Optimal Surveillance
Based on
Exponentially Weighted Moving Averages

Marianne Frisén and Christian Sonesson

Statistical Research Unit, Göteborg University,
Göteborg, Sweden

Abstract: Statistical surveillance is used to detect an important change in a process as soon as possible after it has occurred. The EWMA method is used in industry, economics and medicine. Three optimality criteria of surveillance are studied. The ARL criterion violates commonly accepted inference principles and the drawbacks are demonstrated. The ED criterion is based on the minimal expected delay from change to detection. The full likelihood ratio method is optimal according to this criterion. Approximations of this method turn out to be modifications of the EWMA method. The approximations lead to a formula for the optimal value of the smoothing parameter of the EWMA statistic. The usefulness of this formula is shown. It is demonstrated that, for EWMA, the minimax criterion agrees well with that of the ED criterion but not with that of the ARL criterion.

Key words: ARL; Expected delay Monitoring; Minimax; Quality control; Stopping rule; .

Subject Classifications: 62L15; 62C10, 62C20.

*Address correspondence to: Marianne Frisén, Statistical Research Unit, Göteborg University, Box 660, SE-40530 Göteborg, Sweden;. Fax: 46-31-7731274; E-mail: marianne.frisen@statistics.gu.se.

1. INTRODUCTION

Continual surveillance of time series, with the goal of detecting an important change in the underlying process as soon as possible after it has occurred is needed in many areas. An example is the surveillance of the foetal heart during labor described by Frisén (1992). An abnormality, caused by e.g. a lack of oxygen due to the umbilical cord around the neck of the foetus might happen at any time. If an alarm is given soon after the event has occurred it is possible to rescue the baby by e.g. a Cesarean section. A review of methods for the surveillance in public health is given by Sonesson and Bock (2003). Recent applications using the EWMA method include VanBrackle and Williamson (1999) and Williamson and Hudson (1999) where the detection of an increased incidence of a disease is studied. Environmetric control is described by e.g. Pettersson (1998). Applications in economics, and especially the surveillance of business cycles, are treated in e.g. the special issue (no. 3/4, 1993) of "Journal of Forecasting" and by Andersson et al. (2004). Financial decision strategies using EWMA are described by e.g. Severin and Schmid (1998), Schipper and Schmid (2001) and Schmid and Tzotchev (2004).

Several methods for surveillance have been suggested in the literature. Broad surveys and bibliographies on surveillance are given by Lai (1995), who concentrates on minimax properties of stopping rules and by Woodall and Montgomery (1999), who concentrate on control charts. Here we concentrate on the method for surveillance based on exponentially weighted moving averages, usually called EWMA. It is useful for the analysis of several important issues of inference in connection with surveillance since it has a simple form, can be expected to have good properties, and lacks the discontinuity of the CUSUM method. The EWMA method was introduced in the quality control literature by Roberts (1959) and the method has got much attention. This may be due to papers such as Robinson and Ho (1978), Crowder (1987), Ng and Case (1989), Lucas and Saccucci (1990) and Domangue and Patch (1991) in which positive reports on the quality of the method are given. Also, Srivastava and Wu (1997) pointed out the fact that the EWMA control chart can easily be used to estimate the current mean and to make optimal forecasts.

The concept of optimality in on-line surveillance is discussed by, e.g. Frisén and de Maré (1991) and Lai (1995). In a surveillance situation, there is no fixed data set and not even a fixed hypothesis to be tested. The choice of an optimality criterion is an interesting and important issue. Optimal EWMA is discussed by e.g. Crowder (1987), Lucas and Saccucci (1990) and Srivastava and Wu (1997). These papers all consider an optimality criterion based on the average run length when a change either happens immediately or not at all. Earlier studies on optimal EWMA have mainly focused on this criterion. Here, also other optimality criteria are applied to the EWMA method. The optimality condition based on minimal expected delay from change to detection, suggested by Girshick and Rubin (1952) and Shiryaev (1963) is used. The full likelihood ratio method is optimal according to this criterion. Various approximations of this method turn out to be modifications of the EWMA method. Also, a minimax criterion and its relation to the other two criteria are studied. The criteria are described in Section 4 and analyzed in the following sections.

Most studies on EWMA are for the two-sided case. Important exceptions are the papers by Robinson and Ho (1978), Srivastava and Wu (1993), Wu (1994)and Han and Tsung (2004), where properties of one-sided methods are analyzed. Here we discuss the differences between one- and two-sided EWMA but concentrate on the one-sided case, in order to concentrate on some inferential conclusions. For the one-sided EWMA there are remarkable results which are not present for two-sided EWMA. Only surveillance of a process of independent and univariate observations is examined in order to focus on some critical issues of optimality.

In Section 2 some notations are given and the case studied is specified. In Section 3 some

properties of some variants of the EWMA method are described. In Section 4 the criteria of optimality which are studied are described and analyzed. In Section 5 different constructions of optimal EWMA are described and motivated. In this section a large scale simulation study on different properties for EWMA and competitors are reported. Section 6 contains a discussion. The reliability of the simulation study is reported in the Appendix.

2. NOTATIONS AND SPECIFICATIONS

At each decision time s, s=1, 2, ... we want to discriminate between the two events C(s), which requires an action, and D(s), which does not. The decision at time s is based on $X_s = \{X(t): t = 1,2,...,s\}$, where X(t) is the observation made at time t. The observation may be an average or some other derived statistic. The random process that determines the state of the system is denoted by :={:(t), t = 1,2, ...}. Given :, the observations are assumed independent with the same known standard deviation, Φ. At some unknown time point ϑ, there is a change in the distribution of X(t). As in most literature, the case of a shift in the mean of a Gaussian random variable from a value $:^0$ to another value $:^1$ is considered. Only one-sided alternatives are considered here, where $:^1>:^0$. Here, $:^0$ and $:^1$ are regarded as known values and without loss of generality we set $:^0=0$, the standard deviation Φ=1 and $:^1$ is denoted :. The process suddenly shifts at time ϑ and remains at the new level. That is,

$$:(t) = 0 \text{ for } t = 1, ..., \vartheta-1 \text{ and } :(t) = : \text{ for } t = \vartheta, \vartheta+1, ...$$

The time point, ϑ, when the change in the distributions of X(t) occurs, is regarded as a random variable. We study the case of a constant intensity $< = P(\vartheta=t \mid \vartheta \geq t)$.

The distribution of ϑ is thus geometric with $P(\vartheta=t) = <(1-<)^{t-1}$. We aim to discriminate between the two events C(s) = {ϑ#s} and D(s) = {ϑ>s} by the set of observations X_s. We will consider different ways of constructing alarm sets A(s) with the property that, when X_s is a subset of A(s), there is an indication that the event C(s) has occurred. The time of the first alarm is $t_A = \min\{s: X_s \delta A(s)\}$. For the methods studied here this can be expressed as

$$t_A = \min\{s: p(X_s)>K(s)\},$$

where $p(X_s)$ is an alarm statistic and K(s) is an alarm limit. In the Table below, all the methods and their alarm criteria are presented.

3. THE EWMA METHOD

The EWMA method for surveillance has an alarm statistic based on exponentially weighted moving averages,

$$Z_s = (1-8)Z_{s-1}+8X(s), \quad s=1, 2, ...$$

where $0<8\leq1$ and here, as usual, Z_0 is the target value $:^0$, which is set to zero. The alarm statistic is sometimes referred to as a geometric moving average. It can equivalently be written as

$$Z_s=\lambda(1-\lambda)^s\sum_{t=1}^{s}(1-\lambda)^{-t}X(t).$$

The EWMA statistic gives the most recent observation the greatest weight, and gives all previous observations geometrically decreasing weights. If 8 is near zero, all observations

have approximately the same weight. If 8 is equal to one, only the last observation is considered as in the Shewhart method. Shewhart (1931) suggested that an alarm should be given at

$$t_A = \min\{s: X(s) > L\}, \qquad (3.1)$$

where L is a constant.

Table. Notations for the methods

Method	Expression	Alarmstatistic	Alarmlimit	Formula
Shewhart		X(s)	L	(3.1)
EWMAa	Standard version	Z_s	$L\Phi_Z$	(3.2)
	With normalized weights	$Zw_s = \sum_{t=1}^{s} w_{EWMA}(s,t)X(t)$	$K_{EWMAa}(s)=L\Phi_Z/[1-(1-8)^s]$	(3.3)
EWMAe	Standard version	Z_s	$L\Phi_{Zs}$	(3.4)
	With normalized weights	$Zw_s = \sum_{t=1}^{s} w_{EWMA}(s,t)X(t)$	$K_{EWMAe}(s) = L\Phi_{Zs}/[1-(1-8)^s]$	(3.5)
SCUSUM		$\sum_{t=1}^{s} X(t)/s$	L/s	(5.1)
LCUSUM		$\sum_{t=1}^{s} X(t)/s$	$\mu/2 + L/s$	(5.2)
LR		$\sum_{t=1}^{s} P(\tau=t)\exp\{t\mu^2/2\}\exp\{\mu\sum_{u=t}^{s} X(u)\}$	$\exp\{(s+1)\mu^2/2\}P(\tau>s)\frac{L}{1-L}$	(5.3)
LinLR		$\sum_{t=1}^{s} w_{LinLR}(s,t)X(t)$	$K_{LinLR}(s)$	(5.4)
EWLR		Zw_s	$K_{EWLR}(s)$	(5.5)
EWlnLR		Zw_s	$K_{EWlnLR}(s)$	(5.6)

If no change ever occurs, the expected value of the EWMA statistic is $E(Z_s|:(t)/0) = 0$. If the change occurs immediately, we have $E(Z_s|:(t)/:) = :(1-(1-8)^s)$ 6: as s 6 4. The variance of the statistic is $\Phi^2_{Zs}= 8[1-(1-8)^{2s}]/(2-8)$ which tends to $\Phi^2_Z = 8/(2-8)$ when s 6 4. The rate of convergence of the variance depends on 8. For small values of 8 it is very slow. However, the use of the asymptotic variance is very common and it can in fact be preferred to the exact version in some aspects, see for example Sonesson (2003). This procedure will be referred to as the "asymptotic version". When it is necessary to distinguish between the exact and the asymptotic version, the notations EWMAe and EWMAa will be used. Most of the results below concern EWMAa. When not otherwise indicated it is this version which is analyzed. EWMAa will give an alarm at time t_A, where

$$t_A = \min\{s: Z_s > L\Phi_Z\}, \qquad (3.2)$$

where L is a constant. For EWMAe the exact standard deviation is used in the alarm limit instead of the asymptotic. For the comparison with other methods in the following sections it is useful to express the alarm statistic with normalized weights $w_{EWMA}(s,t) = 8(1-8)^{s-t}/[1-(1-$

$8)^s]$, which sum to one. The alarm criterion in (3.2) is equivalent to

$$t_A = \min\{s: Z_s/[1-(1-\lambda)s] > L\sigma_Z/[1-(1-\lambda)s]\}$$

which can be written as

$$t_A = \min\{s: Zw_s > K_{EWMAa}(s)]\} \tag{3.3}$$

where

$$Zw_s = \sum_{t=1}^{s} w_{EWMA}(s,t)X(t)$$

and

$$K_{EWMAa}(s) = L\Phi_Z /[1-(1-8)^s].$$

The EWMAe method gives an alarm if

$$t_A = \min\{s: Z_s > L\Phi_{Zs}\}, \tag{3.4}$$

$$t_A = \min\{s: Zw_s > K_{EWMAe}(s)]\} \tag{3.5}$$

where

$$K_{EWMAe}(s) = L\Phi_{Zs} /[1-(1-8)^s].$$

Small values of 8 will be of special concern in Section 5.1 on ARL optimality. In the other end we have the large values of 8. For 8=1 the EWMA method is identical with the Shewhart method.

Since an alarm for EWMAe (obvious adjustment for EWMAa) is given if the statistic Z_s exceeds the alarm limit, $L\Phi_{Zs}$, where L is a constant, we have that $P(Z_s > L\Phi_{Zs}) = M(-\{L-E(Z_s) /\Phi_{Zs}\})$, where M is the normal cumulative distribution function. For this exact version this probability is constant if the process is in control. Note that this is not the probability of an alarm at s. It is only if 8=1 that successive values of Z_s are independent. We have in that case $ARL^0 = 1/M(-L)$ and $ARL^1 = 1/M(-\{L-:\})$. For other values of 8, there is a dependency between successive values of Z_s, and the dependency increases when 8 decreases. The value of L, which gives a desired ARL^0, is a function of this correlation which in turn is a function of 8. For a fixed value of L the smallest value of ARL^0 is obtained for 8=1 and the largest when it approaches zero. Equivalently, for a fixed value of ARL^0 the largest value of L is obtained when 8=1 and the smallest when it approaches zero.

The use of EWMA for autocorrelated processes is discussed by e.g. Schmid and Schöne (1997), Lu and Reynolds Jr (1999) and Gut and Steinebach (2004). Multivariate EWMA methods are discussed by e.g. Lowry et al. (1992), Tsui and Woodall (1993), Lowry and Montgomery (1995), Prabhu and Runger (1997) and Yeh et al. (2003). Multivariate surveillance by EWMA for time series is discussed by e.g. Kramer and Schmid (1997), Knoth and Schmid (2002) and Rosolowski and Schmid (2003). Here, only surveillance of a process of independent and univariate observations is discussed in order to focus on some critical issues of optimality.

Two-sided EWMA methods can be achieved by running two one-sided surveillance procedures parallel and to make an alarm as soon as one of them signals a change. When the one-sided versions are symmetrical, this is equivalent to making an alarm as soon as

$|Z_s|>L\Phi_Z$. The investigations by Lucas and Saccucci (1990), Crowder (1989) and Srivastava and Wu (1997) use two-sided limits symmetrical around zero. The properties for one- and two-sided EWMA are not easily related because of different relations between successive decisions. Waldmann (1986) suggested that barriers should be used in order to get an approximation of the ARL for the one sided case. The use of $\max\{b, Z_s\}$ (or $\min\{b, Z_s\}$ in the case of negative changes) in the one-sided surveillance has the advantage of the same simple relation between the one- and two-sided versions as for the CUSUM method. A formula by Kemp (1961) can be used for derivation of the ARL properties. Waldmann (1986) suggested a wide barrier in order to get a good approximation of the ordinary method. Champ et al. (1991), Gan (1995) and Gan (1998) also use barriers for the alarm statistic. Barriers avoid bad properties in the "worst possible" case when the change occurs at a time when the earlier observations least favor the detection of the change. Sonesson (2003) examined the inferential differences between two-sided, one-sided with barrier and ordinary one-sided methods. In this paper we will only deal with ordinary one-sided procedures in order to draw attention to certain optimality issues since optimality is most clearly expressed in this situation. However, it should be noticed that the results are not immediately translated to the two-sided case.

4. OPTIMALITY CRITERIA

The performance of a method for surveillance depends on the time point ϑ of the change. Sometimes it is appropriate to express the performance as a function of ϑ, as by Frisén (1992) and Frisén and Wessman (1999). Sometimes, however, a single criterion of optimality is needed. We will here study three such criteria, the first one often used in the quality control literature and in connection with EWMA and the two others often used in literature on general surveillance but seldom in the construction of optimal EWMA. Also other optimality criteria, based for example on the stationary average delay time, SADT, concerning the asymptotic properties when τ tends to infinity, have been suggested in the literature. However, here only three criteria will be studied in detail.

First, a measure which is often used in quality control, and which was suggested by Page (1954), is the average of the run length until the first alarm. The average run length until an alarm, when there is no change in the process under surveillance, is denoted by $ARL^0 = E(t_A|:(s)/0)$. The average run length until detection of a true change (that occurred at the same time as the surveillance started) is denoted by $ARL^1 = E(t_A|:(s)/:) = E(t_A|\vartheta=1)$. In the literature on quality control, optimality is often stated as minimal ARL^1 for fixed ARL^0. This criterion is, for short, named the ARL criterion. Margavio et al. (1995), Woodall and Montgomery (1999) and Carlyle et al. (2000) stated that the use of the ARL criterion usually is recommended in spite of the known fact that the distributions of the alarm time are skewed. Margavio et al. (1995) suggested that the whole distribution should be used. A time dependent limit can be chosen to give the desired distribution. By this, special properties such as fast initial alarms could be designed. However, the possible problem of an influence of ϑ remains. For some situations, the expected delay of an alarm, given that there was no alarm earlier, is only slightly dependent on ϑ but this is not generally true. This will be demonstrated in Section 5.2. Degenerated methods, which would never be used in practice, give minimal ARL^1 for a fixed ARL^0 as demonstrated by Frisén (2003).

Secondly, an important specification of utility is that of Girshick and Rubin (1952) and Shiryaev (1963). The gain of an alarm is a linear function of the difference $t_A-\vartheta$ between the time of the change and the time of the alarm. The loss of a false alarm is a function of the same difference. The criterion of maximization of this utility is named the ED criterion, since the expected delay from a change to the detection is minimized. This criterion will be further discussed in Section 5.2, where the exact definition is presented. Their solution to the

minimization of the expected utility is identical to the LR method described by Frisén and de Maré (1991). The LR method will be used as a benchmark when comparing the expected delay for EWMA and different modifications in Section 5.2. Variants with exponential penalty for delay have been discussed by Poor (1998) and Beibel (2000), but will not be considered here.

The third criterion is the minimax of the expected delay after a change with respect to the time of the change. This will be further discussed in Section 5.3, where the exact definition is presented. It is related to the ED criterion as several possible change times are considered. However, instead of an expected value, which requires a distribution of the time of change, the worst value is used. Thus, minimax solutions, with respect to ϑ, avoid the requirement of information about the distribution of ϑ. Important results on minimax are given by e.g. Pollak (1985), Gordon and Pollak (1995), Yakir (1997) and Yakir et al. (1999). The maximal value of the expected delay is for $\vartheta=1$ for many methods and with a minimax perspective this can be a motivation for the use of ARL^1. However, this argument is not relevant for all methods. It will be demonstrated that the argument is not useful for the EWMA method. A still more pessimistic criterion, the "worst possible case", not only uses the worst value of ϑ but also the worst history of earlier observations. The merits of studies of this criterion have been thoroughly discussed by e.g. Yashchin (1993). Important results based on this minimax criterion were given by e.g. Moustakides (1986), who concludes that the CUSUM method is optimal based on this criterion.

In the studies of the ED criterion and the minimax criterion a fixed value of the total false alarm probability $P(t_A<\vartheta)$ is used here, whereas a fixed value of the ARL^0 is used in the ARL criterion. The relation between these two measures of false alarms is illustrated in Section 5.2. For the minimax criterion ARL^0 is often used in literature, as will be discussed in Section 5.3.

5. OPTIMAL EWMA

The properties of the EWMA methods depend both on the value of 8 in the alarm statistic and on the alarm limit. The aim is to achieve a combination of limits and values of 8, which results in a total optimality. For the EWMA method there is a restriction of how the given formula for the alarm limit should be combined with weights in the alarm statistic. When optimal weights and optimal limits can be combined, better exponentially weighted methods are achieved as will be seen below.

An important feature of a method is how the alarm limit depends on the decision time s. The unavoidable false alarms can be allowed early or late. This error-spending strategy is important for the properties of the method. It might appear as if the error-spending by EWMAe is independent of 8 since $P(Z_s > L\Phi_{Zs} \mid :=0) = M(-L)$ is independent of 8. However the correlation between successive Z_s, and thus the error-spending, depends on 8.

In order to investigate the properties of methods according to the different optimality criteria we need estimates of the values of different measures. Numerical approximations involve problems, as discussed in the next section. Here a simulation study was performed to investigate the properties of methods. The cases studied were chosen to be representative for situations in practice but only cases which could be studied with enough exactness by simulations were chosen. A special study, reported in the Appendix, demonstrates that the technique used, and the very large number of replicates, allow safe conclusions.

5.1 Minimal ARL^1 for a fixed ARL^0

5.1.1 Choice of value of the parameter 8.

ARL^1 and ARL^0 are expectations under the assumption that there are equal distributions for all observations under each of the two alternatives. Statistical inference with the aim of

discriminating between the alternatives that all observations have the expected value 0, or all have the expected value :, should by the ancillarity principle not be based on the time of the observation. Thus, one should not give unequal weight to the late and old observations. However, the ARL-criterion must not necessarily agree with generally accepted principles of inference.

Frisén (2003) demonstrated that there exist methods with equal weights for all observations which are ARL optimal. This is another reason to choose equal weights for the EWMA method. To get equal weights to all observations by the EWMA method, 8 should approach zero.

The variance for the standardized alarm statistic, Zw_s, has the smallest variance when 8 approaches zero. As the expected value of Zw_s is constant (0 or : for the conditions for ARL^0 or ARL^1 respectively) the coefficient of variation of the statistic is minimized. This is still another argument for using equal weights if the aim is only to discriminate between these two alternatives.

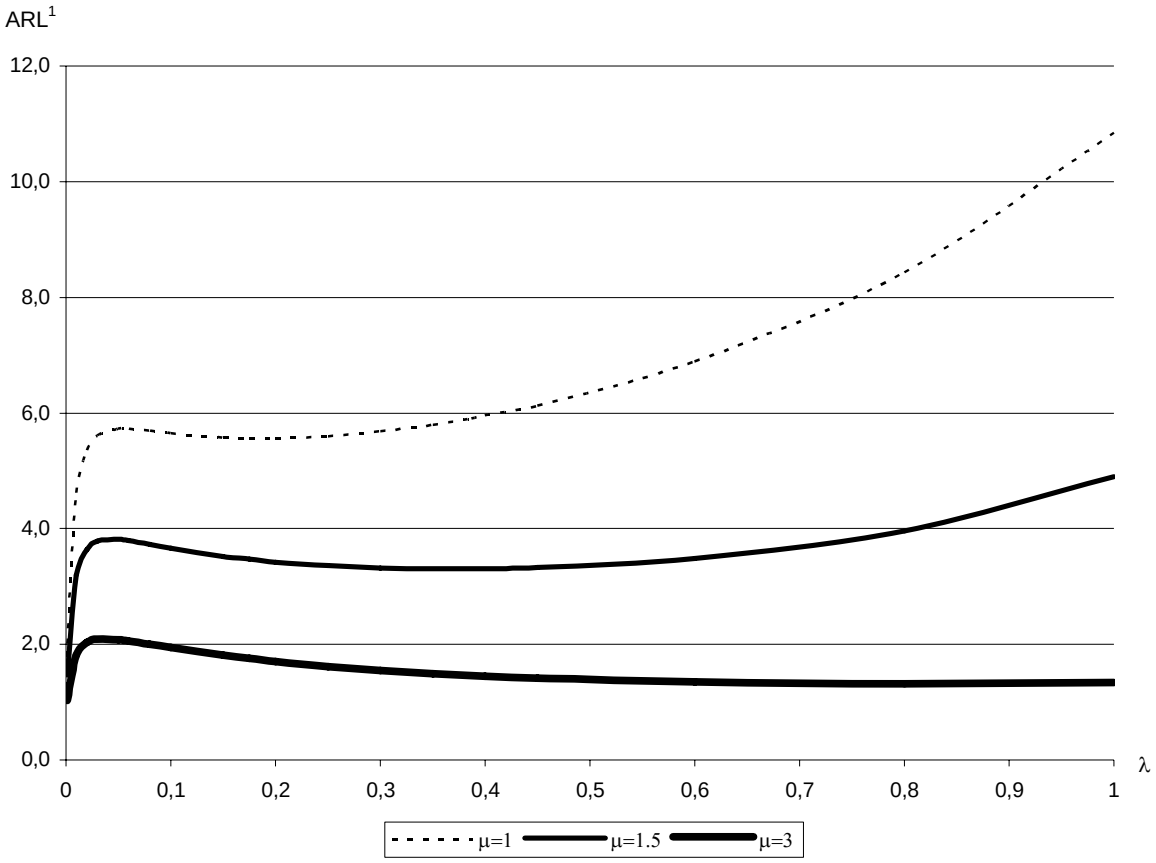

Figure 1. ARL^1 as a function of 8 for EWMA when $ARL^0 = 100$ and :=1, 1.5 and 3.

Figure 1 supports the conclusion that 8 should approach zero to satisfy the ARL criterion, irrespectively of the size of the shift. In that figure the observations for 8=0 and 8=1 are based on theory while the others are based on simulations. Also, Chan and Zhang (2000) observed by simulations that the ARL^1 decreases when 8 decreases. The suggestion was to impose a restriction that the variance of the run length should be small, which it is not for small values of 8. The conclusion that 8 approaching zero is optimal is avoided by this alternative criterion.

One would intuitively give more weight to the later observations as compared to past ones if one might suspect that the past observations might come from the in-control distribution. This is also the solution using the ED criterion (Section 5.2) or the minimax criterion (Section 5.3). These two criteria take into account the possibility of shifts occurring at different time points. However, observe that all observations are out of control when ARL^1 is calculated since $\tau=1$. In that case there is an advantage in giving all observations equal weights.

Different computational techniques to numerically approximate the ARL have been suggested. Here, we mention some which have been directly devoted to find "optimal" values of the parameter 8 in the EWMA method. For one-sided EWMA, Robinson and Ho (1978) used Edgeworth series expansion. According to Lucas and Saccucci (1990) this approximation is inaccurate for small values of 8. The algorithm does not converge for small values of 8. Crowder (1989) used integral calculations suggested by Crowder (1987). Lucas and Saccucci (1990) suggest that the EWMA statistic should be represented as a continuous-state Markov chain, whose properties can be approximated by a finite-state Markov chain following a procedure similar to that of Brook and Evans (1972). Srivastava and Wu (1993) use a continuous time model and an explicit formula for average run lengths for one-sided EWMA. Continuous time approximations are not good for all discrete time situations. Wu (1994) uses a correction by an approximation of the overshoot, which is claimed to be good enough for practical cases but not for optimality considerations. This numerical approximation is used for determination of the value of the parameter which makes the one-sided discrete time EWMA SADT optimal . Srivastava and Wu (1997) give a correction term which is used to determine the parameter for the ARL optimal two-sided EWMA. However this correction term is not satisfactory for very small values of the parameter. Han and Tsung (2004) compare their generalized EWMA, which does not use knowledge of the size of the shift, with the method by Wu (1994) which is named "optimal EWMA" by an analytical comparison of the ARL properties of one-sided EWMA and by a simulation study of the two-sided EWMA.

The results concerning the optimal value of 8, by the numerical approximations in the publications mentioned above, are not in agreement with the theoretical arguments that 8 should approach zero to give a statistic which is suitable for the ARL criterion. The explanation for this might be that the numerical approximations are not good enough for very small values of 8. Very accurate numerical approximations by Knoth (2004) support our result that for one-sided EWMA optimal ARL is achieved when 8 60. Besides, the present result is for the one-sided case and some of the results mentioned above are for the two-sided case.

5.1.2 The alarm limit.

In the preceding section we found that, for ARL optimality, equal weights for all observations should be considered. Still, we have to construct the optimal alarm limit. We will now study alarm limits for EWMA when 8 60.

For the EWMAe method (see formula 3.5) using the exact variance $\Phi^2_{Zs} = 8[1-(1-8)^{2s}]/(2-8)$ we have the alarm limit:

$$K_{EWMAe}(s) = L\Phi_{Zs}/[1-(1-8)^s] = L\,(8[1-(1-8)^{2s}]/(2-8))^{1/2}/[1-(1-8)^s]\ 6\ L/s^{0.5} \text{ as } 8\ 60.$$

The method tends to a repeated likelihood ratio test method (see Siegmund (1985) p 86, 98) as 8 approaches zero. For EWMAa (see formula 3.3) the situation is more complicated since L_{EWMAa} tends to infinity for all s, for a fixed L, when 8 approaches zero. The ratio between the

limit for a certain s and that for s=1, will reflect how the alarm-limit depends on s. The ratio

$$K_{EWMAa}(s)/ K_{EWMAa}(1) \; 6 \; 1/s \text{ as } 860.$$

A related method with an alarm limit which is exactly proportional to 1/s is the "simple CUSUM" (SCUSUM) method described by Frisén (2003). The method is seldom used but is of interest as a limiting case of the EWMA method. The SCUSUM method gives an alarm at

$$t_A = \min\{s: \sum_{t=1}^{s} X(t)/s > L/s\} \qquad (5.1),$$

where L is a constant. Thus, the alarm statistic is the unweighted average of the observations which is also the alarm statistic of the EWMA method when 860. An alarm is triggered when this average exceeds L/s or equivalently: when the simple cumulative sum of the observations exceeds a fixed limit L.

Since the ARL criterion (minimum ARL^1 for a fixed ARL^0) can be criticized, we will also briefly examine how the EWMA satisfies a modification, the criterion of minimum ARL^1 for a fixed false alarm probability, ($P(t_A < \tau)$. The LCUSUM method, demonstrated by Frisén (2003)to be optimal for this criterion is based on an SPRT, and gives an alarm at

$$t_A = \min\{s: \sum_{t=1}^{s} X(t)/s > \mu/2 + L/s\} \qquad (5.2),$$

where L is a constant. Altering the criterion to have a fixed false alarm probability instead of a fixed ARL^0 leads to a much more reasonable "ARL-optimal" method than the "Two-point method" which was demonstrated by Frisén (2003) to be optimal for the usual ARL criterion (for a fixed ARL^0). However, the LCUSUM method is seldom used for surveillance. It is suitable for a test of a hypothesis of τ=1, but also serves as a theoretical benchmark for surveillance methods. In Figure 3 in the next section the minimal ARL^1 for a fixed false alarm probability is illustrated by the low expected delay for ϑ=1. However, the low expected delay is true only for this value of ϑ. The properties for a later change are very bad. This is so also for the SCUSUM method but less pronounced. Thus, to go beyond the class of EWMA methods you can get still better ARL-properties, but even worse properties for large values of ϑ.

None of the EWMA variants have alarm limits which depend on s in the same way as the LCUSUM method when 8 approaches zero. However, when : tends to zero the LCUSUM method approaches the SCUSUM method and in that case EWMAa approaches the LCUSUM method when 8 approaches zero.

It follows from the general theory of SPRT that the LCUSUM method does not have a finite ARL^0. The same follows for SCUSUM from the theory of Brownian motions and the fact that discrete stopping procedures have larger expected stopping times than the continuous version. When 8 approaches zero, the ARL^0 of EWMAa behaves as for SCUSUM and will not be finite. The alarm limit of EWMAa converges to that of EWMAe when s increases if the same value of L is used. The alarm limit for Z_s is constant for EWMAa and increases faster with s for EWMAe than for EWMAa. Thus, the ARL^0 tends to infinity also for EWMAe. Exact ARL optimality for EWMA is thus hard to discuss. However, as L decreases to -4 in order to get a fixed value of ARL^0, $P(t_A=1)$ increases to 1. The results from the simulations, reported in Figure 1 for EWMAa, illustrates that the ARL-properties will

approach the optimal one with $ARL^1=1$ for a fixed value of ARL^0 when 8 60.

Methods which allocate the power to the first time points, like EWMAe, the Fast Initial Response method by Lucas and Saccucci (1990)or a combination of both and a more direct allocation Steiner (1999) will have good ARL^1 properties but worse ones if the change happens later. This is also true for EWMAa with small values of 8 as illustrated in Figure 3.

5.2 Minimal expected delay

The ED criterion as described in general terms in Section 4, belongs to a wide class of utility functions. The LR method (see Frisén and de Maré (1991)), which is based on the full likelihood ratio, maximizes all these utility functions.

Instead of using the ARL^0 as the false alarm measure as in the ARL criterion we now use a fixed value of the false alarm probability, $P(t_A<\vartheta)$. The relation between the ARL^0 and the $P(t_A<\vartheta)$ is illustrated in Figure 2 for some methods. It is clear that you can expect different results in comparisons depending on which of the restrictions you choose. Methods are favored differently by the two restrictions. For large and moderate values of < it can be seen in Figure 2 that the EWMAa method with a large value of 8 is favored by the ARL restriction, while the LR method (to be described below) optimized for small values of < and : is favored by the restriction on $P(t_A<\vartheta)$.

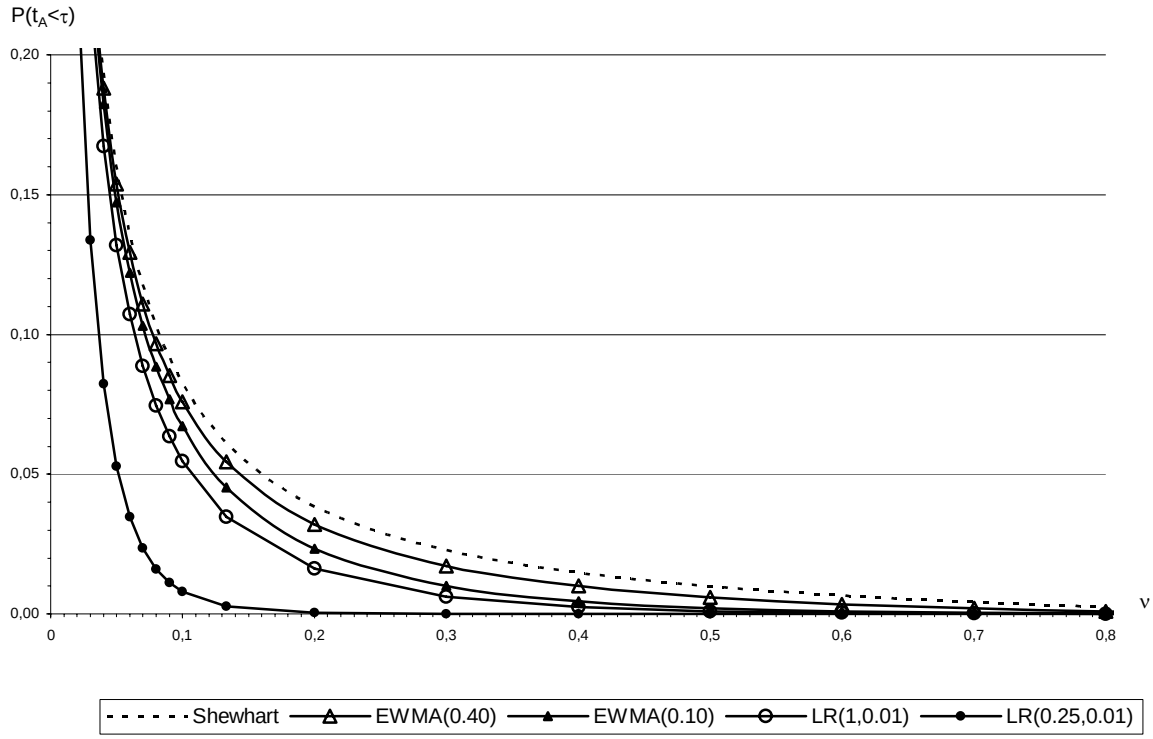

Figure 2. The false alarm probability $P(t_A<\vartheta)$, when $ARL^0 = 100$, as a function of the intensity <.

Let the expected delay from the time of change, $\vartheta=t$, to the time of alarm, t_A, given the time of change, be denoted by

$$ED(t) = E[\max (0, t_A-t) \mid \vartheta=t]$$

To connect with the preceding section, it can be noted that $ED(1)=ARL^1-1$. For most

methods the ED(t) will tend to zero as ϑ increases. The conditional expected delay

$$CED(t) = E[t_A - t \mid \vartheta = t, t_A \exists t] = ED(t) / P(t_A \exists t)$$

on the other hand, will for most methods converge to a constant value. The CED(ϑ) for the EWMA methods, with large values of 8, is fairly independent of ϑ. Only the last observations will have any great influence. Thus the ability to detect a change, given that no alarm has been given earlier, is fairly constant. For 8=1 the EWMA methods equal the Shewhart method, which has a constant CED= ARL^1-1. For large values of 8, the CED curve approaches an asymptote and a constant value of CED for moderate values of τ. However, for small values of 8 the delay for changes which occurred in the beginning differs much from those occurring later on. For small values of 8, CED increases as a function of ϑ. The extreme situation is when λ approaches zero and the CED curve for EWMA approaches that for SCUSUM. In Figure 3 the relation between CED and τ is illustrated for some methods described above or below.

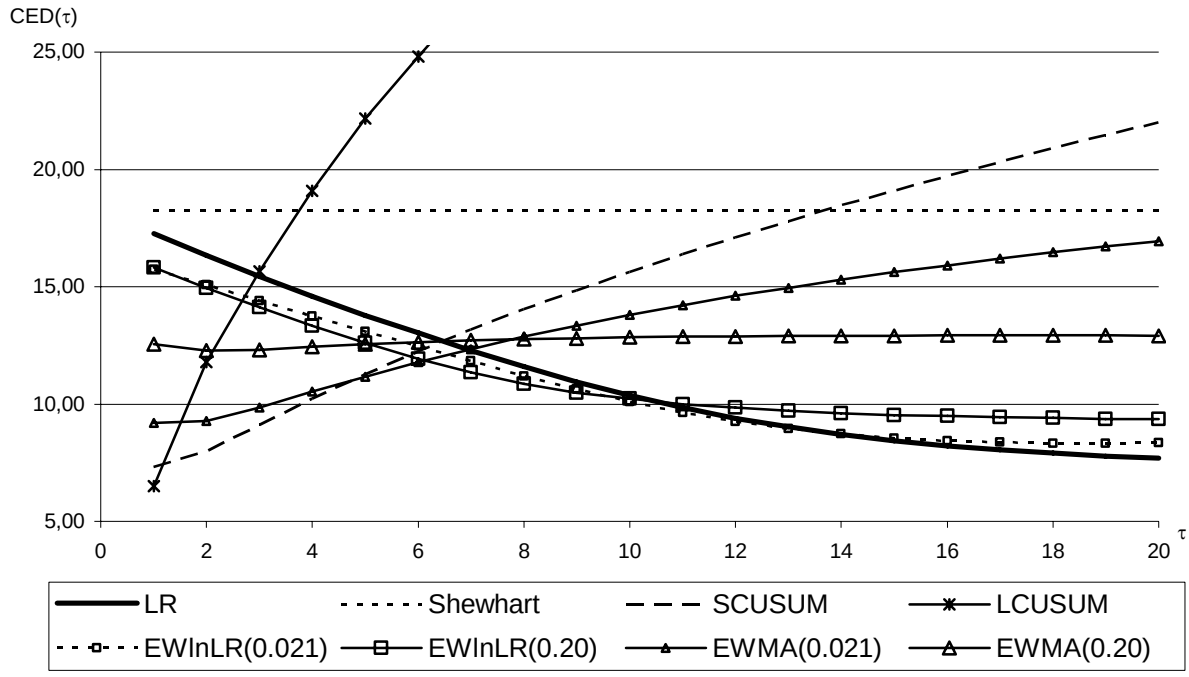

Figure 3. The conditional expected delay CED as a function of ϑ for <=0.01 and :=0.25. The false alarm probability is fixed to 0.75.

The summarizing expected delay

$$ED = E_{\vartheta}[ED(\vartheta)],$$

where the expectation is with regard to the distribution of the time ϑ of the change, is in focus of this section. It satisfies the very general utility function by Shiryaev (1963) where the expected delay of a desired alarm plays an important role. He treated the case of constant intensity of a change where the gain of an alarm is a linear function of the value of the delay, t_A-τ. The loss associated with a false alarm is a function of the same difference. This utility can be expressed as U= E{u(τ, t_A)}, where

$$u(\tau,t_A)=\begin{cases} h(t_A-\tau) & \text{if } t_A<\tau \\ a_1(t_A-\tau)+a_2 & \text{else} \end{cases}$$

The function $h(t_A-\tau)$ is usually a constant (say b), since the cost of alerts and investigations is the same irregardless of how early the false alarm was given. In this case we have

$$U= b\ P(t_A<\tau) + a_1\ ED + a_2\ .$$

Thus, we would have a maximal utility if we have a minimal (a_1 is typically negative) expected delay from the change-point for a fixed probability of a false alarm. The ED criterion seems to be a suitable optimality criterion in many applications because of its generality of including changes occurring at different time points.

The ED is minimized (for a fixed value of the false alarm probability) by the LR method. For the situation specified in Section 2, the LR method gives an alarm for

$$t_A=\min\{s:\ \sum_{t=1}^{s} P(\tau=t)\exp\{t\mu^2/2\}\exp\{\mu\sum_{u=t}^{s} X(u)\} > \exp\{(s+1)\mu^2/2\}P(\tau>s)\frac{L}{1-L} \qquad (5.3),$$

where L is a constant which determines the probability of a false alarm. The alarm criterion can equivalently be expressed by the posterior probability

$$P(\tau \le s|X_s=x_s) > L\ .$$

The special data-analytic interpretation of the values of the statistics of methods with a simple relation to the posterior probability is pointed out by Kenett and Pollak (1996).

The LR method is optimized for the values of : and < used in the alarm statistic. When the values of these two parameters are of interest, they are used as arguments in the notation LR(:,<).

The expected delays for the LR method and the Shewhart method are presented in Figure 4 for some situations. The expected delay of the LR method is used as the benchmark in the following descriptions, since it is the minimal one. The values for the Shewhart method indicate a practical upper limit, even though EWMA with a very small value of 8 has worse values. Thus, the limits between which methods vary for a specific combination of < and $P(t_A<\vartheta)$ can be seen in the figure. For small values of : the intervals are wide, while they are tight for large values. The LR and Shewhart methods converge to have identical properties when : tends to infinity, as proved by Frisén and Wessman (1999).

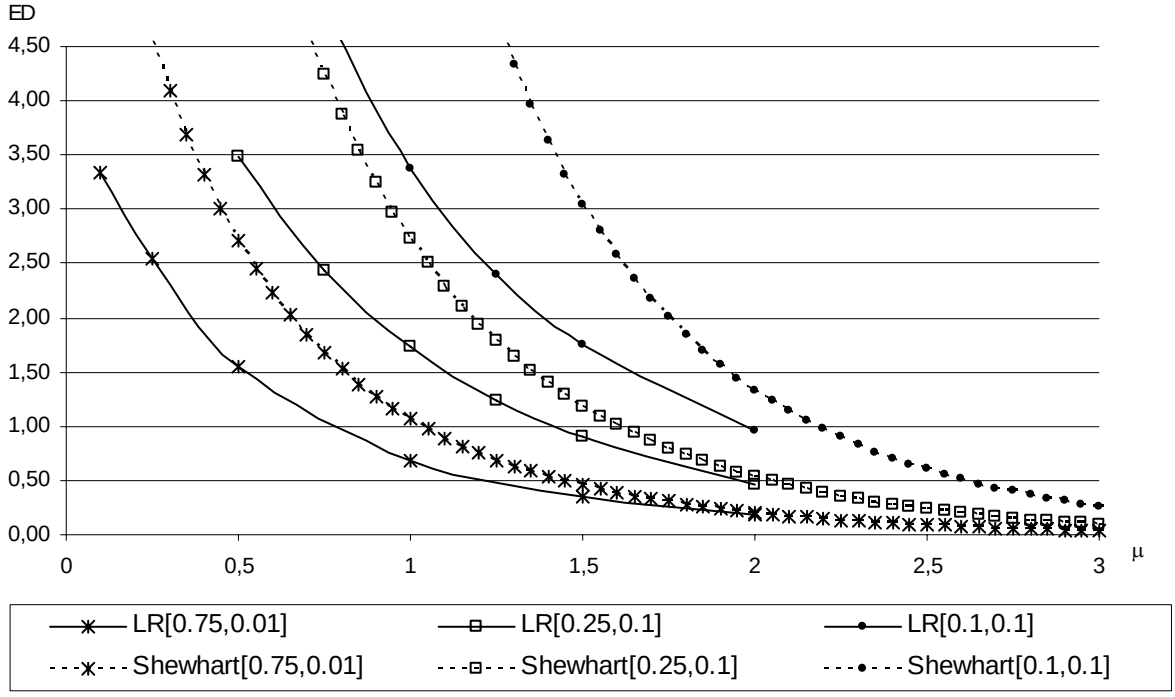

Figure 4. The expected delay as a function of : for the Shewhart and LR methods for different values of intensities and fixed false alarm probabilities. The format "Method(P(t_A<9), <)" is used for the legend to the curves.

Linear approximations of the LR method were suggested by Frisén (2003). Here we will study three variants and compare them to the EWMA method. All the approximations of the LR method are functions of the pivot parameter

$$\lambda^* = 1-\exp(-\mu^2/2)/(1-\nu),$$

which has a specific value as soon as : and < are specified. The method which is best for rare large changes is not the same as the method which is best for small long-term changes. Exact knowledge is rarely available. However, Frisén and Wessman (1999) demonstrated that the LR method is very robust with regard to the parameter <, of the geometrical distribution (no major effects were demonstrated for < less than 0.1).

Here we are specially interested in the methods as modifications of the EWMA method. We will express the methods with the unspecified parameter 8 and examine whether the value 8=8* actually is optimal or if it can be improved.

The first approximation, which is denoted LinLR is achieved by a Taylor approximation of the alarm statistic of the LR method. The LinLR method has a linear alarm statistic with the standardized weights

$$w_{LinLR}(s,t)=\left[\frac{1}{(1-\lambda)^t}-1\right]\frac{\lambda(1-\lambda)^s}{1-(1+s\lambda)(1-\lambda)^s}.$$

The weights are well approximated by exponential weights proportional to 1/(1-8) except possibly for small values of t. The alarm limit is

$$K_{LinLR}(s)=\frac{\lambda\left[\nu(1-L)(1-\lambda)^{s+1}-1+\lambda L\right]}{\nu(1-L)(1-\lambda)\{\mu-\mu(1+s\lambda)(1-\lambda)^{s}\}},$$

where the constant L is determined by false alarm properties. This alarm limit, as well as that of the normalized version of the EWMAa, tends to a constant when s increases. The LinLR method will give an alarm for

$$t_A=\min\{s:\sum_{t=1}^{s} w_{LinLR}(s,t)X(t) > K_{LinLR}(s)\} \qquad (5.4).$$

The exponential weights of the EWMA are appealing. When the weights of the LinLR method are approximated by exponential weights, and we keep the limit of the LinLR method, we have the EWLR method, which gives an alarm at

$$t_A=\min\{s: Zw_s > K_{EWLR}(s)\} \qquad (5.5).$$

A third approximation, EWlnLR, is achieved by a Taylor approximation of the logarithm of the alarm statistic of the LR method and, as a further approximation, exponential weights. An alarm is given at

$$t_A=\min\{s: Zw_s > K_{EWlnLR}(s)\} \qquad (5.6),$$

where

$$K_{EWlnLR}(s)=\frac{\lambda\left[1-(1-\lambda)^{s}\right]\left[L-\ln(1-(1-\lambda)^{s})\right]}{\mu\left[1-(1+s\lambda)(1-\lambda)^{s}\right]}.$$

The three approximations of the LR method can be seen as modifications of the EWMA method. The EWLR and EWlnLR methods keep the alarm statistic of the EWMA but modify the alarm limits. The modifications are negligible for large values of the decision time, s but have an influence by the effect for small values of s. The LinLR method modifies also the alarm statistic for small values of s. The modifications of the EWMA have the purpose to give smaller expected delays.

The CED is illustrated in Figure 3. For the EWlnLR method it can be seen that the dependency of CED on ϑ is very similar to that for LR. The other LR-approximations are not illustrated in the figure but are also very similar to the LR method.

The methods are compared in Figure 5 by the ED relative to that of the LR method for different values of 8 for some situations. In this figure we can find the details, discussed in Section 5.2.1, about how the value of 8 influences the ED for different methods. The diagrams are also used for derivation of the summarizing measures, used to compare the methods in Section 5.2.2.

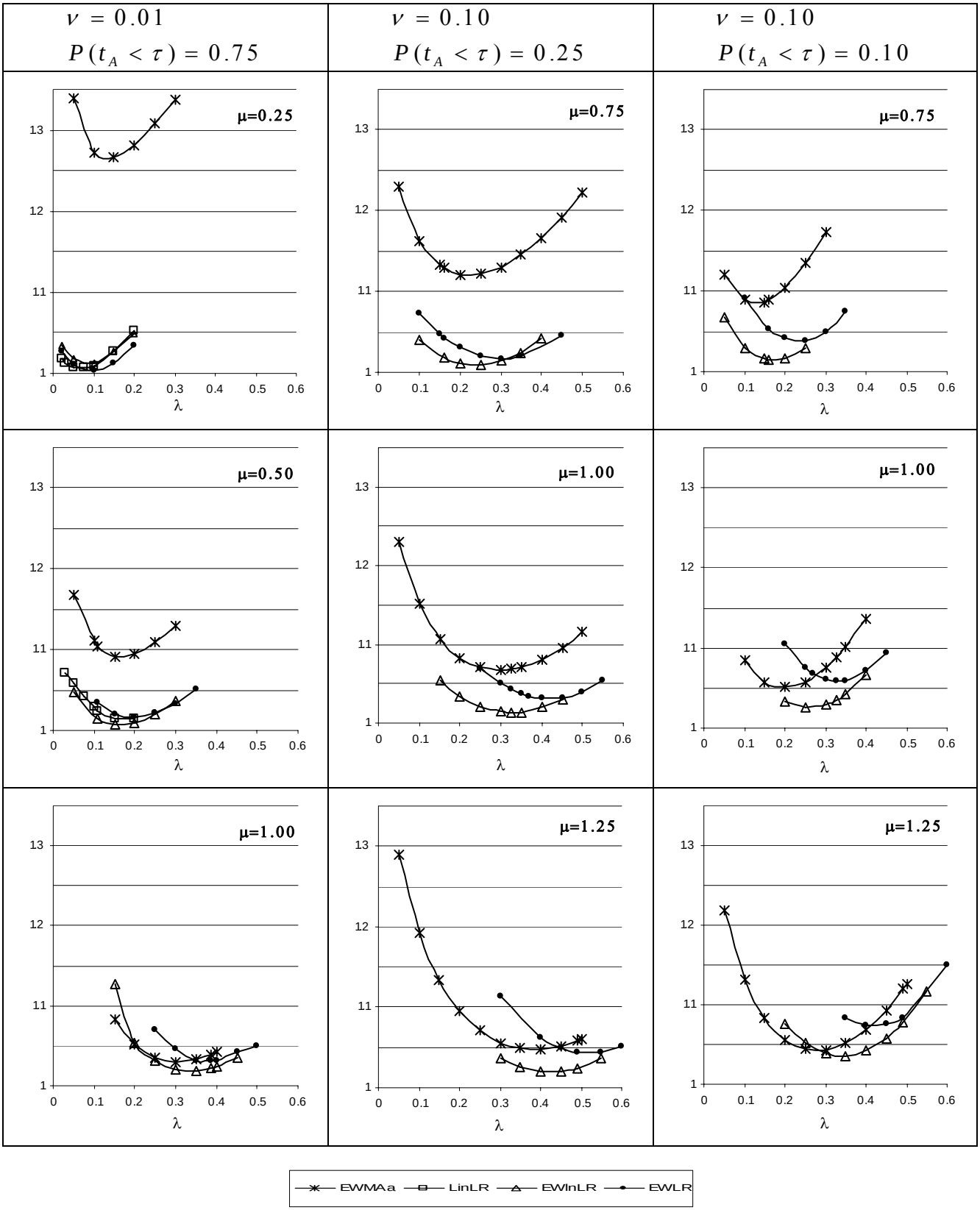

Figure 5. The expected delay relative to the LR method for fixed false alarm probabilities as functions of 8 for different variants of EWMA. The values of < and $P(t_A<9)$ are fixed for each column while the values of : are indicated in each diagram.

5.2.1 Optimal 8.

The choice of 8 is important and the search for the optimal value of 8 has been of great interest in literature, as discussed in Section 5.1.1 concerning the ARL criterion. In Figure 3 it is demonstrated how different values of 8 influence the ability of detection at different time points. Small values of 8 result in good ability to detect early changes while larger values are necessary for changes that occur later. This illustrates the importance of the choice of 8 also for the ED criterion.

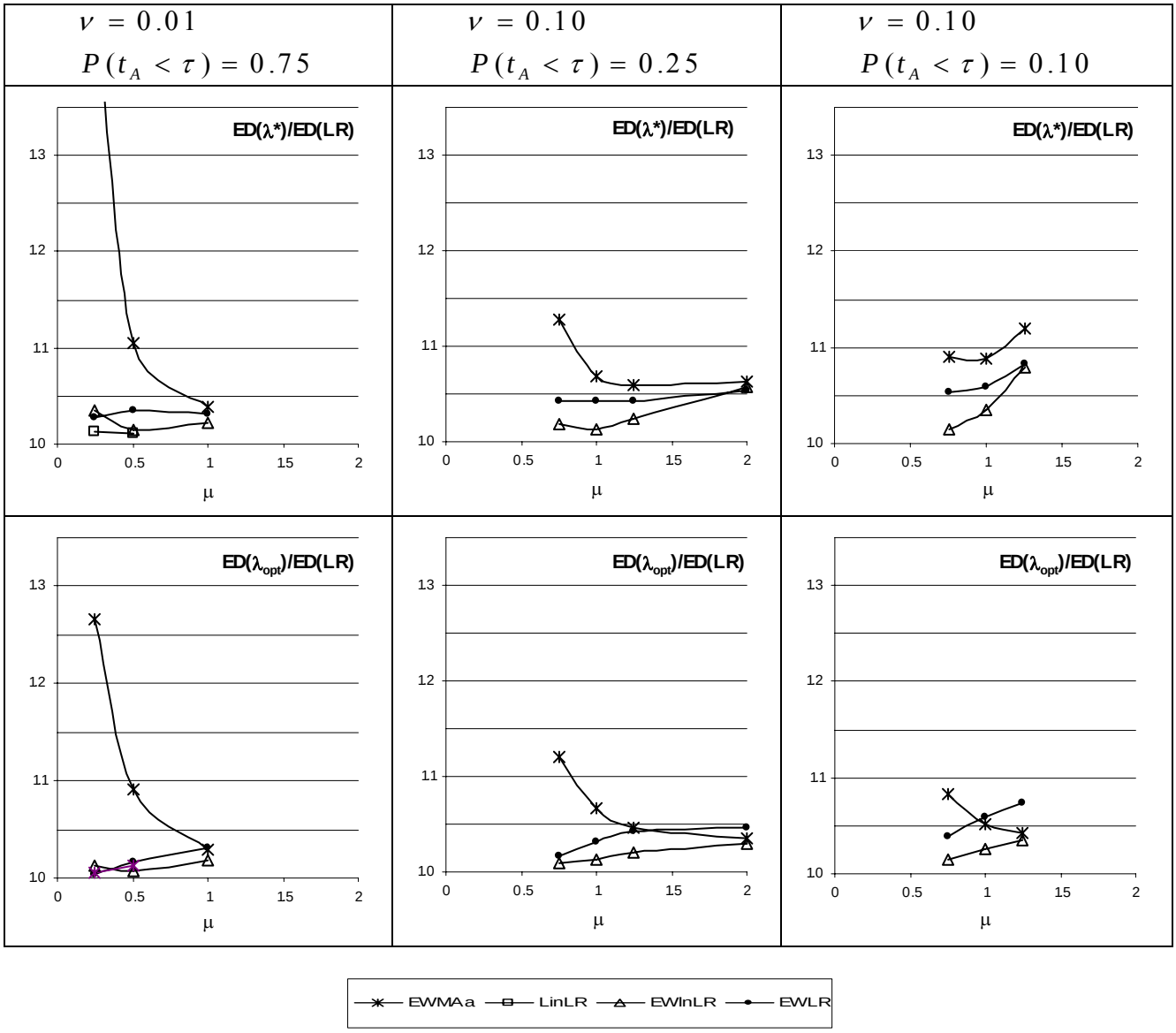

Figure 6. The expected delay, relative to LR, for fixed false alarm probabilities as functions of : for different variants of EWMA. The upper row is for 8= 8* while the bottom row is for the optimal value of 8. The values of < and $P(t_A<9)$ are fixed for each column.

The minima of ED with respect to 8 are rather flat as demonstrated in Figure 5. The minimal value of ED can thus be determined accurately, but there is an uncertainty on the corresponding 8. This also means that an exact value of the optimal value of 8 is not very important. However, an approximate value is important since the optimal value differs much between situations and the ED can be very large far away from the optimal value. The values

of 8 which give the minima of ED with respect to 8 are determined by quadratic interpolation for the neighboring observed values of 8. The optimal value of 8 increases with : and decreases with <.

The formula for 8* is suggested to achieve approximations of LR. Since this formula is a result of several approximations, it can only be expected to give rough approximations of the optimal values of 8 for different situations. However, as can be seen in the upper row of Figure 6, the formula results in very good properties for the modifications (within 5% of the value of the LR-method) for most cases studied. Comparison of the upper and lower row of Figure 6 reveals that the loss by the use of 8* instead of the exactly optimal value of 8 is small and less than 5% for the modifications, for all cases studied. This is also true for the EWMAa for large changes, which are the only cases where EWMAa should be used, as will be demonstrated below. Thus, the formula is useful for the choice of value of 8.

5.2.2 Comparison between methods.

In this section we compare how well the methods perform when the optimal value of 8 for each method and situation is used. These minimal ED values are presented in the bottom row of Figure 6. They are presented in relation to the minimal possible values which are achieved by the LR method.

All the methods studied perform considerably better than the Shewhart method, except for large changes, where the differences are less. The Shewhart method is not included in Figure 5 and 6 since the values are too high to be included in the scale suitable for the other methods.

The EWLR method is very good for small values of :. The minimal value of ED is within 8% of that for LR for all cases studied, and for most cases much smaller. The LinLR method could be expected to be better than the EWLR method since the approximation of the weights is avoided. This approximation could be expected to have most effect for small values of : where small values of 8 are used. However, the improvement compared to EWLR is slight. Thus, there is no indication of a need to modify the EWMA alarm statistic.

EWlnLR has the smallest values of ED, except for very small values of :, where all the modifications are very close to the LR method. The minimal value of ED is within 5% of LR for all cases studied.

The EWMAa method is very bad for small changes. However, for large values of : EWMAa is slightly better than the EWLR method and approximately equal to the EWlnLR method.

The three different cases studied have different false alarm properties. The largest ARL^0 is for < = 0.1 and $P(t_A<\vartheta) = 0.1$. The smallest is for < = 0.01 and $P(t_A<\vartheta) = 0.75$.

For the cases studied, all methods (with the optimal value of 8) have expected delays between the LR and Shewhart methods. The intervals between these two are illustrated in Figure 4.

5.3 Minimax

Examples of CED(ϑ) curves are given in Figure 3 where it is obvious that for EWMA the maximal value is not always for ϑ=1. Thus, ARL^1 does not reflect the maximal value. This means that for EWMA it is not possible to restrict attention to ARL^1 in order to determine the minimax optimal EWMA.. The maximal CED(ϑ) with respect to ϑ was first noted and then the minimal value with respect to 8 was determined. These minimax values of 8 and also the optimal values with respect to the ED criterion are given for some cases in Figure 7. It is clearly seen that the values of 8 by the two criteria are closely related. When the CED curves are flat, as they are for large values of λ, the minimum ED and the minimax values of CED

occur for the same value of 8. As expected, the difference between optimal values of λ with respect to these two criteria diminishes when μ increases as large values of λ are then optimal according to both these criteria. This is not at all the case for the formal ARL criterion where 8=0 is the solution as proved by Frisén (2003). Thus, the exact ARL optimality differs much from the minimax optimality, while the optimal value of 8 for the minimax and the ED optimality are related. The Shewhart method, which is very bad by the other criteria, is nearly as good as the LR method by the minimax criterion. The EWMA method with a value of 8 near 8* is the best of the examined methods by this criterion.

Most minimax studies in the literature are for a fixed ARL^0 and not for fixed $P(t_A<\vartheta)$ as here. This makes a difference as is seen in Figure 2. However, the slight difference in the present situation is in a direction to strengthen the above conclusions.

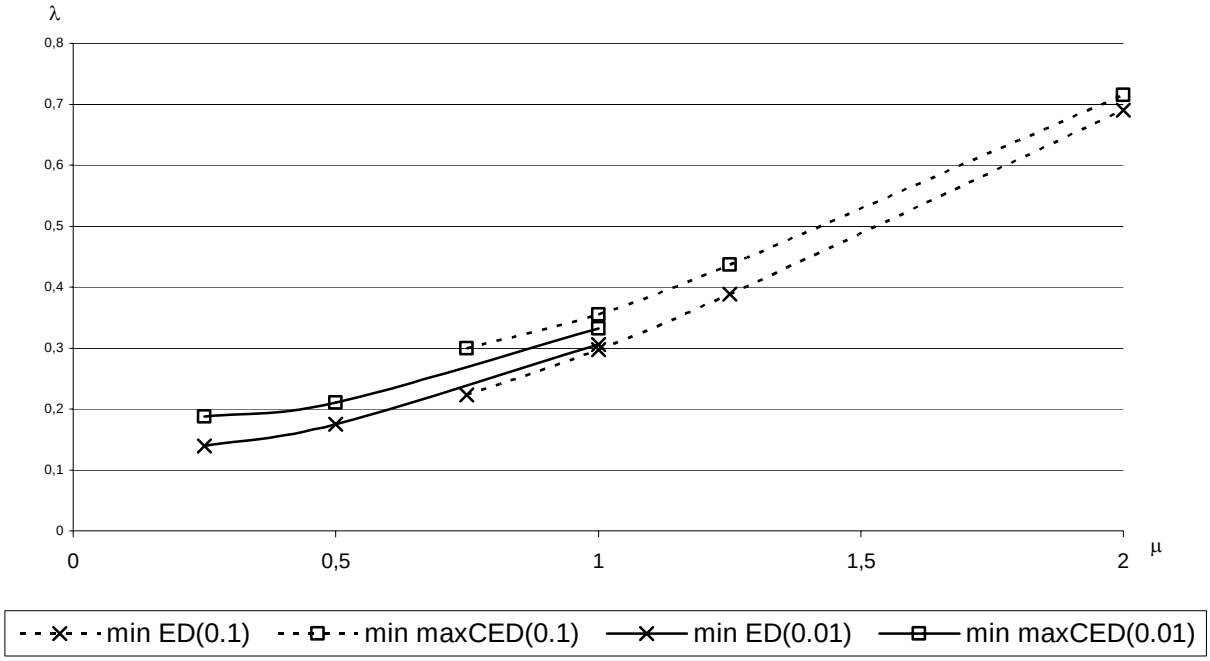

Figure 7. The optimal value of the parameter 8 of the EWMA method for the ED criterion and for the minimax criterion. The symbol "x" is used for the ED criterion and the symbol "ɢ" is used for the minimax criterion. The dotted curves are for the situation with $P(t_A<\vartheta) = 0.25$ and $< = 0.1$. The solid curves are for the situation with $P(t_A<\vartheta) = 0.75$ and $< = 0.01$.

6. DISCUSSION

There are two necessary conditions for a method to satisfy an optimality criterion. One is that the different observations at every fixed decision time, s, have optimal relative weights. The other one is that the alarm limit, as a function of s, is optimal. The alarm limit determines if the unavoidable false alarms will be spent early or late. For the conventional EWMA, the fact that both the weights and the alarm limit are completely determined by the parameter 8 hampers the possibility to optimize both at the same time. This is true for all optimality criteria. In earlier attempts to achieve an optimal EWMA the alarm limit has usually not been questioned and the optimal weighting has been determined given this limit as a function of s. The idea of a constant probability of the EWMAe statistic to exceed the alarm limit might seem appealing. However, for surveillance, this does not correspond to suggested optimality criteria. Besides, the asymptotic version completely destroys this property for small values of

8. Here we examine both the weights and the limits.

The result of this paper that the optimal value (zero) of 8, according to the ARL criterion, implies very bad properties demonstrates that the ARL criterion can be questioned as a formal criterion. Sometimes it is claimed that the ARL criterion which concerns $\vartheta=1$ contains enough information since no great differences between the properties for different values of ϑ exist. This is true for very large values of 8, but certainly not for all values of 8. For the detection of small changes (e.g., :=0.5), and the small values of 8 of EWMA suitable for that case, the conditional expected delay is strongly dependent on the time of the change. This is illustrated in Figure 3. In this figure it is also demonstrated that the ARL criterion differs much from the minimax criterion. We suggest that when immediate change ($\tau=1$) is the main interest, methods for sequential testing of hypotheses such as the SPRT-based LCUSUM method, and not EWMA, should be used.

For the ED criterion it is necessary with some knowledge of the parameter < in the distribution of the time of the change. In practice some knowledge should be available and should influence the choice of method. However, exact knowledge is rarely available. Frisén and Wessman (1999) demonstrated that the LR method is very robust with regard to <, (no major effects were demonstrated for < less than 0.1). The lack of need of a distribution for the change point is an advantage for the ARL- and minimax optimality. However, the robustness of the LR method makes this disadvantage of the ED criterion less important. Besides, it is not self evident that the possibility to optimize a method for a parameter should be seen as a disadvantage even though it is hard to specify which value is of most interest.

The necessity to find the (approximately) optimal value of the parameter 8 in order to get good properties is a problem. For the ED criterion the suggested formula for 8* gives good values of the expected delay for all methods and cases studied here. The only exception is for the EWMAa for very small changes, where the EWMAa should not be used anyway. When conclusions are based on simulations, it is always a question about the generality of the results. As indicated above the results on the usefulness of the approximations of the LR method could not be guaranteed for extreme situations but the cases studied represent a reasonable variation in values of : and < of interest for practical applications.

The EWMAa can be improved, with respect to the expected delay, by modifying the alarm limit to that of EWlnLR. The improvement is substantial for small changes, where it can be even slightly more improved by EWLR or LinLR. The LR method is of course best with respect to expected delay, but the approximations are nearly as good.

For large values of : there is no great difference between the EWMAa and the other methods. Thus, for large values of : there is no need for the modifications. However, the modifications have clearly less expected delay than the EWMAa for small values of :.

EWMAa with 8= 8* approximates the LR-method. Many of the CED-curves for this choice of 8 are rather flat and this adds to the explanation of the similarity between the minimax solution and the ED solution for EWMA. ARL optimality requires a small value of 8 and the steep slope of the CED-curves for these small values explains the lack of correspondence between the minimax- and ARL optimality for the EWMA method.

APPENDIX. ACCURACY OF THE RESULTS FROM THE SIMULATION STUDY.

When conclusions are based on simulations it is necessary to ensure that the number of replicates is enough. In many studies, including this one, different methods are evaluated with respect to some measure, θ under the restriction that another measure, P, has a specified value. In this appendix we describe the statistical technique used for calculation of confidence intervals of the results from the simulations and also exemplify the accuracy of the results displayed in the figures. The measure θ depends on the out of control run length distribution, which in turn makes θ a function, $\theta(L)$, of the constant L used in the alarm limit. The measure θ could for example be the ARL^1, CED or ED for different methods. The measure P, which is specified to P^* for comparability is here ARL^0 or $P(t_A < \tau)$. Also this measure depends on the value of the alarm limit. The aim is to study the accuracy of the estimates of θ for $P = P^*$. By analyzing the accuracy in each step used to produce the results, and combining these, we get conservative confidence intervals for the points in the figures.

We start by studying the accuracy of the value of L. Denote by L^* the value of L such that $P(L^*) = P^*$ and by L' the value of L estimated by our procedure and for which θ is estimated. The procedure used here to estimate L^* is to choose $L_1, L_2, ..., L_n$ and use simulations for each value of L to estimate the values of $P(L_1), P(L_2), ..., P(L_n)$. We approximate $P(L)$ with a linear function locally, $\hat{P}(L) = \hat{a}_1 + \hat{b}_1 L$, and choose L' accordingly to give $\hat{P}(L') = P^*$. A confidence interval CI^1 for L^* (and thus also for $L^* - L'$) can be constructed as consisting of those values of L which would not be rejected by the test of $H_0 : P(L) = P^*$. The very large number of replicates ensures that the normal distribution can be used.

A confidence interval CI^2 for $\theta(L')$ can be constructed using simulations for the chosen value L'. Also in this case we can use a normal approximation due to the large number of replicates.

The last link is to determine how influential the error in L is with respect to the value of θ. Estimates of $\theta(L)$ for some values $L_1, L_2, ..., L_{n'}$ of L around L' are achieved by special simulations for the purpose of determining the accuracy. We approximate $\theta(L)$ locally by a linear function $\theta(L^*) = \theta(L') + b_2(L^* - L')$. A confidence interval CI^3 for b_2 in the regression can be constructed.

The confidence intervals from the three steps described above can then be combined to form a confidence interval for $\theta(L^*)$. Let the confidence interval CI^1 for $L' - L^*$ be of confidence $(1-\alpha_1)$, the confidence interval CI^2 for $\theta(L')$ be of confidence $(1-\alpha_2)$ and the confidence interval CI^3, for b_2 be of confidence $(1-\alpha_3)$. Then, we can combine these intervals to construct a confidence interval of confidence at least $(1-\alpha_1)(1-\alpha_2)(1-\alpha_3)$ for $\theta(L^*)$ by taking $\min\{a + b \cdot c;\ a \in CI^2,\ b \in CI^3,\ c \in CI^1\}$ to be the lower limit and $\max\{a + b \cdot c;\ a \in CI^2,\ b \in CI^3,\ c \in CI^1\}$ to be the upper. We choose $1-\alpha_1 = 1-\alpha_2 = 1-\alpha_3 = 0.983$ which results in a confidence interval for $\theta(L^*)$ of confidence at least 95%. Confidence intervals for $\theta(L^*)$ constructed in this way for some cases studied in the paper will now be exemplified.

Accuracy of the results in Figure 1: Figure 1 shows ARL^1 as a function of λ for the EWMAa method when ARL^0=100. We exemplify the accuracy in Table A.1. For the determination of the alarm limit L' we have used 500.000 replicates for 25 neighbouring

values of L. The confidence intervals for b_2 are based on 10 values of L with 10.000.000 replicates each. The estimates of $\theta(L')$ are based on 1.000.000 replicates. The conclusion that λ should approach zero in order to minimize ARL^1 when ARL^0=100 is thus well supported. The width of the confidence interval is short compared with the thickness of the lines in Figure 1.

Table A.1. Examples of the components for the confidence intervals for different values of Figure 1.

	Confidence interval for L^*-L'		*Confidence interval for* b_2		*Confidence interval for* $ARL^1(L')$		*Confidence interval for* $ARL^1(L^*)$	
λ	*lower*	*upper*	*lower*	*upper*	*lower*	*upper*	*lower*	*upper*
0.001	-0.00022	0.00029	18.75	21.73	2.00	2.02	2.00	2.02
0.010	-0.00044	0.00045	7.07	8.08	4.62	4.64	4.61	4.64
0.100	-0.00055	0.00058	2.42	4.49	5.64	5.66	5.64	5.67

Accuracy of the results in Figure 3: Figure 3 shows CED as a function of τ for the case when μ= 0.25, ν=0.01 and $P(t_A < \tau)$=0.75. As an example we calculate a 95% confidence interval for $CED(1)$ for the LR method. For this case the determination of the alarm limit L' was based on 500.000 replicates for 24 neighbouring values of L. This resulted in a confidence interval for L^*-L' of (-0.06, 0.04). Again, the confidence interval for b_2 was based on 10 values of L with 10.000.000 replicates each. This results in a confidence interval of (0.39, 0.44) for b_2. A confidence interval of (17.23, 17.28) was constructed using 1.000.000 replicates for the estimate of $CED(1)$ using the value L' in the alarm limit. The width of the resulting 95% confidence interval for $CED(L^*)$, (17.2067, 17.2887), is small compared with the line width in Figure 3.

Accuracy of the results in Figure 4: For the same case as exemplified for Figure 3 we can construct a conservative confidence interval for

$$ED(L^*) = \sum_{t=1}^{\infty} P(\tau = t) ED(t, L^*)$$

by using the confidence intervals calculated for ED(t) by the technique given for CED(t) above. Under the assumption that all the steps given above result in an approximately normally distributed estimator we can estimate the variance for the estimator of ED(t). The variance of the weighted sum ED(L*) has less variance than the component with the largest variance. The simulations indicate that this is the case for τ=1. Since $ED(1) = CED(1)$, we can use the results from the example above and we have a 95% confidence interval for the ED for the LR method as (2.4976, 2.5796). The width of this confidence interval is small compared with the line width in Figure 4.

The confidence intervals for the results in this paper are small enough using the chosen number of replicates to ensure that the results are reliable. In most cases the confidence intervals are shorter than the thickness of the lines in the figures.

ACKNOWLEDGEMENTS

We thank Professor Hans van Houwelingen, the Associate Editor and a referee for constructive comments. This work has been partly supported by the Swedish Council for Research in the Humanities and Social Sciences, Grant F0473/2000.

REFERENCES

Andersson, E., Bock, D. and Frisén, M. (2004) Detection of Turning Points in Business Cycles, *Journal of Business Cycle Measurement and Analysis,* 1: 93-108.

Beibel, M. (2000) A Note on Sequential Detection with Exponential Penalty for the Dealy, *The Annals of Statistics,* 28: 1696-1701.

Brook, D. and Evans, D. A. (1972) An Approach to the Probability Distribution of Cusum Run Length, *Biometrika,* 59: 539-549.

Carlyle, W. M., Montgomery, D. C. and Runger, G. C. (2000) Optimization Problems and Methods in Quality Control and Improvement, *Journal of Quality Technology,* 32: 1-19.

Champ, C. W., Woodall, W. H. and Mohsen, H. (1991) A Generalized Quality Control Procedure, *Statistics & Probability Letters,* 11: 211-218.

Chan, L. K. and Zhang, J. (2000) Some Issues in the Design of Ewma Charts, *Communications in Statistics. Simulation and Computation,* 29: 207-217.

Crowder, S. (1987) A Simple Method for Studying Run-Length Distributions of Exponentially Weighted Moving Average Charts, *Technometrics,* 29: 401-407.

Crowder, S. V. (1989) Design of Exponentially Weighted Moving Average Schemes, *Journal of Quality Technology,* 21: 155-162.

Domangue, R. and Patch, S. C. (1991) Some Omnibus Exponentially Weighted Moving Average Statistical Process Monitoring Schemes, *Technometrics,* 33: 299-313.

Frisén, M. (1992) Evaluations of Methods for Statistical Surveillance, *Statistics in Medicine,* 11: 1489-1502.

Frisén, M. (2003) Statistical Surveillance. Optimality and Methods., *International Statistical Review,* 71: 403-434.

Frisén, M. and de Maré, J. (1991) Optimal Surveillance, *Biometrika,* 78: 271-80.

Frisén, M. and Wessman, P. (1999) Evaluations of Likelihood Ratio Methods for Surveillance. Differences and Robustness., *Communications in Statistics. Simulation and Computation,* 28: 597-622.

Gan, F. F. (1995) Joint Monitoring of Process Mean and Variance Using Exponentially Weighted Moving Average Control Charts, *Technometrics,* 37: 446-453.

Gan, F. F. (1998) Designs of One- and Two-Sided Exponential Ewma Charts, *Journal of Quality Technology,* 30: 55-69.

Girshick, M. A. and Rubin, H. (1952) A Bayes Approach to a Quality Control Model., *Annals of Mathematical Statistics,* 23: 114-125.

Gordon, L. and Pollak, M. (1995) A Robust Surveillance Scheme for Stochastically Ordered Alternatives, *The Annals of Statistics,* 23: 1350-1375.

Gut, A. and Steinebach, J. (2004) Ewma Charts for Detecting a Change-Point in the Drift of a Stochastic Process., *Sequential Analysis,* 23: 195-237.

Han, D. and Tsung, F. (2004) A Generalized Ewma Control Chart and Its Comparison with the Optimal Ewma, Cusum and Glr Schemes, *The Annals of Statistics,* 32: 316-339.

Kemp, K. W. (1961) The Average Run Length of the Cumulative Sum Chart When a V-Mask Is Used, *Journal of the Royal Statistical Society B* 23: 149-153.

Kenett, R. and Pollak, M. (1996) Data-Analytic Aspects of the Shiryaev-Roberts Control Chart: Surveillance of a Non-Homogeneous Poisson Process, *Journal of Applied Statistics,* 23: 125-137.

Knoth, S. (2004) The Art of Evaluating Monitoring Schemes - How to Measure the Performance of Control Charts? In *VIIIth International workshop on Intelligent Statistical Quality Control.*Warsaw, Poland.

Knoth, S. and Schmid, W. (2002) Monitoring the Mean and the Variance of a Stationary Process, *Statistica Neerlandica,* 56: 77-100.

Kramer, H. and Schmid, W. (1997) Ewma Charts for Multivariate Time Series, *Sequential Analysis,* 16: 131-154.

Lai, T. L. (1995) Sequential Changepoint Detection in Quality-Control and Dynamical Systems, *Journal of the Royal Statistical Society B,* 57: 613-658.

Lowry, C. A. and Montgomery, D. C. (1995) A Review of Multivariate Control Charts, *IIE Transactions,* 27: 800-810.

Lowry, C. A., Woodall, W. H., Champ, C. W. and Rigdon, S. E. (1992) A Multivariate Exponentially Weighted Moving Average Control Chart., *Technometrics,* 34: 46-53.
Lu, C. W. and Reynolds Jr, M. R. (1999) Ewma Control Charts for Monitoring the Mean of Autocorrelated Processes, *Journal of Quality Technology,* 31: 166-188.
Lucas, J. M. and Saccucci, M. S. (1990) Exponentially Weighted Moving Average Control Schemes: Properties and Enhancements, *Technometrics,* 32: 1-12.
Margavio, T. M., Conerly, M. D., Woodall, W. H. and Drake, L. G. (1995) Alarm Rates for Quality-Control Charts, *Statistics & Probability Letters,* 24: 219-224.
Moustakides, G. V. (1986) Optimal Stopping Times for Detecting Changes in Distributions, *The Annals of Statistics,* 14: 1379-1387.
Ng, C. H. and Case, K. E. (1989) Development and Evaluation of Control Charts Using Exponentially Weighted Moving Averages, *Journal of Quality Technology,* 21: 242-250.
Page, E. S. (1954) Continuous Inspection Schemes, *Biometrika,* 41: 100-114.
Pettersson, M. (1998) Monitoring a Freshwater Fish Population: Statistical Surveillance of Biodiversity., *Environmetrics,* 9: 139-150.
Pollak, M. (1985) Optimal Detection of a Change in Distribution, *The Annals of Statistics,* 13: 206-227.
Poor, V. H. (1998) Quickest Detection with Exponential Penalty for Delay, *The Annals of Statistics,* 26: 2179-2205.
Prabhu, S. S. and Runger, G. C. (1997) Designing a Multivariate Ewma Control Chart, *Journal of Quality Technology,* 29: 8-15.
Roberts, S. W. (1959) Control Chart Tests Based on Geometric Moving Averages, *Technometrics,* 1: 239-250.
Robinson, P. B. and Ho, T. Y. (1978) Average Run Lengths of Geometric Moving Average Charts by Numerical Methods, *Technometrics,* 20: 85-93.
Rosolowski, M. and Schmid, W. (2003) Ewma Charts for Monitoring the Mean and the Autocovariances of Stationary Gaussian Processes, *Sequential Analysis,* 22: 257-285.
Schipper, S. and Schmid, W. (2001) Sequential Methods for Detecting Changes in the Variance of Economic Time Series, *Sequential Analysis,* 20: 235-262.
Schmid, W. and Schöne, A. (1997) Some Properties of the Ewma Control Chart in the Presence of Autocorrelation, *The Annals of Statistics,* 25: 1277-1283.
Schmid, W. and Tzotchev, D. (2004) Statistical Surveillance of the Parameters of a One-Factor Cox-Ingersoll-Ross Model, *Sequential Analysis,* 23: 379-412.
Severin, T. and Schmid, W. (1998) Statistical Process Control and Its Application in Finance, in *Risk Measurement, Econometrics and Neural Networks*, Bol, G., Nakhaeizadeh, G. and Vollmer, C.-H. eds. pp. 83-104, Heidelberg: Physica Verlag.
Shewhart, W. A. (1931) *Economic Control of Quality of Manufactured Product,* London: MacMillan and Co.
Shiryaev, A. N. (1963) On Optimum Methods in Quickest Detection Problems, *Theory of Probability and its Applications,* 8: 22-46.
Siegmund, D. (1985) *Sequential Analysis. Tests and Confidence Intervals.,* New York: Springer.
Sonesson, C. (2003) Evaluations of Some Exponentially Weighted Moving Average Methods, *Journal of Applied Statistics,* 30: 1115-1133.
Sonesson, C. and Bock, D. (2003) A Review and Discussion of Prospective Statistical Surveillance in Public Health, *Journal of the Royal Statistical Society A,* 166: 5-21.
Srivastava, M. S. and Wu, Y. (1993) Comparison of Ewma, Cusum and Shiryaev-Roberts Procedures for Detecting a Shift in the Mean., *The Annals of Statistics,* 21: 645-670.
Srivastava, M. S. and Wu, Y. (1997) Evaluation of Optimum Weights and Average Run Lengths in Ewma Control Schemes, *Communications in Statistics. Theory and Methods,* 26: 1253-1267.
Steiner, S. H. (1999) Ewma Control Charts with Time-Varying Control Limits and Fast Initial Response, *Journal of Quality Technology,* 31: 75-86.
Tsui, K. L. and Woodall, W. H. (1993) Multivariate Control Charts Based on Loss Functions, *Sequential Analysis,* 12: 79-92.
Waldmann, K.-H. (1986) Bounds for the Distribution of the Run Lenght of Geometric Moving Average Charts, *Applied Statistics,* 35: 151-158.
VanBrackle, L. and Williamson, G. D. (1999) A Study of the Average Run Length Characteristics of the National Notifiable Diseases Surveillance System, *Statistics in Medicine,* 18: 3309-3319.
Williamson, G. and Hudson, G. (1999) A Monitoring System for Detecting Aberrations in Public Health Surveillance Reports, *Statistics in Medicine,* 18: 3283-3298.
Woodall, W. H. and Montgomery, D. C. (1999) Research Issues and Ideas in Statistical Process Control, *Journal of Quality Technology,* 31: 376-386.
Wu, Y. (1994) Design of Control Charts for Detecting the Change Point, in *Change-Point Problems*, Carlstein,

E., Muller, H.-G. and Siegmund, D. eds. pp. 330-345, Hayward, CA: IMS.

Yakir, B. (1997) A Note on Optimal Detection of a Change in Distribution, *The Annals of Statistics,* 25: 2117-2126.

Yakir, B., Krieger, A. M. and Pollak, M. (1999) Detecting a Change in Regression: First-Order Optimality, *The Annals of Statistics,* 27: 1896-1913.

Yashchin, E. (1993) Statistical Control Schemes - Methods, Applications and Generalizations, *International Statistical Review,* 61: 41-66.

Yeh, A. B., Lin, D. K. J., Zhou, H. and Venkataramani, C. (2003) A Multivariate Exponentially Weighted Moving Average Control Chart for Monitoring Process Variability, *Journal of Applied Statistics,* 30: 507-536.