopping times \( N_{CS}^{\infty} \) and \( N_{SR}^{\infty} \) are asymptotically (as \( A \to \infty \)) exponential, in which case ARL2FA, the global FAR metric, is obviously appropriate. See, e.g., Pollak...
and Tartakovsky (2008a). For testing the accuracy of the exponential (geometric) approximation for the stopping times \( N_A^{CS} \) and \( N_A^{SR} \) under the no-change hypothesis, we used quantile-quantile plots (QQ-plots – theoretical versus experimental quantiles). The obtained results for an exponential example (change in the mean from 1 to \( 1 + \theta \)) are presented in Figure 1(a) and (b). Linearity is a strong

Figure 1. QQ-plots for the stopping times of the CUSUM and S–R procedures: \( \theta = 1 \), \( \text{ARL}_{2\text{FA}} = 100 \): (a) CUSUM; (b) Shiryaev–Roberts.
indication that the stopping times are indeed exponentially distributed, even though $\text{ARL}_{2FA}(N_A) \approx 100$, which is considered small (i.e., high FAR). Figure 2(a) and (b) illustrate the behavior of the local conditional probabilities of false alarm $\text{CPFA}_m(N_A) = \mathbb{P}(N_A < k + m \mid N_A \geq k)$ versus $k$ for the fixed stretch size $m$. The dashed curves are the values of CPFA obtained experimentally for different values of $m$. The solid lines are the values calculated by the “geometric” approximation $\text{CPFA}_m(N_A) = 1 - (1 - p_A)^m$ with $p_A = 1/\text{ARL}_{2FA}(N_A)$, which

![Figure 2](image-url)

**Figure 2.** Local PFA for S–R and CUSUM procedures: $\theta = 1$, $\text{ARL}_{2FA} = 100$: (a) CUSUM; (b) Shiryaev–Roberts.
follows from Pollak and Tartakovsky (2008a) and Tartakovsky et al. (2008). It is seen that even for small ARL2FA, geometric approximation works fairly well. This allows one to justify the use of the ARL2FA as a measure of false alarms and to evaluate supremum PFA, \( PFA_m(T) = \sup_{k} P_{\infty}(T < k + m \mid T \geq k) \), which is a proper measure of the FAR, as discussed in Tartakovsky (2005).

Whereas for i.i.d. models the distribution of stopping times of conventional detection procedures is asymptotically exponential if there never is a change, which justifies using the ARL2FA as a measure of the FAR, for non-i.i.d. models this is, in essence, an open problem. In general, we cannot even claim that large values of ALR2FA guarantee small values of the supremum local false alarm probability \( PFA_m(T) \), which is a desirable, if not a necessary, property. Therefore, even if the ARL2FA is finite, it may not be an appropriate measure of the false alarm rate!

On the other hand, if one is interested in the class of procedures \( \Lambda_x = \{ T : PFA_m(T) \leq x \} \) for which the local false alarm probability \( PFA_m(T) \) does not exceed a predefined number \( x \), the CUSUM and S–R detection procedures are asymptotically uniformly optimal if \( A = A_x = C_m/x \) (for every \( v = k \geq 1 \)):

\[
\inf_{T \in \Lambda_x} E_x(T - k \mid T \geq k) \sim E_x(N^{SR}_A - k \mid N^{SR}_A \geq k) \sim E_x(N^{CS}_A - k \mid N^{CS}_A \geq k) \sim \frac{\log(1/x)}{\mathcal{J}}, \quad x \to 0,
\]

where \( C_m \) is a constant depending on \( m \) (which is different for CUSUM and S–R tests) and \( \mathcal{J} \) is the Kullback–Leibler information number (cf. Tartakovsky, 2005). This result also holds for general non-i.i.d. models with \( \mathcal{J} \) replaced by \( Q \) whenever \( n^{-1} \mathcal{J}_n \to Q \), almost surely under \( D_x \), with a certain rate of convergence, where \( \mathcal{J}_n = \log(dP_x(X^n)/dP_{\infty}(X^n)) \) is the LLR for the data \( X^n = (X_1, \ldots, X_n) \).

Note that Lai (1998) considers the problem of minimizing \( E_x(T - k) \) (for every \( k \geq 1 \)) under constraints imposed on the supremum unconditional probability of false alarms \( \sup_{k} P_{\infty}(k \leq T < k + m) \leq x \) and shows that certain detection procedures are uniformly asymptotically optimal as \( x \to 0 \) for quite general non-i.i.d. models when \( m = m_x \to \infty \) with a certain rate.

The fact that the distribution \( P_{\infty}(T < x) \) may be far from being geometric for non-i.i.d. models (even for sufficiently low FAR) is illustrated by Figure 3, where we show QQ-plots for the CUSUM stopping time for the autoregressive data. If this is the case, the ARL2FA is perhaps not a good measure of the FAR. This means that even if Lai’s essential supremum conditions hold, which is true for a variety of non-i.i.d. examples, and, thus, the CUSUM and S–R procedures are asymptotically optimal for large ARL2FA, this optimality property may not be too valuable. Therefore, local PFA measures for the FAR proposed by Lai (1998) and Tartakovsky (2005) are important alternatives to ARL2FA.

3. BAYESIAN PROBLEM SETTING

Consider the Bayesian setting with the prior distribution \( \pi_k = \mathbb{P}(v = k) \) and the constraint on the average PFA

\[
PFA(T) := \sum_{k=1}^{\infty} \pi_k \mathbb{P}_k(T < k) = \sum_{k=1}^{\infty} \pi_k \mathbb{P}_{\infty}(T < k) \leq x, \quad 0 < x < 1.
\]
It has been recently shown by Tartakovsky and Veeravalli (2005) (see Baron and Tartakovsky, 2006, for continuous time) that Shiryaev’s Bayesian detection procedure (cf. Shiryaev, 1963) $T_k = \min\{n : G_n \geq A_n\}$ with $A_n = (1 - \alpha)/\pi$ is asymptotically optimal (as $\alpha \to 0$) in terms of minimizing AD2D(T) = $E(T - \nu | T \geq \nu)$ for an arbitrary prior distribution $\pi_k$ and very general non-i.i.d. models. Here $G_n$ is the average likelihood ratio defined in Tartakovsky and Veeravalli (2005). In fact, this result holds whenever the normalized LLR $n^{-1} \sum_{k=1}^{n} \ell_k$ converges almost surely to a positive finite number $Q$ with a certain rate of convergence (e.g., complete convergence).

We now show that the FAR problem and related asymptotic optimality issues do not exist for Bayesian problems.

Although all argument holds for a continuous mixture (e.g., for the half-Cauchy prior for the prechange parameter), for the sake of simplicity, consider a mixture of two distributions, assuming that there are three densities $g_1(X_n), g_2(X_n),$ and $f_1(X_n)$. The problem is to detect the change from the mixture density

$$f_0(X^n) = \beta \prod_{i=1}^{n} g_1(X_i) + (1 - \beta) \prod_{i=1}^{n} g_2(X_i)$$

to the density $f_1$, where $0 < \beta < 1$ is a mixing probability. Therefore, the observations are dependent with joint density $f_0(X^n)$ before the change occurs and i.i.d. with density $f_1$ after the change occurs.

Consider the geometric prior $\pi_k = \rho(1 - \rho)^{k-1}, 0 < \rho < 1$. (All results can be generalized for an arbitrary prior.)
Denote \( R_j(n) = \log [f_j(X_n) / g_j(X_n)] \); \( \mathcal{J}_j = \mathbb{E}_j R_j(1), \ j = 1, 2; \ \xi_i = \prod_{m=1}^i [g_1(X_m) / g_2(X_m)]; \ \nu = \beta/(1 - \beta) \). It may be shown that the LLR

\[
\zeta_n^k := \frac{1}{n} \sum_{i=k}^n \log \frac{f_1(X_i)}{f_0(X_i | X_{i-1})} = \sum_{i=k}^n R_2(i) + \log \frac{1 + v\xi_{k-1}}{1 + v\xi_n}. \tag{3.1}
\]

Assume that \( \mathcal{J}_1 > \mathcal{J}_2 \), in which case expectation \( \mathbb{E}_k \log [g_1(X_m) / g_2(X_m)] \leq 0 \) for \( k < m \) and, hence,

\[
\xi_n = \xi_{k-1} \prod_{i=k}^n [g_1(X_i) / g_2(X_i)] \xrightarrow{\mathbb{P}_k \text{-a.s.}} 0 \quad \text{as } n \to \infty,
\]

which according to Tartakovsky and Veeravalli (2005) implies the lower bound

\[
\inf_{\{T| \mathcal{H}_1(T) \leq \alpha\}} \text{AD}2\text{D}(T) \geq \left( \frac{|\log x|}{\mathcal{J}_2 + |\log(1 - \rho)|} \right) (1 + o(1)) \quad \text{as } x \to 0.
\]

Next, using (3.1), it can be shown that, for any \( k \), the statistic \( \log G_n \) can be written in the following form:

\[
\log G_n = \sum_{i=k}^n R_2(i) - (n - k + 1) \log(1 - \rho) + Y^k_n,
\]

where \( \{Y^k_n\}_{n \geq k} \) are slowly changing. By the nonlinear renewal theorem (cf. Siegmund, 1985; Woodroofe, 1982),

\[
\text{AD}2\text{D}(T_{\alpha}) = \left( \frac{|\log x|}{\mathcal{J}_2 + |\log(1 - \rho)|} \right) (1 + o(1)) \quad \text{as } x \to 0,
\]

and it follows that the Shiryaev detection procedure \( T_{\alpha} \) is asymptotically optimal.

On the other hand, the minimax property of the CUSUM and S–R tests does not hold in the example considered under the ARL2FA constraint.

### 4. ALTERNATIVE APPROACHES TO MEI’S EXAMPLE

Consider a couple of alternative approaches to the problem proposed by Dr. Mei without introducing any mixtures.

Recall that the problem is to detect the change in the mean of the i.i.d. Gaussian sequence, \( \theta \to \lambda \), where \( \theta < 0 < \lambda \). Assume, for simplicity, that \( \lambda \) is known (but \( \theta \) is not known). Write \( \text{AD}2\text{D}_\theta \) for either Pollak’s supremum or Lorden’s essential supremum average detection delay when \( \theta \) is the true value of the prechange parameter.

The simplest but not the most efficient way is to use the CUSUM test tuned to \( \theta = 0 \), which is the worst point (AD2D\( _\theta \) is maximized for \( \theta = 0 \)). This procedure
delivers the optimum for \( \sup_{\theta \leq \theta_0} \text{AD2D}_\theta \) in the class \( \{ T : \inf_{\theta \leq \theta_0} \text{ARL2FA}_\theta \geq \gamma \} \), but for large negative \( \theta \) the performance is poor compared to the optimal CUSUM tuned to the true value of \( \theta \).

In order to improve performance, choose \( \theta_M < \cdots < \theta_2 < \theta_1 = 0 \) and let \( \widetilde{N}^\text{CS} = \min \{ n : W_n > b_i \} \) be the stopping time of the CUSUM test tuned to \( \theta = \theta_i \). The point \( \theta_M \) is chosen in such a way that the \( \text{AD2D}_{\theta_M} \) is small enough and further optimization is not important.

Consider the multichart CUSUM \( N = \min_{1 \leq i \leq M} \widetilde{N}^\text{CS} \). It may be shown that this procedure is asymptotically optimal at points \( \theta = \theta_i \), \( i = 1, \ldots, M \) if thresholds \( b_i \) are selected as \( b_i = \log[\gamma M(\lambda - \theta_i)^2 v_i^2/2] \), where \( v_i \) are computable constants, \( 0 < v_i < 1 \):

\[
\inf_{\{ T : \inf_{\theta \leq \theta_0} \text{ARL2FA}_\theta \geq \gamma \}} \text{AD2D}_{\theta_i}(T) \sim \text{AD2D}_{\theta_i}(N) \sim \frac{2 \log \gamma}{(\lambda - \theta_i)^2}, \quad \gamma \to \infty.
\]

On the other hand, the multichart CUSUM will give a reasonably good approximation to the best possible performance at the intermediate points \( \theta \neq \theta_i \), and, therefore, may be considered as a reasonable candidate for practical applications. The same asymptotic performance can be obtained by using a multichart S–R detection test.

Yet another possible (and asymptotically efficient) solution can be constructed based on the maximal invariant sequence \( Y_n = X_n - X_1, \quad n \geq 2 \). Specifically, we conjecture that building likelihood ratios for \( Y_n \) and applying the corresponding invariant S–R test \( N_A \) will allow one to obtain an asymptotically optimal solution (as \( \alpha \to 0 \)) with respect to the average detection delay \( E_{\text{AMS}}(T - k \mid T \geq k) \) uniformly for every \( k \geq 2 \) in the class of invariant detection procedures \( \Delta_A = \{ T : \sup_k P_{\infty} < k + m \mid T \geq k \} \leq \alpha \) that confines the supremum local PFA. In fact, because the invariant S–R statistic \( R_n \) is a non-negative submartingale with mean \( E_{\infty} R_n = n \), it follows that \( P_{\infty}(N_A < k + m \mid N_A \geq k) \leq m/A \). Choose \( m_\alpha = O((\log \alpha)) \) and \( A = A_\alpha \) as a solution of the equation \( m_\alpha/A_\alpha = \alpha \). Generalizing an argument in Tartakovsky (2005) may lead to the desired asymptotic optimality result. This problem will be addressed elsewhere. Note also that the global ARL2FA metric may not be a good choice for the FAR, because the sequence \( \{ Y_n \}_{n \geq 2} \) is not i.i.d.

5. DETECTION OF A CHANGE OCCURRING AT A FAR HORIZON

In many surveillance applications (such as detecting targets with various sensors and intrusions in computer networks), one is interested in rapid detection of changes that occur not from the very beginning but further off. In this case, a real change-point \( v \) is a lot greater than \( \gamma \), the mean time between false alarms, and the most reasonable strategy is to apply a procedure (CUSUM, S–R, etc.) repeatedly after many false detections have been experienced. In the i.i.d. case, this problem setting is well understood. In particular, it follows from Pollak and Tartakovsky (2008b) (see also Shiryaev, 1963, for a Brownian motion) that the S–R procedure is optimal for every \( \gamma > 1 \) (for known pre- and postchange distributions).

It is not clear, however, whether this kind of approach can be used for non-i.i.d. observations. This is related not only to the FAR issue, but to the nature of the problem itself.
6. FINAL COMMENTS

The issues raised by Dr. Mei are very important and lead to a series of other questions, the most general one being “Whether it is worth trying to develop a general minimax change-point detection theory that goes far beyond the conventional i.i.d. assumption under constraints imposed on global FAR metrics, in particular on ARL2FA?” Many problem settings that are currently being used fit well into i.i.d. data models, although, perhaps, not so well into general non-i.i.d. models. Local FAR metrics (one of which is the local probability of false alarm $PFA_m(T) = \sup_{k} \mathbb{P} \{T < k + m | T \geq k\}$) seem to be more appropriate when dealing with non-i.i.d. observations. Also, in extreme situations a “case-by-case” approach may be the most appropriate. On the other hand, this issue does not exist for a Bayesian problem setting.

Finally, Dr. Mei’s attempt to develop an alternative asymptotic theory that would allow to cover general stochastic models, although very important, can be considered only as the very first step. Any theory is useful if, and only if, a corresponding performance fits well into prelimit performance estimates that can be computed using numerical methods or Monte Carlo experiments. In this sense a problem is still open. Could we suggest simple and exhaustive performance metrics that can be easily evaluated and used by practitioners in a variety of applications for non-i.i.d. data models?

ACKNOWLEDGMENTS

The research was supported in part by the U.S. Office of Naval Research grant N00014-06-1-0110 and by the U.S. Army Research Office MURI grant W911NF-06-1-0094 at the University of Southern California.

REFERENCES


Discussion on “Is Average Run Length to False Alarm Always an Informative Criterion?” by Yajun Mei

Benjamin Yakir
Department of Statistics, Hebrew University, Jerusalem, Israel

Abstract: In his interesting paper Mei criticized the use of the expected run length as the mean for controlling the rate of false detection and proposed alternative measures. In this paper we join Mei’s attack on the traditional constraint by claiming that the rate of detection is a local phenomena, and hence should be balanced against local constraints on false detection. We propose local probabilistic constraints on the rate of false detection and demonstrate their usefulness in the detection of a shift in a normal mean with an unknown baseline.

Keywords: Asymptotic optimality; Change-point detection; False detection rate; Invariant statistics.

Subject Classifications: 62L15; 60G40.

1. INTRODUCTION

The title of Mei’s paper is composed of a question regarding the general appropriateness of the average run length (ARL) as a measure of false detection. The author’s short answer to the question is “No,” with which we agree. Yet, here we would like to propose a different rational for that answer and an alternative remedy for fixing the difficulties associated with the traditional approach of constraining false alarms. In the next section we introduce and motivate our alternative proposal. In the section that follows we demonstrate its applicability in the context of monitoring with invariant statistics for a shift in a normal mean with unknown baseline.
2. MONITORING THE FALSE DETECTION RATE

Using the ARL to false alarm as the mean for monitoring false detections is not the only choice that became the default in the context of optimal online change-point detection. Another such choice is the consideration of nonlocal alternatives for modeling the postchange distribution. The implication of that other default in simple settings is that the rate of detection is on the log-scale in comparison to the rate of false alarms. Thence, when considering the monitored sequence on the scale of the expected time to false alarm, the delay in detection is a function of the local behavior of the process in the vicinity of the change-point and not a function of its global behavior. The translation of a constraint given in terms of ARL to false alarm, which is a global measure, to lower bounds on the delay in detection is not direct and depends on specific properties of the monitored sequence. This adds another layer of complexity to the proofs when attempting to extend optimality statement to more complex settings.

A more direct relation between the rate of false alarms and the delay in detection may be obtained if the constraint on false detection is also given in terms of the local behavior of the monitored sequence. In a recent paper (Siegmund and Yakir, 2008), an attempt was made to generalize the asymptotic theory of optimal change-point detection to the setting where the postchange parameter is unknown. In that paper it was proposed to allow only stopping rules \( N \) that possess a probabilistic constraint on the local rate of false alarm. Specifically, denote by \( A \) a large index. In order for \( N \) (a sequence of stopping times indexed by \( A \)) to be permissible, there must be some \( l > 0 \) and \( m \ll A \), both may depend on \( A \), for which \( \mathbb{P}_a(N \leq l + m | N > l) \leq (m/A)(1 + o(1)) \), as \( A \to \infty \). (The distribution \( \mathbb{P}_a(\cdot) \) corresponds to no change.) A likelihood-ratio identity may be used to relate the null probability of stopping in the interval \( (l, l + m) \) to the postchange distribution of the delay in detection in the same interval. The relation may be used to produce a lower bound on the expected delay. A permissible Shiryayev–Roberts stopping rule asymptotically obtains the lower bound and is thus optimal.

The traditional formulation constrains the rate of false alarms using the condition \( \mathbb{E}_a(N) \geq A \). An argument by contradiction shows that the constraint we propose is weaker. If a monitoring procedure is optimal among all stopping rules that satisfy the weaker constraint and if the same stopping rule also satisfies the traditional constraint, which is the case for the Shiryayev–Roberts stopping rule, then it is optimal in the traditional formulation as well.

The essential feature in the examples analyzed by Mei is the fact that the baseline level of the monitored parameter is unknown. Clearly, because the prechange observations may be used to estimate that baseline, the consequences are more severe the sooner the change takes place. Therefore, a fundamental question is the rate by which the performance of a monitoring procedure deteriorates as a function of the decrease in the number of prechange observations. This deterioration depends on local behavior at the initiation of monitoring. The ARL to false alarm, which is affected mainly by the distribution of larger values of the stopping rule, will not be effective in controlling the properties of the smaller values of that rule. Generally, stopping rules are functions of likelihood ratios. These likelihood ratios are inherently nonstationary in the discussed setting. Consequently, the ability to extract local properties of the monitoring procedure from its global properties is further reduced in comparison to the settings that involve stationary likelihood
ratios. This is demonstrated beautifully via the pathological examples that the author provides that clearly show the inappropriateness of a global constraint such as the mean.

As an alternative we propose to consider probabilistic constraints on the local rate of false detection at the origin and relate the constraints to the speed of detection directly. We think that this is a useful way to extract insights on the fundamental question. In order to demonstrate that point, we provide in the next section a telegraphic assessment of the problem of online monitoring for a shift in a normal mean with an unknown initial baseline. We have selected this example, rather than the main example of the discussed paper, because we had investigated the model of invariant normal shift in the past and are more familiar with its details. In the near future we hope to be able to submit a more detailed investigation of the general issue of unknown baseline, which should include the examples presented in Mei’s paper.

3. DETECTING A CHANGE IN NORMAL MEAN WITH AN UNKNOWN BASELINE

3.1. The Model and Its Likelihood Ratios

Assume one monitors for a shift in mean, when the initial mean is unknown, a sequence \( Y_0, Y_1, \ldots \) of independent Gaussian observations with unit variance. Invariance considerations propose using likelihoods of the maximal invariant statistics: \( X_i = Y_i - Y_0, \ i \geq 1 \). For the determination of the distribution of the invariant statistics we may assume, without loss of generality, that the baseline expectation of the \( Y \)'s equals zero and use the representation of the likelihood ratio for a shift of size \( \lambda \) at time \( k \):

\[
e^{\ell_k(n)} = \frac{f_k(X_1, \ldots, X_n)}{f_\infty(X_1, \ldots, X_n)} = e^{\sum_{i=k}^{n-1} X_i - \frac{1}{2} \lambda^2 (n-k+1)} \mathbb{E}_m \left[ e^{\lambda (n-k+1) Y_0} \mid \mathcal{F}_n \right],
\]

where we set \( \mathcal{F}_n = \sigma(X_1, \ldots, X_n) \). Denoting \( \overline{Y}_k = (1/k) \sum_{i=0}^{k-1} Y_i \), one may obtain for a given shift:

\[
\ell_k(n) = \lambda \sum_{i=k}^{n} \left\{ X_i + \mathbb{E}(Y_0 \mid \mathcal{F}_n) - \frac{\lambda}{2} [1 - (n - k + 1)\text{Var}(Y_0 \mid \mathcal{F}_n)] \right\}
\]

\[
= \frac{\lambda k}{n+1} \sum_{i=k}^{n} \left\{ Y_i - \overline{Y}_k - \frac{1}{2} \lambda \right\}.
\]

For a mixture-type log-likelihood ratio, one may integrate the likelihood \( \exp\{\ell_k(n)\} \), as a function of \( \lambda \), against a mixture distribution and take the logarithm of the outcome. Specifically, if one uses a zero mean normal density with variance \( \varsigma^2 \) and if one sets \( \overline{Y}_n = \frac{1}{n} \sum_{i=0}^{n-1} Y_i \), then one gets:

\[
\ell_k(n) = \left[ \frac{k(n-k+1)}{n+1} \right] \left\{ \overline{Y}_n - \overline{Y}_k \right\}^2 - \frac{1}{2} \log \left( 1 + \frac{k(n-k+1)^2}{n+1} \right),
\]

(3.2)
Discussion 409

The Shiryayev–Roberts detection rule takes the form: $N_\delta = \inf\{n : \sum_{k=1}^n \exp\{\ell_k(n)\} \geq A\}$. Because the cumulative sum of likelihoods is a submartingale, one may obtain the probabilistic constraint on the rate of false alarms: $\mathbb{P}_\infty(N_\delta \leq m) \leq m/A$. Let $N$ be any other stopping time with respect to the sequence of invariant statistics. We will require that $N$ obeys the same probabilistic constraint and investigate the resulting implication on the delay in detection when a change takes place at time $k$, for $k$ relatively small. As it turns out, when the shift equals $\lambda$, it is convenient to relate $k$ to the index $A$ by taking $k = t \log A(2/\hat{\lambda}^2)$. In the next subsections we show that if $t < 1$ then the detection delay is bounded from below by $A$ to some power and if $t > 1$ then a logarithmic rate of detection is feasible.

3.2. Subcritical Prechange Sample Sizes

Assume $t < 1$. Select $\epsilon > 0$ and take $x = (1+\epsilon)t \log A$. From the likelihood-ratio identity:

$$\frac{m}{A} \geq \mathbb{P}_\infty(N \leq m) = \mathbb{E}^\mathbb{P}_\infty[\epsilon_k(N) ; N \leq m] \geq e^{-\epsilon t} \mathbb{P}_\infty^k(\ell_k(N) \leq x, N \leq m),$$

for any permissible $N$. Also, because $\mathbb{P}_\infty^k(\ell_k(N) \leq x, N \leq m) \geq \mathbb{P}_\infty^k(N \leq m) - \mathbb{P}_\infty^k(\max_{n \leq m} \ell_k(n) > x)$,

$$\mathbb{P}_\infty^k(N \leq m) \leq \frac{m}{A^{1-(1+\epsilon)}} + \mathbb{P}_\infty^k(\max_{n \leq m} \ell_k(n) > (1+\epsilon)t \log A).$$

(3.3)

The probability on the right-hand side of (3.3) is $\mathbb{P}_\infty^k(\tau_k \leq m)$, where

$$\tau_k = \inf\{n \geq k : \epsilon_k(n) \geq x\}$$

$$= \inf\left\{n \geq k : \lambda \sum_{i=k}^n (Y_i - T_k) - (1+\epsilon/2)A \geq (1+\epsilon)t \log A \right\}.$$ 

(3.4)

This probability may be assessed by conditioning on the value of $\bar{T}_k$ and bounding the conditional probability by the probability that a Brownian motion with drift $-A[\bar{T}_k + A\delta/2]$ ever reaches the upper boundary $\lambda x$. After integrating the resulting upper bound with respect to the distribution of $\bar{T}_k$, we get the inequality:

$$\mathbb{P}_\infty^k\left(\max_{n \leq m} \epsilon_k(n) > (1+\epsilon)t \log A\right) \leq \frac{\phi[(\epsilon t/2 \log A)^{1/2}]}{(1+\epsilon)(t \log A)^{1/2} - \lambda \epsilon / \sqrt{2}} < A^{-2\delta}.$$ 

(3.5)

From the combination of (3.3) and (3.5) one may conclude that for any $\delta > 0$, and with probability that converges to one, the accumulation of at least $A^{1-t-\delta}$ observations is needed before declaring a change.

3.3. Logarithmic Regimes

Next we consider cases where the accumulated sample size prior to the change is above the critical level. We apply the likelihood identity again, this time in combination with Jensen’s inequality, to obtain:

$$\frac{m}{A} \geq \mathbb{P}_\infty(N \leq k - 1 + m \mid N \geq k) = \mathbb{E}^\mathbb{P}_\infty^k[\epsilon_k(N) ; N - k + 1 \leq m \mid N \geq k]$$

$$\geq \mathbb{P}_\infty^k(N - k + 1 \leq m \mid N \geq k) e^{-\mathbb{E}^\mathbb{P}_\infty^k[\epsilon_k(N) ; N - k + 1 \leq m \mid N \\geq k]} \mathbb{P}_\infty^k(N - k + 1 \leq m \mid N \geq k).$$
Denote $p_k = \mathbb{P}_k^*(N - k + 1 \leq m \mid N \geq k)$, take logarithm, and rearrange the terms to get:

$$\mathbb{E}[\ell_k(N); N - k + 1 \leq m \mid N \geq k] \geq p_k \log A - p_k \log m + p_k \log p_k.$$  \hspace{1cm} (3.6)

Using (3.1) and straightforward algebra, we get a representation of $\ell_k(N)$ as a sum of three elements:

$$\ell_k(N) = \left[ \frac{2/\lambda^2}{N - k + 1} + \frac{1}{t \log A} \right]^{-1} + \lambda \left[ \frac{k}{N + 1} \right] \sum_{i=k}^{N} (Y_i - \lambda) - \lambda \left[ \frac{k(N - k + 1)}{N + 1} \right] y_k.$$

Taking expectation of the first element and applying Jensen’s inequality produces:

$$\mathbb{E}[\ell_k(N)] \leq \left( \frac{2/\lambda^2}{\mathbb{E}[N - k + 1 \mid N \geq k]} + \frac{1}{t \log A} \right)^{-1}.$$  \hspace{1cm} (3.7)

The expectation of the other two terms can be shown to be $o(\log A)$, provided that the expected delay in detection is $o((\log A)^2)$. Combining that with (3.6) and (3.8), for a large enough $m$, leads to:

$$\left( \frac{(2/\lambda^2) \log A}{\mathbb{E}[N - k + 1 \mid N \geq k]} + \frac{1}{t} \right)^{-1} \geq 1 + o(1),$$

and, as a result,

$$\mathbb{E}[N - k + 1 \mid N \geq k] \geq \frac{t}{1 - 1/\lambda^2} \log A (1 + o(1)).$$  \hspace{1cm} (3.8)

It can be shown that for $t > 4$ the lower bound is obtained in expectation by the Shiryayev–Roberts stopping rule $N_A$. For $1 < t \leq 4$ the lower bound is obtained in probability, but probably not in expectation. Unfortunately, thus far we were not able to produce a rigorous proof of the very last statement.

REFERENCES

Discussion on “Is Average Run Length to False Alarm Always an Informative Criterion?” by Yajun Mei

Shelemyahu Zacks
Department of Mathematical Sciences, Binghamton University,
Binghamton, New York, USA

Abstract: In addition to the ARL$_\infty$, the discussion emphasizes the importance of determining PFA($\tau$) and CED($\tau$). An example of such computations is given. When the parameters before and after the change-point are unknown, it is recommended to employ the full Bayesian framework.

Keywords: Average run length; Conditional expected delay; CUSUM; Probability of false alarm; Statistical process control.

Subject Classifications: 62L10; 62L15; 60G40.

1. GENERAL

The average run length to false alarm, ARL$_\infty$, is an important characteristic of stopping times, for statistical process control, when the stopping times are finite with probability 1 and the change point, $\tau$, is at infinity. Notice that in this case the probability of false alarm is 1. However, ARL$_\infty$ by itself is not sufficient.

When the change point $\tau$ is finite, it is important to characterize a stopping time $T$ by the conditional expected delay $\text{CED}(\tau) = E_\tau \{ (T - \tau)^+ | T \geq \tau \}$. This is a function of the location $\tau$ of the change-point. Whereas ARL$_\infty$ is analogous to type I error in testing hypotheses, CED($\tau$) is analogous to the power function in testing hypotheses. Moreover, when $\tau < \infty$, it is important to characterize $T$ by its probability of false alarm, $\text{PFA}(\tau) = P_\infty \{ T < \tau \}$. The computation of PFA($\tau$) and CED($\tau$) requires the determination of the distribution of $T$, when $\tau = \infty$ and when $\tau < \infty$. Generally it is a difficult problem. In the case of the Shewhart $3 - \sigma$ control chart for the mean, the distribution of $T$ under $\tau$ is easily determined.
However, in the case of a CUSUM stopping time, the problem is not trivial. Zacks (1981) derived this distribution for one-sided CUSUM procedure, when the observed random variables are non-negative integer valued. In the paper of Zacks (2004), we have a derivation of the distribution of the stopping time for a one-sided CUSUM procedure based on a Poisson process. In the textbook *Modern Industrial Statistics* of Kenett and Zacks (1998), the PFA($\tau$) and the CED($\tau$) are computed for various control schemes (CUSUM, Shiryaev–Roberts) by simulations. In the review article of Zacks (1990), the distribution of the run length with a finite $\tau$ is illustrated for the case of one-sided CUSUM on Poisson random variables. We illustrate these computations here on an artificial example with a nonoptimal stopping rule, for simplicity.

Let $\{N(t), t \geq 0\}$ be a homogeneous Poisson process with intensity

\[ \mu_t = \begin{cases} 
\lambda, & \text{if } t \leq \tau \\
\phi, & \text{if } t > \tau. 
\end{cases} \]

The stopping time for detecting the change-point $\tau$ is

\[ T = \inf \{t > 0 : N(t) = 30\}. \]

The distribution of $T$, when $\tau = \infty$, is Erlang(30, $\lambda$). When $\lambda = 0.1$, then ARL$_{\infty}$ = 300. The distribution of $T$ for a finite $\tau$ is

\[ F_T(t; \tau, \lambda, \phi) = \begin{cases} 
1 - P(29; \lambda t), & \text{if } t \leq \tau \\
1 - \sum_{j=0}^{29} p(j; \lambda, \tau) P(29 - j; \phi(t - \tau)), & \text{if } t > \tau.
\end{cases} \]

Here $p(j; \mu)$ is the probability density function (p.d.f.) of the Poisson distribution with mean $\mu$, and $P(j; \mu)$ the corresponding cumulative distribution function (c.d.f.) Accordingly, the probability of false alarm is

\[ \text{PFA}(\tau) = 1 - P(29; \lambda \tau), \]

and the conditional expected delay is

\[ \text{CED}(\tau) = \frac{1}{P(29; \lambda \tau)} \sum_{j=0}^{29} p(j; \lambda \tau) \cdot \frac{30 - j}{\phi}. \]

In Table 1 we present some values of PFA and of CED for $\lambda = 0.1$ and $\phi = 0.2, 0.5, 1, 2, 3$, for $\tau = 10, 50, 100, 200, 220, 250$.

Such a table gives a good picture of the operating characteristics of the stopping variable $T$.

## 2. THE CASE OF UNKNOWN PRECHANGE PARAMETER

The author considers the case where before the change $X \sim N(\theta, 1)$ and after the change $X \sim N(\lambda, 1)$ where $\theta < 0 < \lambda$, $\theta$ unknown and $\lambda$ known. He proceeds then
with a semi-Bayesian approach, by using a mixture measure $P_j = \int_{-\infty}^{0} P_0 \pi(\theta) d\theta$ for the joint density of the r.v.’s $X_1, X_2, \ldots, X_{\tau-1}$. Furthermore, the author continues using the CUSUM procedure. Notice that the CUSUM procedure was designed for independent observations. I believe that cases of unknown prechange or postchange parameters should be dealt within a Bayesian framework. In the context of the change-point problem, there are many papers in the literature that treat the unknown parameters problem in a Bayesian framework. See for example, Zacks and Barzily (1981), Kenett and Zacks (1994), Brown and Zacks (2006), and more. One should mention that important characteristics of the posterior processes, such as recursive determination of posterior probabilities, do not exist when the densities before and after the change are not completely known. In many cases the sequentially sufficient statistics are the trivial ones and one has to retain all the observations. This might be inconvenient. See Zacks (1990) for further details.

REFERENCES


<table>
<thead>
<tr>
<th>$\tau$</th>
<th>10</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>220</th>
<th>250</th>
</tr>
</thead>
<tbody>
<tr>
<td>PFA</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0218</td>
<td>0.0602</td>
<td>0.1821</td>
</tr>
<tr>
<td>$\phi$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CED</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. PFA and CED versus change-point $\tau$
Author’s Responses

Yajun Mei
H. Milton Stewart School of Industrial and Systems Engineering,
Georgia Institute of Technology, Atlanta, Georgia, USA

Abstract: In this rejoinder I briefly summarize my thoughts on appropriate measures of performance for evaluating change-point detection schemes, particularly the false alarm criterion. Then I address some specific issues in the light of the discussion pieces from eight experts in this field.

Keywords: Average run length; CUSUM; Expected false alarm rate; Quantile run length; Statistical process control; Surveillance.

Subject Classifications: 62L10; 62L15; 60G40.

1. INTRODUCTION

First of all, I would like to thank all the discussants for their valuable insights and for commenting on identifying measures of performance that are useful and appropriate for evaluating change-point detection schemes, especially when the observations are not independent. In the paper, I focused on inappropriateness of the average run length (ARL) to false alarm criterion in certain scenarios. I also proposed a couple of alternative false alarm criteria specifically for the examples considered in the paper. According to an old Chinese saying, however, this paper was meant “to throw out a brick to attract a jade.” From this respect, I think that this paper has achieved its goal. The discussants have provided deeper and broader discussions on the issues I raised as well as offered many fruitful ideas on alternative performance criteria. These discussions and ideas will be invaluable for further research in the development of the updated change-point detection theory and for my current and future research in particular.
For the sake of brevity, I will restrict the remainder of this piece to those issues that I feel need emphasis or to which I feel I can add some useful input. To better organize my rejoinder, I divide my remarks into two sections. Section 2 discusses the background of the paper and some issues related to the examples in the paper. Section 3 considers alternative mathematical formulations for change-point detection.

2. THE BACKGROUND AND EXAMPLES

Let me start with a story about one of my papers, Mei (2006). It used a simpler version of the main example in the present paper to show that Page’s CUSUM or Shirayev–Roberts procedures can be asymptotically suboptimal under a standard minimax formulation for dependent observations. When I submitted its earlier version to the journal, a reviewer correctly pointed out that the suboptimal properties were caused by the standard minimax formulation and the issue did not appear under the Bayesian formulation. However, the reviewer thought that this was a strong evidence against Lorden’s “worst-case” detection delay criterion! The true reason, of course, is due to the ARL to false alarm criterion.

This story is intended not as a criticism of the reviewer’s initial reactions, but as evidence supporting my assertion that the ARL as a false alarm criterion has been well accepted in the field, and few researchers questioned its appropriateness (although some concerns have been raised before). This motivated me to write the present paper to highlight some drawbacks of the ARL to false alarm criterion. With this in mind, it is time to respond to some of the points raised by the discussants on the examples in the paper.

Professors Schmid and Knoth and Professor Pollak questioned the half-Cauchy prior used in my main example. The Cauchy prior was chosen to highlight the main message that the ARL to false alarm criterion may lead one to find a detection scheme that focuses larger changes (θ negatively large) instead of desired smaller changes. Because larger changes play a more important role under the ARL to false alarm criterion, from the asymptotic optimality viewpoint, our example with the mixture prechange distribution $f$ essentially becomes the problem of detecting a change in the mean from $-\infty$ to $\lambda$ ($>0$). Thus it is not surprising that we can construct a detection scheme $T$ with a finite detection delay that can have infinite ARL to false alarm $E_f(T)$, because $E_{\theta<0}(T) = \infty$. In retrospect, however, the choice of half-Cauchy prior may be too extreme and may actually mask our main message. Probably I should have used a more reasonable half-normal distribution as a prior $\pi(\theta)$, in which case the scheme $T^*(a)$ in (3.5) of the paper would still have an infinite ARL to false alarm with a finite detection delay. The only difference is that Page’s CUSUM procedure $T_{CM}(\theta_0, b)$ would have finite ARL to false alarm. No matter whether we use a half-Cauchy prior or a half-normal prior, or no matter whether the ARL to false alarm is finite or infinite, our main message remains the same: the ARL to false alarm criterion is not appropriate in our examples.

Professor Tartakovsky showed that the problematic issues associated with the ARL to false alarm criterion do not exist under a Bayesian formulation. Specifically, in a simpler version of our main example, Page’s CUSUM and Shirayev–Roberts procedures are asymptotically optimal under the Bayesian formulation, although they are asymptotically suboptimal under the ARL to false alarm criterion. Here
is a more intuitive explanation of this phenomena. In the example, asymptotically optimal detection schemes under the ARL to false alarm criterion need to detect larger changes efficiently, whereas those under the Bayesian formulation need to detect smaller changes efficiently. It turns out that Page’s CUSUM and Shiryayev–Roberts procedures are very inefficient to detect larger changes compared to the optimal scheme with exact knowledge of prechange parameters, but they are very efficient to detect smaller changes. In other words, a detection scheme that is suboptimal under the ARL to false alarm criterion does not necessarily mean that it is a poor detection scheme. This is consistent with our message of the appropriateness of performance criteria.

I fully agree with Professor Zacks on the importance of investigating the conditional expected delay and the probability distribution of false alarms. For any given detection scheme \( T \), we should analyze its properties under both \( P_v \) and \( P_{\infty} \) measures. Ideally, \( P_v \) distribution of \((T - v)^+\) will be “stochastically small” for every \( v \geq 1 \), whereas \( P_{\infty} \) distribution of \( T \) will be “stochastically large.” To define a rigorous mathematical optimality problem, this viewpoint is simplified under the standard minimax formulation. The \( P_v \) distribution of \((T - v)^+\) is generally simplified to Lorden’s “worst-case” detection delay or the “average” detection delay proposed by Shiryayev or Pollak. The \( P_{\infty} \) distribution of \( T \) typically leads to \( E_{\infty}(T) \). Such a simplified formulation is known to be useful for change-point problems with independent observations. The aim of my paper was to illustrate that this simplification could be flawed for dependent observations, because it may no longer reflect the \( P_v \) or \( P_{\infty} \) distribution of \( T \) in meaningful ways. From this perspective, Professor Zacks and I actually discussed the same issue, although from different angles.

I thank Professors Aston, Fuh, and Luo for providing a deeper insight to the issues I raised for hidden Markov models (HMMs). They show that the key issues involve irreducibility and communicating class restrictions. They also distinguished the difference between the structure change and the parameter change in HMM. However, all ARLs in their numerical studies (except one in Table 2 with \( \mu = 0 \)) are what I would like to call short ARL, which is the detection delay when the change-point occurs at time \( v = 1 \). On the other hand, the focus of my paper was what I would like to call long ARL, which characterized the frequency of false alarm. Although their numerical simulations illustrate that the short ARL seems to be a reasonable criterion of detection delay, it is not clear whether the long ARL is still appropriate as a false alarm criterion, or specifically, whether the detection scheme for HMM is still asymptotically exponentially distributed under \( P_{\infty} \). It will be helpful to conduct some numerical experiments similar to what Professor Tartakovsky has done in his discussions.

Professors Lai and Chan claimed in Section 3 of their discussions that it is possible to achieve a finite detection delay under a probability constraint on false alarm: \( P_{\infty}(T < \infty) \leq \alpha \). The tricky point here is to identify the definition of detection delay! Professors Lai and Chan’s claim is true for individual detection delay \( E_{v, \lambda}^{(v)}(T - v \mid T > v) \) for each and every given change-point \( v \). In our paper, however, the detection delay \( D_g(T) \) is defined by taking the superum over all change-point \( v \geq 1 \), e.g.,

\[
D_g(T) = \sup_{1 \leq v < \infty} E_{0, \lambda}^{(v)}(T - v \mid T > v)
\]
or \( D_g(T) = \text{Lorden's "worst-case" detection delay} \ E_g(T) \). Also see Moustakides (2008) for a general framework for the appropriate definition of detection delays. Under our definition of detection delay in \( D_g(T) \), Professors Lai and Chan’s claim does not hold for independent observations. In fact, as pointed out by Pollak and Siegmund (1975, Section 6), for independent observations, under the probability constraint, it is \textit{impossible} that \( E_{g_0}^{(0)}(T - v | T > v) \) remain bounded as \( v \) goes to \( \infty \), implying that \( D_g(T) = \infty \). Yakir (1996) further showed that \( E_{g_0}^{(0)}(T - v | T > v) \) generally grows like \( \log(v) \) as \( v \) goes to \( \infty \) under the probability constraint. For \textit{dependent} observations, Professors Lai and Chan’s claim is rendered as a conjecture under our definition of detection delay in \( D_g(T) \), and it will be interesting to investigate whether one can construct a scheme with finite detection delay \( D_g(T) \) under the constraint \( P_T(T < \infty) \leq \alpha \). New approaches or examples are needed to prove or disprove this conjecture, since all schemes in this paper will raise a false alarm with probability 1.

Finally, it is informative to note that the mixture prechange distribution, or the assumption that the prechange parameter is a random variable, occurs frequently in biomedical applications. In these applications, it is standard to borrow the knowledge at the population level to make decisions at the individual level, because it is impossible to build an accurate model at the individual level. For instance, McIntosh and Urban (2003) applied the aspiration of change-point detection to cancer screening based on longitudinal observations of a biomarker, where the prechange parameter for a particular subject is assumed to be a random variable with a distribution function estimated from normal subjects.

3. POSSIBLE ALTERNATIVE FORMULATIONS

One consensus topic on alternative mathematical formulations of change-point detection is a Bayesian formulation, in which the change-point \( v \) is assumed to be a random variable, because my example with mixture prechange distribution has an obvious Bayesian flavor. When an unknown parameter \( \theta \) is present in the distribution of observations, the standard asymptotic optimality method is generally to assume a prior distribution \( \pi(\theta) \), consider the corresponding Bayes risk function. Then, investigate the asymptotic Bayes solutions by fixing \( \pi(\theta) \) and by letting some cost parameter \( c \) converge to 0. Such an approach has been successfully applied to change-point detection problem with composite/unknown postchange distributions.

Motivated by the results in this paper, a new asymptotic approach to the Bayesian formulation is to assume that the prior distribution \( \pi(\theta) = \pi_c(\theta) \) depends on the cost parameter \( c \) as \( c \) goes to 0! Such an approach is particularly valuable when the prechange distribution involves unknown parameter \( \theta \). It is worth emphasizing that the scenario with composite/unknown prechange distributions is significantly different from the scenario with composite/unknown postchange distributions. This is illustrated by Professor Yakir in Section 2 of his discussions. The following is another intuitive explanation. For a given family of detection schemes \( \{N(a)\} \), the detection delays have the same order, that is, the limit of \( D_{\pi_1}(N(a))/D_{\pi_2}(N(a)) \) exists as \( a \to \infty \). However, the ARLs to false alarm are in the different order, i.e., the limit of \( E_{\pi_1}(N(a))/E_{\pi_2}(N(a)) \) is either 0 or \( \infty \) as \( a \to \infty \), and only the logarithms of the ARLs to false alarm are in the same order. Because the
ARL to false alarm is closely related to the Bayesian formulation as discussed by Professor Pollak in his discussions, a suitable prior distribution for the prechange parameter $\theta$ may be

$$\pi_c(\theta) \sim e^{-p(\theta)D_c},$$

(up to a constant), where $p(\theta)$ may be chosen appropriately and the constant $D_c$ goes to $\infty$ as $c \to 0$. Such a choice of $\pi_c(\theta)$ takes into account that for a given scheme $T$, the ARL to false alarm $E_a(T)$ is generally exponentially increasing function of $\theta$. In a semi-Bayesian context, the corresponding mathematical problem is to balance the trade-off between the detection delay $D = D_a(T)$ and the ARL to false alarm $E_f(T) \sim \int_0^\theta e^{-p(\theta)D_c} E_a(T) d\theta$.

Professors Pollak, Tartakovsky, and Yakir proposed a new false alarm criterion in their respective discussions by considering the conditional probability criterion based on $P_{\infty}(T \leq k + m \mid T \geq k)$. Professors Tartakovsky and Yakir also discussed this new criterion from the viewpoint of global versus local constraints of false alarms, and illustrated its promising potential through some concrete examples. I think this new conditional probability criterion is useful and informative and deserves more attention in both theory and application of change-point detection, as it goes parallel to the detection delay $E_f(T - v \mid T \geq v)$.

Professors Brodsky and Darkhovsky expressed a similar idea of conditional probability criterion, but with their proposed supremal probability of false decision. For a stopping time $T$ defined by $T = \inf \{n : W_n \geq a\}$, the standard definition of the probability of false alarm in the first $n$ observations is

$$P_{\infty}(T \leq n) = P_{\infty}\left(\max_{1 \leq i \leq n} W_i \geq a\right),$$

whereas Professors Brodsky and Darkhovsky propose to consider

$$\max_{1 \leq i \leq n} P_{\infty}(W_i \geq a).$$

When the $W_n$’s are the CUSUM statistics and the $X_n$’s are i.i.d. under the probability measure $P_{\infty}$, it is well-known (e.g., Iglehart, 1972) that under $P_{\infty}$, $W_n$ converges to a finite limiting random variable $W_\infty$, but $\max_{1 \leq i \leq n} W_i$ grows like $\log(n)$ as $n$ goes to $\infty$. Thus, for any given detection threshold $a$, the probability of false alarm under the standard definition is $P_{\infty}(\max_{1 \leq i \leq n} W_i \geq a) = 1$, whereas the supremal probability of false decision is well-defined as $P_{\infty}(W_\infty \geq a) \in (0, 1)$. Hence, the supremal probability of false decision is a well-defined probability and seems to be an appealing alternative measure of false alarms.

Professors Lai and Chan proposed unconditional probability of false alarms and developed a general asymptotic theory based on $\sup_{k} P_{\infty}(k \leq T \leq k + m)$ and $E_a(T - v)^+$. However, I find it difficult to treat $E_a(T - v)^+$ as an appropriate criterion of detection delay, and I am not sure how useful this asymptotic theory may be in practice. Nevertheless, I do agree with Professors Lai and Chan that it is meaningful and useful to consider $P_{\infty}(T \leq m)$, which can be thought of as a special case of either conditional or unconditional probability of false alarms, $P_{\infty}(T \leq k + m \mid T \geq k)$ or $P_{\infty}(k \leq T \leq k + m)$, when $k = 0$.

Professor Yakir provided a detailed analysis on using the maximal invariant statistics to tackle the change-point problems with unknown prechange
distributions. If we denote by $\theta$ the value of unknown prechange distribution, then the properties of the detection scheme $N_A$ constructed from the invariance statistics do not depend on $\theta$ when there is no change! That is, $P_{\theta}(N_A \leq m + k \mid N_A \geq k)$ or $E_{\theta}(N_A)$ does not depend on $\theta$ when $\theta \in \Theta$, the space of prechange parameters. Translating into a minimax formulation, this is equivalent to placing a constraint on

$$\inf_{\theta \in \Theta} P_{\theta}(N_A \leq m + k \mid N_A \geq k) \text{ or } \inf_{\theta \in \Theta} E_{\theta}(N_A).$$

In the examples considered by Professor Yakir, such an approach, of course, gives a better formulation than the “mixture” approach that considers $\int_{\Theta} E_{\theta}(N_A) \pi(\theta) d\theta$ as a criterion.

When the prechange and postchange distributions are “separable,” our proposed expected false alarm rate (EFAR) seems to be a good competitor. I am glad to hear from Professors Schmid and Knoth that their numerical study confirms that $\text{EFAR}_\theta(T) \approx q/\xi_q(T)$, when $E_{\theta}(T)$ is large and $q$ is in some small subintervals. This is consistent with our heuristic arguments on a close relationship between EFAR and the quantile run length. It will be interesting to investigate more rigorously the accuracy level of this approximation or its speed of convergence under the nonasymptotic scenarios.

4. CONCLUSIONS

I hope that this paper will stimulate further theoretical developments for changepoint detection beyond the customary simplest model, because it is clear that further research is needed to improve the theory, methodology, and algorithms for changepoint detection. I will also be extremely pleased if practitioners and researchers will pay more attention to the measure of performance itself in real-world applications.

ACKNOWLEDGMENT

I would like to thank Professor Mukhopadhyay for his encouragements and valuable comments throughout this work.

REFERENCES


