Research Report
Department of Statistics
Göteborg University
Sweden

Characterization of methods for surveillance by optimality

Marianne Frisén

Research Report 1999:13
ISSN 0349-8034

Mailing address:	Fax	Phone	Home Page:
Dept of Statistics	Nat: 031-773 12 74	Nat: 031-773 10 00	http://www.handels.gu.se/stat
P.O. Box 660	Int: +46 31 773 12 74	Int: +46 31 773 10 00	
SE 405 30 Göteborg			
Sweden			

Characterization of Methods for Surveillance by Optimality

Marianne Frisén

Abstract. Different criteria of optimality are discussed. The shortcomings of some criteria of optimality are demonstrated by their implications. The correspondences between some criteria of optimality and some methods are examined. The situations and parameter values for which some commonly used methods have certain optimality properties are thus illuminated. A linear approximation of the full likelihood ratio method, which satisfies several criteria of optimality, is presented. This linear approximation is used for comparisons with the exponentially weighted moving average method. Via these comparisons it is possible to illuminate the influence of different criteria of optimality on the "optimal" parameter of a method. A uniform presentation of methods, by expressions of likelihood ratios, facilitates the comparisons between methods.

Key words andphrases: Quality control, Warning system, Monitoring, Control chart, Likelihood ratio, EWMA, CUSUM, Moving average, Shewhart.

Marianne Frisén is Professor, Department of Statistics, Göteborg University, Box 660, SE-40530 Göteborg, Sweden.

1. INTRODUCTION

There is a need of continual observation of time series, with the goal of detecting an important change in the underlying process as soon as possible after it has occurred. Surveillance, statistical process control, monitoring and change-point detection are different names for this. The timeliness and also the simplicity of decisions is taken into account in the vast literature on quality control charts. Also, the literature on stopping rules is useful and relevant here. The inferential problems involved are important for the applications and interesting from a theoretical view since they are linking together different areas of statistical theory.

Some broad surveys and bibliographies are found in, e.g. Zacks (1983), Basseville and Benveniste (1986), Vardeman and Cornell (1987) and Lai (1995). In the survey by Kolmogorov et al (1990) and the collection of papers edited by Telksnys (1986) the early results on optimal stopping rules by Kolmogorov and Shiryaev are reported and used in further research. Also the book by Brodsky and Darkhovsky (1993) on nonparametric methods on change-point problems is in the same spirit. This literature treats both the case of a fixed period and the case of continual observation. The survey by James et al (1987) treats the fixed period case.

In recent years there have been a growing number of papers in economics, medicine, environmental control and other areas dealing with the need of methods for surveillance. Applications in medicine are described in, e.g. the special issue (no. 3, 1989) of "Statistics in Medicine" and by Frisén (1992). 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 Frisén (1999).

In some applications the whole process will be stopped as soon as an alarm occurs. This is the case at the surveillance of the foetal heart during labour. At an alarm the baby is rescued by Caesarean section. We call this active surveillance. In the case when our actions at an earlier time point do not affect the distributions we say that we have passive surveillance. This can be the case in flood warning systems when alarms do not affect the height of the flood wave. Most of the discussion concerns active surveillance, but the differences with respect to stochastic properties between the active and passive surveillance will be pointed out.

There are a large number of papers which claim to give the optimal method of surveillance. However, the optimality criteria differ in important aspects. Most commonly used methods are optimal in some respect. Here, the aim is to make a characterization of the methods by the optimality properties they have. In Table 1 some schematic characterizations are given. The motivations will be given in the text.

In Section 2 some notations are given. Also, specification is made of the most commonly discussed case of a shift in the mean of a normal distribution. This simple case is used throughout this paper, in order to be specific and concentrate on principle properties, even though some results are valid also for other cases. In Section 3 some criteria of optimality are described and analyzed. In Section 4 methods derived from optimality criteria as well as some commonly used methods are described. The two groups of methods are compared in order to characterize the commonly used methods by their optimality properties. In Section 4.1 the full likelihood ratio method, LR, which fulfills important optimality criteria is described. In Section 4.2 a linear approximation, LLR, of the LR method is derived. This approximation is used in Section 4.3 to determine the optimal value of the parameter of the exponential moving average method, EWMA, and also to discuss for which situation EWMA will be a good method. Different variants of CUSUM methods are analyzed in Section 4.4 with respect to their optimality. In Section 5 some concluding remarks are given.

2. NOTATIONS AND SPECIFICATIONS

The variable under surveillance is denoted by $X = \{X(t): t = 1, 2, ...\}$, where $X(t)$ is the observation made at time t. This observation may be an average or some other derived statistic. For the case of surveillance of the foetal heart rate, described in Frisén (1992), $X(t)$ is a recursive residual of a measure of variation. The random process that determines the state of the system is identified by $\mu(t)$, $t = 1, 2, ...$

The critical event of interest at decision time s is denoted $C(s)$. As in most literature on quality control, the case of a shift in the mean of a Gaussian random variable from an acceptable value μ^0 (say zero) to an unacceptable value μ^1 is considered. Only one-sided procedures are considered here. It is assumed that if a change in the process occurs, the level suddenly moves to another constant level, $\mu^1>\mu^0$, and remains on this new level. That is $\mu(t) = \mu^0$ for $t= 1, ..., \tau-1$ and $\mu(t) = \mu^1$ for $t= \tau, \tau+1, ...$ For each decision time s, $s=1, 2, ...$ we want to discriminate between the two events $C(s) = \{\tau\leq s\}$ and $D(s) = \{\tau>s\}$. $C(s)$ implies $\mu(s)=\mu^1$ and $D(s)$ implies $\mu(s)=\mu^0$.

We will consider different ways of constructing alarm sets A(s) with the property that, when $X_s = \{X(t): t \leq s\}$ is a subset of A(s), there is an indication that the event C(s) has occurred. The time of the first alarm is $t_A = \min\{s: X_s \subset A(s)\}$.

Here μ^0 and μ^1 are regarded as known values and the time point τ where the critical event occurs is regarded as a generalized random variable with the probabilities π_t=P(τ=t) and with the probability, π_n that no change ever occurs

$$\pi_n = 1 - \sum_{t=1}^{t=\infty} \pi_t.$$

The intensity, ν_t ,of a change is

$$\nu_t = P(\tau = t | \tau \geq t).$$

The aim is to discriminate between the states of the system at each decision time s, s=1, 2, ... by the set of observations $X_s = \{X(s): t \leq s\}$ under the assumption that X(1) - μ(1), X(2) - μ(2), ... are independent normally distributed random variables with mean zero and with the same known standard deviation σ. For clarity, and to get shorter formulas standardization to μ^0=0 and σ=1 is used and the size of the shift after standardization is denoted by μ. The case μ>0 is described here. The case μ<0 is treated in the same way. Two-sided procedures can be constructed from a combination of the one-sided ones but optimality will only be discussed for the one-sided case.

3. OPTIMALITY CRITERIA

The performance of a method for surveillance depends on the time τ of the change. Alarm probabilities will typically not be the same for early changes as for late changes. Sometimes it is appropriate to express the measure of the performance as a function of τ, as in Frisén (1992), Frisén and Wessman (1999) and the self instructive free computer program by Frisén and Cassel (1994). However, sometimes a precise criterion of optimality is needed. In order to obtain a measure, which is independent of the value of τ, several approaches have been used:

1. In the literature on quality control it is often assumed that the change occurred at the same time as the surveillance started, that is τ=1. This approach is discussed in Section 3.1 on ARL.
2. In the literature on statistical theory it is often assumed that the surveillance has been started a very long time before a possible change (e.g. Lindgren 1985, Pollak and Siegmund 1991, Srivastava and Wu 1993).
3. A probability distribution of τ is considered and an averaging measure with respect to this distribution is used. Error probabilities are described in Section 3.2 and expectations and utilities are described in Section 3.3.
4. A minimax criterion with the worst possible value of τ is used (Section 3.4).

3.1 ARL

A measure that is often used in quality control is the average of the run length until the first alarm. See, e.g. Wetherill and Brown (1990). This idea was already suggested by Page (1954). The average run length until an alarm, when there is no change in the system under surveillance, is denoted ARL^0. The average run length until detection of a true change (that occurred at the same time as the surveillance started) is denoted ARL^1. The part of the definition in the parenthesis is seldom spelled out, but seems to be generally used in the literature on quality control.

Optimality can be defined as minimal ARL^1 for fixed ARL^0. This criterion will shortly be called "the criterion of minimal ARL^1 ". This criterion is usually used in the literature on quality control and is sometimes used also in more general statistical literature. A consequence of this criterion, which makes it unsuitable for many applications, will now be demonstrated. Some might consider the consequence self-evident, but since it is in contradiction with much of the literature a detailed proof is given.

Proposition 3.1.1. There exist values c_s such that a surveillance system with alarm at

$$t_A = \min\{s: \sum_{t=1}^{s} X(t) > c_s\}$$

gives the minimal ARL^1 for fixed ARL^0 for the case specified in Section 2.

Proof. First, some properties of surveillance systems based on t_A above are derived. Let, in this proof, C(s) = {τ = 1} and D(s) = {τ = ∞} with the notation that τ = ∞ is the event that no change ever happens. As a technical tool, passive surveillance with the alarm set denoted by $_pA(.)$, is used to start with. Then, with the specifications in Section 2, the likelihood ratio method (Section 4.1) has the alarm set

$$_pA(s) = \{X_s\text{: } f_{Xs}(x_s \mid C) / f_{Xs}(x_s \mid D) > a_s\}$$

$$= \{X_s\text{: } \exp\left\{\frac{1}{2}\mu^2\right\} \exp\{\mu \sum_{t=1}^{s} X(t)\} > b_s\} = \{X_s\text{: } \sum_{t=1}^{s} X(t) > c_s\}$$

where a_s, b_s and c_s are constants.

At active surveillance, where the surveillance is stopped at the first alarm, it follows from Theorem 3.1 in Frisén and de Maré (1991) that

$$_aA(s) = {}_pA(s) \cap {}_pA^c_{s-1}$$

where $_aA(.)$ is the alarm set at active surveillance, $A^c_{s-1} = A^c(1) \cap A^c(2) \cap \ldots \cap A^c(s-1)$ and $A^c(.)$ is the compliment of A(.). We have that

$$_aA(s) = \{X_s\text{: } \sum_{t=1}^{s} X(t) > c_s\} \cap \{X_s\text{: } \sum_{t=1}^{r} X(t) \leq c_r \text{ , } r=1,\ldots s-1\}$$

$$= \{X_s\text{: } s = \min\{i\text{: } \sum_{t=1}^{i} X(t) > c_i\} \}.$$

Thus, the monitoring system in the proposition is identical to that of a certain known likelihood-based one. Theorem 2.1 in Frisén and de Maré (1991) (see also Section 4.2 here and de Maré (1980)) states that the likelihood ratio method has the property that for each decision time s it gives the maximal probability of alarm P(A(s)| C(s)) for a fixed false alarm probability P(A(s)| D(s)).

Now, we use the properties derived above to examine the optimality condition. Both ARL^1 and ARL^0 are expected values under the condition that μ(t) has the same value for all t. The condition $\mu(t) \equiv 0$ is equivalent to the condition that no change ever happens, that is τ = ∞, with our notation.

$$ARL^0 = E(t_A|\, \mu(t)\equiv 0) =$$

$$= \sum_{t=1}^{\infty} t\ P(t_A=t|\ \tau=\infty)\ = \sum_{t=1}^{\infty} t\ P({}_aA(t)|\ D(t)).$$

$$ARL^1 = E(t_A|\, \mu(t)=\mu) =$$

$$= \sum_{t=1}^{\infty} t\ P(t_A=t|\ \tau=1)\ =\ \sum_{t=1}^{\infty} t\ P({}_aA(t)|\ C(t)).$$

The constants, c_s, can be chosen to match any given set of false alarm probabilities and thus any given ARL^0. For these fixed values of c_s the likelihood ratio method with

$$t_A\ =\ \min\{s:\ \sum_{t=1}^{s} X(t) > c_s\}$$

gives maximal detection probability for the fixed value of $P({}_aA(s)|\, D(s))$ for all s and thus minimal ARL^1. □

Thus, only methods which give equal weight to all observations satisfy the optimality criterion of minimal ARL^1 for fixed ARL^0. Such methods are not very often used in quality control. Examples of such methods are the simple CUSUM variants described in Section 4.4, where also the drawbacks of these methods are discussed. The Proposition 3.1.1 thus demonstrates that the optimality criterion could be questioned. There are a great number of papers in the literature on quality control where the aim is to find the parameters of a method which is "optimal" in the sense that the ARL^1 is minimized for a fixed ARL^0.

In applications where the criterion of minimal ARL^1 is the proper one (in spite of the drawbacks given above) it is not sufficient to know the alarm statistic for each decision time s. You would also have to determine the alarm limit c_s for this statistic for each s.

Proposition 3.1.2. The surveillance system with alarm at

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

where L is a constant, gives the minimal ARL^1 in the class of methods with the same total false alarm probability.

Proof. In Frisén and de Maré (1991) it was demonstrated that the sequential probability ratio test (SPRT) of D= $\{\tau>s\}$ against C=$\{\tau=t\}$ without an acceptance limit and with a constant rejection limit will give the shortest expected delay E(t_A-τ |C) for a given total false alarm probability. With the conditions of Section 2 and with t=1 the SPRT will be

$$\prod_{t=1}^{s} \exp[-\frac{1}{2}(\{x(t)-\mu\}^2-\{x(t)\}^2)]>G \Rightarrow \sum_{t=1}^{s} x(t) > L + s(\mu)/2$$

where G and L are constants. The expected delay E(t_A-τ |C), which is minimal, is equal to ARL^1 - 1, since t=1. Thus, also ARL^1 is minimal.□

For comparison with other linear methods it is convenient to use weights for the observations, which have the sum equal to one. With such weights the method which gives the minimum ARL^1 for a fixed false alarm probability, but which does not have a finite ARL^0, has the alarm condition

$$\sum_{t=1}^{s} x(t)/s > \mu/2 + L/s.$$

This method will be further discussed in Section 4.4 as a CUSUM method and is there named LCUSUM. By choosing L small enough in this method, the finite value of ARL^1 can be made arbitrary close to one. Still, for this L the ARL^0 will not be finite and thus greater than any fixed value.

Proposition 3.1.3. There exists alarm limits c_s, to the alarm statistic of Proposition 3.1.1, which fulfil "the criterion of minimal ARL^1" by having ARL^1 arbitrary close to the minimal value, one, for a fixed ARL^0.

Proof. Denote the fixed desired value of ARL^0 with A. The method with the alarm limits $c_1 = L$, $c_i = \infty$ for $i = 2,3,...k-1$ and $c_k = -\infty$, where L is a constant and $k = [A - \Phi(-L)] / \Phi(L)$, have ARL^0 $= 1-\Phi(L-0) + k\Phi(L-0) = A$ and ARL^1 $= 1-\Phi(L-\mu) + k\Phi(L-\mu) = 1- \Phi(L-\mu)\ [\Phi(L) + A - \Phi(-L)] / \Phi(L)$, which has the limit one when L tends to minus infinity, since $\Phi(L-\mu)/\Phi(L)$ has the limit zero.□

The above demonstration of the possibility to fulfill the criterion of minimal ARL^1 for a fixed ARL^0, is not intended as a recommendation of how to proceed in practical applications, but is a demonstration of shortcomings of the criterion.

Sometimes optimality is defined as minimal ARL^1/ARL^0. This ratio might be useful as a rough indicator but has drawbacks as a formal optimality criterion. The skewness of the run length distributions (especially if there is a change) and other facts make it easy to construct situations where obviously inferior methods satisfy this criterion. Below the shortcoming of this criterion is illustrated by the often used Shewhart method which gives an alarm as soon as x(s) exceeds a limit G (Section 4.7).

Proposition 3.1.4. For the Shewhart method (see Section 4.7), ARL^1/ARL^0 is decreasing when the limit G increases.

Proof. The method has ARL^0 $=1/(1-\Phi(G))$, ARL^1 $=1/(1-\Phi(G-\mu))$ and thus a ratio

$$ARL^1/ARL^0 = [1-\Phi(G)]/[1-\Phi(G-\mu)]$$

which is decreasing when G increases. □

Thus, in the class of Shewhart methods, the greatest possible limit G should be used. This demonstrates that the optimality criterion of minimal ARL^1/ARL^0 should not be used without care.

3.2 Error probabilities

An important optimality criterion is the maximal detection probability P(A(s)| C(s)) for a fixed false alarm probability P(A(s)| D(s)), and a fixed decision time s, when C = { $\tau \leq s$} and D = { $\tau > s$}. The LR method of Section 4.1 satisfies this criterion which in short will be called "the maximum detection probability criterion". Different error rates were discussed by de Maré (1980) and Frisén and de Maré (1991).

3.3 Utilities

An important specification of utility is that of Girshick and Rubin (1952) and Shiryaev (1963). They treat the case of constant intensity where the gain of an alarm is a linear function of the expected value of the delay, τ-t_A, between the time of the change and the time of the alarm. 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

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

The function h(τ-t_A) could be a constant b, in which case

$$U= b\,P(A(.)|\,D(.)) + a_1\,E[\tau\text{-}t_A|\,\tau\leq t_A] + a_2\,.$$

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 false alarm rate. The criterion will for short be named "the criterion of minimal expected delay". The full likelihood ratio method LR (Section 4.1) satisfies the criterion for the case specified in Section 2.

3.4 Minimax

Minimax solutions with respect to τ avoid the requirement of information about the distribution of τ. Pollak (1985) gives an approximate solution to the criterion of minimal expected difference, τ-t_A, between the time of the change and the time of the alarm for the worst value of τ. The solution is a randomized procedure. The start of the procedure is made in a way that avoids the properties to be dependent on τ. For many applications however, it would be more appropriate with a method depending on the distribution of τ than one depending on an ancillary random procedure. Both dependencies decrease with time.

Moustakides (1986) uses a still more pessimistic criterion by using not only the worst value of τ but also the worst possible outcome of $X_{\tau-1}$ before the change occurs. The CUSUM method, described in section 4.5, provides (except for the first time point) a solution to the criterion proposed by Moustakides.

Ritov (1990) considers a loss function which is not identical to that of Shiryaev but depends on τ and t_A besides τ - t_A. The worst possible distribution $P(\tau=s+1|\ \tau>s;\ X_s)$ is assumed for each time s. With this assumption of a worst possible distribution (based on earlier observations) CUSUM minimizes the loss function.

3.5 Evaluation functions

Optimality criteria are useful, but sometimes a single criterion is not enough and a function is useful for the evaluation. Some examples of this will be given below.

3.5.1 *Delay of an alarm*

A measure related to the ARL[1], but more general, is the conditional expected delay as a function of τ

$$E(t_A-\tau|t_A\geq\tau,\ \tau=t).$$

Differences in shapes of these curves for different methods (Frisén and Wessman 1999) demonstrate the need for other measures than the conventional ARL. For τ=1 the values of this function equal the values of (ARL^1- 1).

The expectation of the delay, also with the respect to the distribution of τ, is:

$$E(t_A - \tau | t_A \geq \tau).$$

This expression is used in the utility functions in Section 3.3. When the distribution of τ is geometrical with the intensity ν, it is sometimes useful to express the expected delay as a function of ν as in Frisén and Wessman (1999).

In some applications there is a limited time available for rescue actions. Then, the expected value of the difference τ-t_A is not of main interest. Instead of using the expected value as in Section 3.3 and 3.4, the probability that the difference does not exceed a preassigned limit is used. The limit, say d, is the time available for successful detection. Bojdecki (1979) considers the supremum (with respect to τ) of

$$P(|t_A - \tau| \leq d).$$

See Section 4.6 for discussion of consequences of this optimality criterion.

A related expression is the probability of successful detection,

$$PSD(\tau,d) = P(t_A - \tau \leq d | t_A \geq \tau).$$

3.5.2 Predictive value

The predictive value PV(s) = P(C(s)| A(s)) of an alarm at time s has been suggested as a criterion of evaluation by Frisén (1992). The predictive value tells us how probable a change is when we have an alarm. Thus, it gives important information about which action that would be appropriate. It simplifies matters if the same action can be used whenever an alarm occurs. Thus, a constant predictive value with respect to time is a good property.

The relation between the predictive value and the posterior distribution PD(s) = P(C(s) | X_s) is different for passive and active surveillance. This is important since the method of giving an alarm as soon as the posterior distribution exceeds a fixed limit is often advocated. See, e.g. Smith et. al. (1983) and Harrisson and Veerapen (1994).

Proposition 3.5.2.1. At passive surveillance the method based on the posterior distribution with ${}_pA(s)= [X_s; PD(s)>c]$ implies a lower limit of the predictive value, $PV(s) > c$.

Proof. $PV(s) = P(C(s) \mid A(s)) = P(C(s) \mid X_s; P(C(s) \mid X_s) > c) > c$. □

At passive surveillance the predictive value typically increases to one as time s increases, since $P(C(s)) = P(\tau \leq s)$ tends to one. As an example, the predictive value for the Shewhart method, when τ has a geometric distribution with intensity v will be given. For the Shewhart method, the alarm probabilities $\alpha = P(t_A{=}t \mid t_A \geq t, D)$ and $(1-\beta) = P(t_A{=}t \mid t_A \geq t, C)$ do not depend on time which simplifies formulas.

$$PV(s) = P(C(s) \mid A(s)) = P(C(s) \cap A(s))/P(A(s))$$

$$= \sum_{\tau=1}^{s} (1-\gamma)^{\tau-1} \nu(1-\beta) \Big/ \left[(1-\nu)^s + \sum_{\tau=1}^{s} (1-\gamma)^{\tau-1} \nu(1-\beta) + \right]$$

$$= \left[\nu(1-\beta)(1-(1-\nu)^{s-1}) \right] \Big/ \left[\nu(1-\beta)(1-(1-\nu)^{s-1}) - \alpha\nu(1-\nu)^{s-2} \right]$$

which tends to one when s tends to ∞.

At active surveillance, the process is stopped if ${}_aA(1)$ occurs. Otherwise the complement ${}_aA^c(1)$ occurs and for s=2, 3, ... write ${}_aA^c_{s-1} = {}_aA^c(1) \cap {}_aA^c(2) \cap \ldots {}_aA^c(s)$. In this active case, the simple relation in the Proposition 3.2.1 above is no longer true. Instead $PV(s) = P(C(s) \mid {}_aA(s) \cap {}_aA^c_{s-1})$.

At active surveillance the predictive value typically has an asymptote less than one since we have that the probability of the first alarm at s decreases with s for large s. The formula of the asymptote for the Shewhart method is given in Frisén (1992). Graphs of the predictive value for different methods are given in Frisén and Wessman (1999). The predictive value is not monotonically increasing for all methods.

There is a great difference between a single decision and a sequence of decisions. At a single decision the posterior distribution might give sufficient information. For a sequence of decisions, characteristics of the sequence, such as constant predictive value, become of interest.

4. METHODS

In Figure 1 the alarm set of some methods, which will be described below, are illustrated for the decision time s=2. The purpose of the figure is to illustrate the geometrical differences of the alarm sets.

In Table 1 some main characterizations of some methods are schematically described. The number of parameters which can be used to optimize for different situations is one important difference. Many methods for surveillance are based in one way or another on likelihood ratios. For the comparison, expressions in terms of the basic likelihood ratios are also given in Table 1.

4.1 The likelihood ratio method

A method constructed by Frisén and de Maré (1991) to meet several optimality criteria, e.g. those of Sections 3.2 and 3.3, will first be presented. The general method uses combinations of likelihood ratios. Although methods based on likelihood ratios have been suggested earlier, for other reasons, the use in practice is (yet) rare. The likelihood ratio method will be used as a "benchmark". Commonly used methods are compared with it in order to clarify their optimality properties.

Here, the likelihood ratio method is applied to the shift case specified in Section 2. The "catastrophe" to be detected at decision time s is $C = \{ \tau \leq s\}$ and the alternative is $D = \{ \tau > s\}$.

The likelihood ratio method has an alarm set consisting of those X for which the likelihood ratio exceeds a limit:

$$f_{X_s}(x_s \mid C) / f_{X_s}(x_s \mid D) = p(x_s) > G_s.$$

For the case of $C = \{ \tau \leq s\}$ this can be expressed as

$$\sum_{t=1}^{s} w(t)L(t) > G_s$$

where $w(t) = P(\tau=t)/P(\tau\leq s)$ and L(t) is the likelihood ratio for the case when τ=t.

$$L(t) = f_{X_s}(x_s \mid \tau=t) / f_{X_s}(x_s \mid D)$$

For the case of normal distribution, $C(s)=\{\tau\leq s\}$ and $D(s)=\{\tau>s\}$, as specified in Section 2, we have

$$p(x_s) = g(s)\; p_s(x_s)$$

where

$$g(s)=\frac{\exp(-(s+1)\mu^2/2)}{P(\tau\leq s)}$$

does not depend on the data and

$$p_s(x_s)=\sum_{i=1}^{s}\pi_i \exp\left\{\frac{1}{2}i\mu^2\right\}\exp\{\mu\sum_{t=i}^{s}x(t)\}$$

is a nonlinear function of the observations.

In order to achieve the optimal error probabilities described in Section 3.2, an alarm should be given as soon as $p(x_s) > G_s$.

In the case of geometric distribution of τ the condition of "minimal expected delay", as described in Section 3.3, is achieved if an alarm is made as soon as the posterior distribution exceeds a fixed limit (Shiryaev 1963).

$$P(\tau\leq s|X_s=x_s)>K \quad \Rightarrow \quad p(x_s)>\frac{P(\tau>s)}{P(\tau\leq s)}\;\frac{K}{1-K}$$

where K is a constant. Thus, the optimality is achieved by the likelihood ratio method with the additional requirement

$$G_s = K\, P(\tau > s) \,/\, (1-K)\, P(\tau \leq s).$$

The method for this limit, that thus gives alarm for the first s where

$$\sum_{i=1}^{s}\pi_i \exp\left\{\frac{1}{2}i\mu^2\right\}\exp\{\mu\sum_{t=i}^{s}x(t)\} \;>\; G_s/g(s)=\exp((s+1)\mu^2/2)P(\tau>s)\frac{K}{1-K}$$

will here be called the LR method. A usual assumption is that τ has a geometric distribution with $\pi_t=(1-\nu)^{t-1}\nu$. The shape of the alarm set for this case is illustrated in Figure 1. The alarm is given for the first s where

(1) $$\sum_{i=1}^{s}(1-\nu)^{i-1}\nu\ \exp\left\{\frac{1}{2}i\mu^2\right\}\exp\{\mu\sum_{t=i}^{s}x(t)\} > \exp((s+1)\mu^2/2)(1-\nu)^s\frac{K}{1-K}\ .$$

When ν tends to zero both the weights w(t) and the limit G_s of the LR method tend to constants. Shiryaev (1963) and Roberts (1966) suggested the method, which is now called the Shiryaev-Roberts method, for which an alarm is triggered at the first time s, for which

$$\sum_{t=1}^{s} L(t) > G$$

where G is a constant. The method has an approximately constant predictive value (Frisén and Wessman 1999), which makes it easier to interpret alarms which happens late or early during the surveillance.

The posterior distribution PD(s) = P(C(s) | X_s) has been suggested as an alarm criterion by, e.g. Smith et al (1983). When there are only two states, C and D, this criterion leads to the LR method. Sometimes the use of the likelihood ratio or equivalently the use of the posterior distribution is named “the Bayes´ method”. This name is avoided here since it might give wrong associations. Here no use of Bayesian inference will be made. Bayes’ theorem is used and τ is considered as a stochastic variable but no results are dependent on the perspective of Bayesian inference.

4.2 Linear approximation of the likelihood ratio method

To obtain a method which is easier to use, and also to clarify the connection with other methods, a linear approximation of the alarm-function p_s is of interest. The exponential functions of the partial sums of the observations will be approximated by linear functions. The situation, often studied in the literature, with late changes and thus expected values of the partial sums which are close to zero, is considered. By approximation by Taylor expansions

$$\exp\left\{\mu\sum_{t=i}^{s}x(t)\right\} \approx 1+\mu[\sum_{t=i}^{s}x(t)]$$

and with a = exp($\mu^2/2$) the following linear approximation is achieved:

$$p_s(x_s) \approx p_s{}^*(x_s) = \sum_{i=1}^{s} \pi_i [1 + \mu \sum_{t=i}^{s} x(t)] \exp(i\mu^2/2) =$$

$$= \sum_{i=1}^{s} \pi_i a^i + \mu \sum_{i=1}^{s} \pi_i a^i \sum_{t=i}^{s} x(t) =$$

$$= m(s) + \mu \sum_{t=1}^{s} x(t) m(t),$$

where the weights for the observations are

$$m(t) = \sum_{i=1}^{t} a^i \pi_i.$$

The linear approximation of the LR method is here denoted as the LLR method. It will give an alarm as soon as

$$p_s{}^{**}(x_s) = \sum_{t=1}^{s} x(t) m(t)$$

exceeds the limit

$$[G_{s.}/g(s) - m(s)]/\mu$$

$$= [a^{s+1} \; P(\tau > s) \; \frac{K}{1-K} \; - \; m(s)]/\mu$$

$$= [\exp((s+1)\mu^2/2) \; P(\tau > s) \; \frac{K}{1-K} \; - \; \sum_{k=1}^{s} \pi_k \exp(k\mu^2/2)]/\mu.$$

The sum of the weights is

$$\sum_{t=1}^{s} m(t) = \sum_{t=1}^{s} \sum_{i=1}^{t} a^i \pi_i = \sum_{t=1}^{s} \pi_t \sum_{i=t}^{s} a^i.$$

With adjustment to make the sums of the weights equal to one we have the weights

$$w_L(t)=m(t)/\sum_{t=1}^{s}m(t)=\sum_{i=1}^{t}a^{i}\pi_i/\sum_{t=1}^{s}\pi_t\sum_{i=t}^{s}a^{i}$$

and the LLR method can be expressed as

$$\sum_{t=1}^{s}w_L(t)x(t) > [\frac{a^{s+1}}{\sum_{t=1}^{s}\pi_t\sum_{i=t}^{s}a^{i}}\,P(\tau>s)\,\frac{K/\mu}{1-K} - w_L(s)]/\mu.$$

If the intensity is constant, when τ has a geometric distribution $\pi_t=(1-\nu)^{t-1}\nu$ and then, with $b=a(1-\nu) = (1-\nu)\exp(\mu^2/2)$, we have

$$m(t) = [\nu/(1-\nu)]\sum_{i=1}^{t}b^{i} = \frac{b\nu}{(b-1)(1-\nu)}(b^{t}-1)$$

$$\sum_{t=1}^{s}m(t) = \frac{b\nu}{(b-1)(1-\nu)}\,\frac{b(b^{s}-1)-s(b-1)}{b-1}$$

and

$$w_L(t) = (b^{t}-1)\,\frac{b-1}{b(b^{s}-1)-s(b-1)} \propto (b^{t}-1)$$

For geometrically distributed τ the alarm criterion for the LLR method becomes

$$\text{(2)}\qquad \sum_{t=1}^{s}w_L(t)x(t)> [\frac{b^{s}}{b(b^{s}-1)-s(b-1)}\{\frac{(b-1)^2K}{\nu(1-K)}-(b-1)\}+\frac{b-1}{b(b^{s}-1)-s(b-1)}]/\mu$$

For large s the limit tends to

$$[\frac{b^{s}}{b(b^{s}-1)}\{\frac{(b-1)^2K}{\nu(1-K)}-(b-1)\}]/\mu$$

which is proportional to $b^s/(b^s-1)$. For s=2 the alarm set is illustrated in Figure 1.

4.3 Exponentially weighted moving average

A method for surveillance based on exponentially weighted moving averages, usually called EWMA, was introduced in the quality control literature by Roberts (1959). Recently the method has got much attention. This may be due to papers by Robinson and Ho (1978), Crowder (1987), Ng and Case (1989), Lucas and Saccucci (1990) and Domangue and Patch (1991) in which positive reports of the quality of the method are given.

The alarm statistic is

$$Z_s = (1-\lambda)Z_{s-1}+\lambda x(s),\ s=1, 2, \ldots$$

where $0<\lambda<1$ and in the standard version of the method Z_0 is the target value μ^0, which here is chosen to zero.

The statistic is sometimes referred to as a geometric moving average since it can equivalently be written as

$$Z_s = \lambda\sum_{j=0}^{s-1}(1-\lambda)^j x(s-j) = \lambda(1-\lambda)^s\sum_{t=1}^{s}(1-\lambda)^{-t}x(t) \propto \sum_{t=1}^{s}k^t x(t)$$

where $k=1/(1-\lambda)$ is a constant > 1.

An out-of-control alarm is given if the statistic Z_s exceeds an alarm limit, usually chosen as $L\sigma_Z$, where L is a constant and σ_Z the limiting value, as s tends to infinity, of the standard deviation. This method will give an alarm for the first s for which

$$Z_s = \lambda\sum_{j=0}^{s-1}(1-\lambda)^j x(s-j) > L\sigma_Z,$$

or equivalently

$$(3) \qquad \sum_{t=1}^{s} w_E(t)x(t) > L_E$$

with weights $w_E(t)= k^{t-1}(k-1)/(k^s-1)$ which sum to one and with $L_E = L\sigma_Z\, k^s/(k^s-1)$.

EWMA gives the most recent observation the greatest weight, and gives all previous observations geometrically decreasing weights. If λ is equal to one, only the last observation is considered and the resulting method is the Shewhart method described in Section 4.7. If λ is near zero, all observations have approximately the same weight.

Proposition 4.3.1 There does not exist any λ which makes the EWMA exactly optimal in the sense of Sections 3.2 or 3.3.

Proof. The likelihood method, which satisfies the optimality criteria above, gives alarm when a nonlinear function of the observations exceeds a fixed limit, while the EWMA method gives alarm when a linear function exceeds a fixed limit. □

Since the EWMA method has two parameters, λ and L, these can be chosen to equal any other linear method when s=2, as in Figure 1. It is thus not included separately in that figure. When s>2 differences appear and similarities with the linear LLR method will now be examined.

Proposition 4.3.2 The weights of the observations in the alarm function of the EWMA method cannot be exactly identified with the weights by the LLR method for the case of constant intensity. For late observations approximate identification is achieved with $\lambda = 1 - \exp(-\mu^2/2)/(1-\nu)$, when this is positive.

Proof. At constant intensity ν

$$\pi_i = (1-\nu)^{i-1}\nu \quad i=1, 2, \ldots$$

The weights, m(t) of the LLR method are found in Section 4.2. The relative weights are

$$m(t+1)/m(t) = (1-b^{t+1})/(1-b^t) = b + (1-b)/(1-b^t).$$

The relative weights are thus not constant for the LLR method as they are for the EWMA method. For large values of u the relative weight tends to b when b>1. When $m(t+1)/m(t) = k= 1/(1-\lambda) = b = (1-\nu)\exp(\mu^2/2)$ and thus $\lambda = 1 - \exp(-\mu^2/2)/(1-\nu)$. □

The comparison between the weights of the LLR method and the weights of the EWMA method with $\lambda = 1 - \exp(\mu^2/2)/(1-\nu)$ is made in Figure 2. In the beginning of the surveillance the EWMA puts more weight to the older observations than the LLR method. However, already for decision time s=10 the differences between the two methods are without importance. For s=15 it is not possible to see any difference in the scale of the figure. The approximately optimal values of λ are given as a function of μ in Figure 3. It is given for different values of the intensity, ν. The effect of ν is moderate. Instead, the size, μ, of the shift which is to be detected, has a major effect. This effect is slightly more pronounced for the greater intensities. The approximation in Proposition 4.3.2 requires that μ is not too small if ν is very large. Also by the formula in Proposition 4.3.2, it is seen that the parameter λ of the EWMA method rapidly increases to one which makes the method identical to the Shewhart method.

So far only one time s of decision has been considered. The LLR method is an approximation of the LR method which satisfies "the maximum detection probability criterion" for each value of s. The EWMA method with proper weights can thus be expected to have good properties according to Section 3.2 for each value of s. For a full comparison of the methods it is necessary also to consider how the limits for alarm depend on s for different methods. The limit for the EWMA method depends on the decision time s as $k^s/(k^s-1)$ as was seen in the beginning of this section. With λ chosen as in Proposition 4.3.2 we have k=b. In Section 4.2 it was demonstrated that also the limit for the LLR method depends on s as $b^s/(b^s-1)$ for large s. Thus, for this choice of λ the EWMA method approximately fulfills also the optimality condition of Section 3.3 of a minimal expected delay.

According to Proposition 3.1.1 λ should approach zero in order to give equal weight to all observations and thus give an alarm statistic which can give a minimal ARL^1 for a fixed value of ARL^0. However, the alarm limit does not depend on s in the way required (see Proposition 3.1.2 and 3.1.1). Papers which determine optimal λ according to the criterion of minimal ARL^1 (e.g. Crowder (1989), Lucas and Saccucci (1990) and Srivastava and Wu (1997)) recommend considerably smaller values than those derived here for the minimal expected delay.

4.4 Simple cumulative sums

Sometimes CUSUM is used as a unifying notation for methods based on the cumulative sum of the deviations between a reference value and the observed values. In the simplest form there is an alarm as soon as the cumulative sum of differences from the target value μ^0 which here, by standardization, is set to zero

(4)
$$C_s = \sum_{t=1}^{s} x(t)$$

exceeds a fixed limit. This method is sometimes called the simple CUSUM. It will here be denoted as SCUSUM. The SCUSUM method gives optimal error probabilities for $\tau=1$ in the case specified in Section 2. However, Frisén (1992) demonstrated that when $\tau>1$, SCUSUM cannot compete with other methods with the same ARL^0. The probability of successful detection within a short time is lower. Also, the predictive value of an alarm is strongly decreasing with the time of the alarm. As is seen in Figure 1 the shape of the alarm set is quite different from the optimal one to minimize the expected delay.

Another simple method based on cumulative sums is the method which gives an alarm when the likelihood ratio for $C=\{\tau=1\}$ against $D=\{\tau>s\}$ exceeds a fixed constant. As was demonstrated in Proposition 3.1.2 we have an alarm at

$$(5) \qquad t_A = \min\{s: \sum_{t=1}^{s} x(t) > L + s\mu/2\}.$$

This method, which gives an alarm as soon as C_t exceeds a linear function of s is here called the LCUSUM method. The method is a sequential probability ratio test without the limit for acceptance. The alarm set of the method can also be expressed by the likelihood ratio condition $L(1) > G$, where G is a constant and L(1) as before is the likelihood ratio for $C=\{\tau=1\}$. For the SCUSUM method the limit for L(1) depends on s. The LCUSUM method has minimal $E(t_A-\tau)$ when $\tau=1$among methods with the same total false alarm probability. In Figure 1, where the alarm limit for s=2 is illustrated, the LCUSUM is identical to the SCUSUM since the only difference is how the limit for alarm depends on the decision time s.

For both SCUSUM and LCUSUM, the data from all earlier points in the time series have the same weights as the last one. As soon as only $\tau=1$ is considered (as in the criterion that minimizes the ARL^1 for fixed ARL^0) these weights are the optimal ones. For most applications this is not considered rational. The most often suggested optimality criterion in the literature on quality control does thus lead to a type of method which is seldom used in practice.

4.5 CUSUM

The variant of cusum tests, which is most often advocated, is called the CUSUM or V-mask. It can be based on a diagram of the cumulative sums of deviations from the target value. In the two-sided case a V-shaped mask is moved over the diagram until some earlier observation is outside the limits of the mask and an alarm is given. The two legs of the V are usually placed symmetrically to the horizontal line. The apex of the V is placed on the same level as the last observation but at a distance to the right of the observation. There is thus an alarm for the first s for which

(6) $$|C_s - C_{s-i}| > h + ki \quad \text{for some } i=1, 2, ..., s,$$

where $C_0 = 0$ and h and k are chosen constants. The parameter k determines the slopes of the legs in the V-shaped mask and h determines the location of it. The distance between the apex and the last observation is h/k if the axes have the same scale. In that case the angle of the V-shaped mask is 2*arctan(k). In V-masks with a very narrow angle there is no big difference between the weights of recent and old observations and there are similarities to the simple cusum test. With a wide angle the last observation has a heavy weight and there are similarities to the Shewhart test. By the CUSUM method (in contrast to the simple variants of Section 4.4) the information from earlier observations is handled quite differently depending on the position in the time series.

Sometimes (e.g. Siegmund 1985 and Park and Kim 1990) the CUSUM test is presented in a more general way by likelihood ratios (which in the normal case reduce to C_t-C_{t-i}). Observe however that this is not the LR method described above. The CUSUM method is the result of a natural (but not optimal) combination of methods. Each of these is optimal to detect a change that occurs at a specific time point. The alarm condition of the method can be expressed by the likelihood ratios for $C=\{\tau=t\}$ as

$$\max(L(t); t=1, 2,.., s) > G,$$

where G is a constant.

The optimal value of the parameter k of (6) is usually claimed to be $k=(\mu^0+\mu^1)/2$, which after our standardization reduces to $\mu/2$. The chain of references (if any) usually ends with Ewan and Kemp (1960). In that paper they conclude from a nomogram that this value seems to be about the best. The likelihood ratio method for $C=\{\tau=i\}$ gives alarm for

$$\sum_{t=i}^{s} x(t) > c + (s-i)\mu/2.$$

where c is a constant. Thus, also here we have the slope $\mu/2$. That this slope is optimal in each step does explain why it "seems to be about the best". However it does not prove that it is optimal for the sequence of decisions.

The CUSUM, with $k= \mu/2$ satisfies certain minimax conditions (Moustakides 1986 and Ritov 1990) as was discussed in Section 3.4. In Figure 1 the alarm limit of the CUSUM method is seen to be a two-phase linear approximation of the nonlinear limit of the LR method.

4.6 Moving average

The moving average method gives an alarm as soon as

(7) $$C_s - C_{s-d} > L$$

where d is a fixed window width and L is a constant. The alarm set can also be expressed by the likelihood ratios L(t) as

$$L(s\text{-}d) > G$$

where G is a constant.

The method can be shown to be a special case of the solution of Bojdecki (1979) to the maximization of

$$P(|\tau - t_A|) \leq d)$$

where t_A is the time of alarm. See Section 3.5 for discussion on this optimality criterion.

4.7 Shewhart

This method, which is much used in quality control, was suggested by Shewhart already (1931) An alarm is triggered as soon as an observation deviates too much from the target. The stopping rule is that we have an alarm as soon as

(8) $$x(s) > G.$$

The limit G for a fixed ARL^0, is calculated by the relation: $P(X(s)>G|\mu(s)=\mu^0)=1/ARL^0$. For illustration of the alarm set at decision time s=2 see Figure 1. More expanded descriptions are found in many textbooks like Wetherill and Brown (1990).

The alarm statistic of the LR method

$$f_{X_s}(x_s \mid C) / f_{X_s}(x_s \mid D)$$

reduces to that of the Shewhart method when the "catastrophe" to be detected at decision time s is $C = \{ \tau = s\}$ and the alternative is $D = \{ \tau > s\}$. The alarm set can be expressed by the condition

$$L(s) > G$$

where G is a constant. Thus the Shewhart method has optimal error probabilities for these alternatives for each decision time s.

For large shifts, it was demonstrated by Frisén and Wessman (1999) that the LR method and the CUSUM method converge to the Shewhart method.

5. CONCLUDING REMARKS

The performance of a system of surveillance depends on the time of the change τ. To get an index with a single value, either a summarizing measure over the distribution of τ, or evaluation for a specific value of τ, can be used. Suggested optimality criteria based on specific values of τ are those based on τ=1, "$\tau=\infty$" or τ= "worst possible value". In Roberts (1959 and 1966) the value τ=8 was used, but that was because of technical reasons. The solution to an optimality criterion based on τ= "worst possible value" is a randomized procedure. Recent suggestions are to make the minimax criterion still more pessimistic by also assuming the worst possible outcome. Optimization and evaluations for the steady state case "$\tau=\infty$" are of great value, but as with other asymptotic results it is not enough for all applications.

In quality control, optimality criteria based on ARL, with the same distribution for all time points, is the common choice. Sometimes the criterion is expressed as the ratio ARL^1/ARL^0. As was noted in Proposition 3.1.4, this has unreasonable implications. More often the criterion is stated as minimal ARL^1 for a fixed ARL^0. As was noted in Proposition 3.1.1 this criterion implies methods where all observations have the same weight. The shortcomings of such methods were pointed out in Section 4.4 and they are not often recommended. Instead, methods which have all weight on the last observation (Shewhart) or gradually less weight on the older observations (EWMA and CUSUM) are commonly recommended in the literature on quality control. Methods which have good properties when $\tau=1$ might not be as good if the change occurs later. If the problem is to discriminate between the hypothesis $\mu(t)=0$ for all t and the hypothesis $\mu(t)=\mu$ for all t, when sequential methods for tests of hypotheses (such as the power one SPRT method of Proposition 3.1.2) are appropriate. Only the situations where a change is expected to happen after an unknown time, τ, require the special methods for surveillance.

A summarizing optimality criterion is achieved by using an assumption on the distribution of τ. Exact information about the distribution might be lacking. However, the drawbacks with the criteria based on ARL demonstrate the importance of any information on the distribution of τ.

Criteria based on the posterior distribution have an intricate relation both to the LR method and to the predictive value of an alarm. These relations were analyzed in Section 3.5.2 for passive and active surveillance.

The LR method is nonlinear with respect to the data. Commonly used methods are equivalent to the LR method only at extreme cases where the nonlinearity disappears. The linear approximation, LLR, is here used mainly for the comparison with other linear methods. The comparison is used to demonstrate how the parameter λ of the EWMA method should depend on the size of the shift and intensity of the change to be detected. The value of λ which approximately satisfies "the criterion of minimal expected delay" depends strongly on μ but very little on ν. The result that the method with this choice of λ approaches the Shewhart method when μ increases is in agreement with the results by Frisén and Wessman (1999) that when the LR, Shiryaev-Roberts and the CUSUM methods are optimized for large shifts they are very much alike the Shewhart method. This λ differs, both in the level and in the slope of the function of μ, in an expected way from the results where "the criterion of minimal ARL^1" is used.

The EWMA method has continuously decreasing weights for older observations. The CUSUM method has a discrete adaptive way of including old observations. This can explain the good minimax properties for the CUSUM method. The EWMA method has bad "worst possible" properties according to Yashchin (1987). The best thing would be to have continuous adaptive weights. That is actually what the LR method gives.

The simple cumulative sum methods SCUSUM and LCUSUM satisfy optimality conditions for C={τ=1}. They are linear, but with equal weight to all observations in contrast to the linear approximations of the LR method which give more weight to later observations.

The limits for the alarm functions in Figure 1 are not comparable with respect to false alarm probability. The false alarm probability P(t_A=s|D) depends on s in different ways for the different methods. Thus the area under the curves cannot be interpreted. However, the shapes of the curves demonstrates geometrically some characteristics. The linear methods LLR and EWMA (with two and one adjustable parameter respectively) can approximate the nonlinear LR method rather well. The two-phase linear CUSUM method which has an adjustable parameter also approximates the smooth LR method rather well. However the Shewhart and the SCUSUM methods which do not have any adjustable parameter except the limit can only approximate the LR method for very special cases.

The robustness with respect to the choice of parameters is also of interest. The properties of different methods when the actual shift μ or intensity v is not the same as those M and V for which the method was optimized have been examined. Srivastava and Wu (1993) studied the asymptotic effect of different true μ for a fixed parameter M. Järpe and Wessman (1999) studied the same effect for small samples. Frisén and Wessman (1999) studied the small sample properties for different values of M for a fixed μ to examine the robustness to the choice of parameter value M. The theorems and the figures demonstrate that the choice of a large value of M makes the properties of the methods more alike. For large values of M all methods behave as the Shewhart method. Heuristically, a method designed to detect a large shift with a small expected delay should allocate nearly all weight to the single last observation. A consequence is that with specification to a large value of the shift the choice of method is not very important. The similarity is pronounced for M larger than 2 for μ=1. This confirms the results by Mevorach and Pollak (1991) that the Shiryaev-Roberts method and the CUSUM method have similar properties for the cases M=5 and M=7 for μ=1. The study by Frisén and Wessman (1999) confirms the conjecture by Roberts (1966) about the robustness with respect to differences between the assumed and true intensities V and v.

Here, the simplest and in the literature most commonly discussed situation has been treated in order to concentrate on principal inferential matters. However, also many other situations are of interest for applications. Some examples will now be given about results for such situations. Multivariate surveillance is of interest in many applications. Wessman (1998) has examined the case there many processes are monitored for a common change point and demonstrated that univariate surveillance can be used. The case of spatial surveillance has been studied by, e.g. Järpe (1999) where also a reduction to ordinary univariate surveillance was demonstrated to be the proper solution. Here the case of independent observations is described but also the case of autocorrelated observations has been studied by, e.g. Alwan (1992). Here normally distributed observations were studied but the surveillance of the frequency of events has been studied by, e.g. Radelli and Gallus (1989).

ACKNOWLEDGMENT

This work was supported by the Swedish Council for Research in the Humanities and Social Sciences. Graduate student Christian Sonesson has contributed to Proposition 3.1.3.

REFERENCES

Alwan, L.C. (1992). Effects of autocorrelation on control chart performance. *Communications in Statistics - Theory and Methods* **21** 1025-1049.

Basseville, M. and Benveniste A. (1986). *Detection of abrupt changes in signals and dynamical systems*. Berlin: Springer.

Bojdecki, T. (1979). Probability maximizing approach to optimal stopping and its application to a disorder problem, *Stochastics* **3** 61-71.

Brodsky, B. E. and Darkhovsky B. S. (1993). *Nonparametric methods in change point problems*. Dordrecht: Kluwer Academic Publishers.

Crowder, S. V. (1987). A simple method for studying run-length distribution of exponentially weighted moving average charts. *Technometrics* **29** 401-407.

Crowder, S.V. (1989). Design of exponentially weighted moving average schemes. *J. 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.

Ewan W. D. and Kemp K. W. (1960). Sampling Inspection of Continuous Processes with no Autocorrelation between Successive Result. *Biometrika* **47** 363-380.

Frisén, M. (1992). Evaluations of methods for statistical surveillance, *Statistics in Medicine* **11** 1489-1502.

Frisén M. (1999). Statistical Surveillance of Business Cycles. Submitted.

Frisén M. and Cassel C. (1994). Visual evaluations of statistical surveillance. Research report, 1994:3, Department of Statistics, Göteborg University.

Frisén, M. and de Maré, J. (1991). Optimal surveillance, *Biometrika* **78** 271-280.

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.

Girshick, M. A. and Rubin, H. (1952). A Bayes approach to a quality control model. *Annals of Mathematical Statistics* **23** 114-125.

Harrison, P. J. and Veerapen, P. J. (1994). A Bayesian Decision Approach to Model Monitoring and Cusums. *Journal of Forecasting* **13** 29-36.

James, B., James, K. L. and Siegmund D. (1987). Tests for a change-point. *Biometrika* **74** 71-83.

Järpe, E. (1999). Surveillance of spatial patterns. To appear in *Communications in Statistics. Theory and methods* **28**.

Järpe, E. and Wessman, P. (1999). Some power aspects of methods for detecting different shifts in the mean. To appear in *Communications in Statistics. Simulation and Computation* **28**.

Kolmogorov, A. N., Prokhorov, Y. V. and Shiryaev, A. N. (1990). Probabilistic-statistical methods of detecting spontaneously occurring effects. Proceedings of the Steklov Institute of Mathematics, 1-21.

Lai, T.L. (1995). Sequential changepoint detection in quality control and dynamic systems. *J. of the Royal Statistical Society B* **57** 613-658.

Lindgren, G. (1985), Optimal prediction of level crossings in Gaussian processes and sequences, *Ann. Prob* **13** 804-24.

Lucas, J. M. and Saccucci, M. S. (1990), Exponentially weighted moving average control schemes: properties and enhancements, *Technometrics* **32** 1-12.

Maré, J. de (1980), Optimal prediction of catastrophes with application to Gaussian processes, *Ann. Prob.* **8** 841-850.

Mevorach, Y. and Pollak, M. (1991), "A small sample size comparison of the CUSUM and Shiryaev-Roberts approaches to changepoint detection," *American Journal of Mathematical and Management Sciences* **11** 277-298.

Moustakides, G. V. (1986), Optimal stopping times for detecting changes in distributions, *Ann. Statist.* **14** 1379-87.

Ng, C. H. and Case, K. E. (1989), Development and Evaluation of Control Charts Using Exponentially Weighted Moving Averages, *J. Quality Technology*, 21, 242-250.

Page, E. S. (1954), Continuous inspection schemes, *Biometrika*, 41, 100-114.

Park, C. S. and Kim, B. C. (1990) A CUSUM chart based on log probability ratio statistic, *J. Korean Statistical Society* **19**, 160-170.

Pettersson,M. (1998). Monitoring a freshwater fish population: Statistical surveillance of biodiversity. *Environmetrics* **9** 139-150.

Pollak, M. (1985) Optimal stopping times for detecting changes in distributions. *Annals of Statistics* **13**, 206-227.

Pollak M. and Siegmund D. (1991) Sequential detection of a change in a normal mean when the initial value is unknown. *Annals of Statistics* **19** 394-416.

Ritov, Y. (1990) Decision theoretical optimality of the CUSUM procedure, *Annals of Statistics* **18** 1464-1469.

Radelli, R. and Gallus, G. (1989) On Detection of a Change in the Dynamics of Rare Health Events. *Communications in Statistics - Theory and Methods* **18** 579-590.

Roberts, S. W. (1959), Control Chart Tests Based on Geometric Moving Averages, *Technometrics* **1** 239-250.

Roberts S. W. (1966) A comparison of some control chart procedures. *Technometrics* **8** 411-30.

Robinson, P. B. and Ho, T. Y. (1978), Average Run Lengths of Geometric Moving Average Charts by Numerical Methods, *Technometrics*, 20, 85-93.

Shewhart, W. A. (1931). Economic Control of Quality Control. Reinhold Company, Princeton N.J.

Shiryaev, A. N. (1963), On optimum methods in quickest detection problems, *Theory Probab. Appl.*, 8, 22-46.

Siegmund, D. (1985). *Sequential analysis. Tests and confidence intervals*, Springer.

Smith, A. F. M., West, M., Gordon, K.., Knapp, M. S. and Trimble, M. G. (1983). Monitoring kidney transplant patients. *The Statistician* **32** 46-54.

Srivastava, M..S. and Wu, Y (1993). Comparison of EWMA, CUSUM and Shiryayev-Roberts procedures for detecting a shift in the mean. *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.

Telksnys, L. (1986). *Detection of changes in random processes*. New York: Springer.

Vardeman, S. and Cornell, J. A. (1987). A partial Inventory of Statistical Literature on Quality and Productivity through 1985. *J. Quality Technology* **19** 90-97.

Wessman, P. (1998). Some Principles for surveillance adopted for multivariate processes with a common change point. *Communications in Statistics - Theory and Methods* **27** 1143-1161.

Wetherill, G.B. and Brown, D.W. (1990). *Statistical process control*, London: Chapman and Hall.

Yashchin, E. (1987). Some aspects of the theory of statistical control schemes, *IBM J. Res. Develop.* **31** 199-205.

Zacks S. (1983). Survey of classical and Bayesian approaches to the change-point problem: Fixed sample and sequential procedures of testing and estimation. in *Recent advances in statistics*, 245-269.

LEGENDS TO TABLE AND FIGURES

Table 1. Schematic characterization of methods by optimality properties described in the text.

Figure 1. Alarm limits at decision time s=2 for some methods described in the text and in Table 1. The values ν=0.01 and μ=1 were used for those methods which can be optimized.

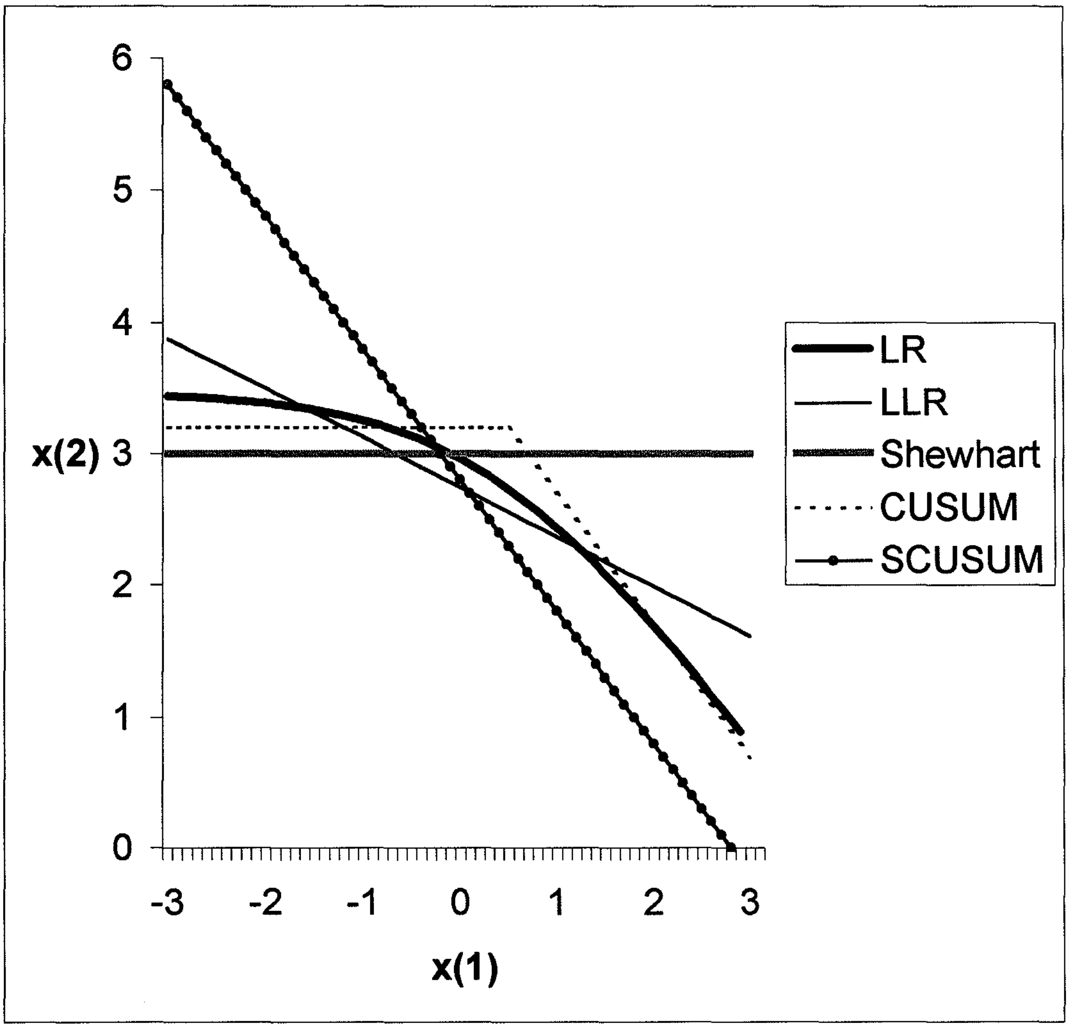
Figure 2. Connections with straight lines of the weights w(t) of the observations x(t). The weights of the EWMA method are calculated for $\lambda = 1 - \exp(\mu^2/2)/(1-\nu)$. The LLR method is optimized for the case when the change τ has a geometric distribution with intensity ν=0.01 and the shift is μ=1 and the same values are used for λ. The pairs of curves are for decision times s = 5 and 10.

Figure 3. The value of the parameter λ of the EWMA method, which according to Proposition 4.3.2 is approximately optimal, as a function of the shift, μ, for the intensities, ν, 0,10, 0,05 and the limiting value 0.

Table 1

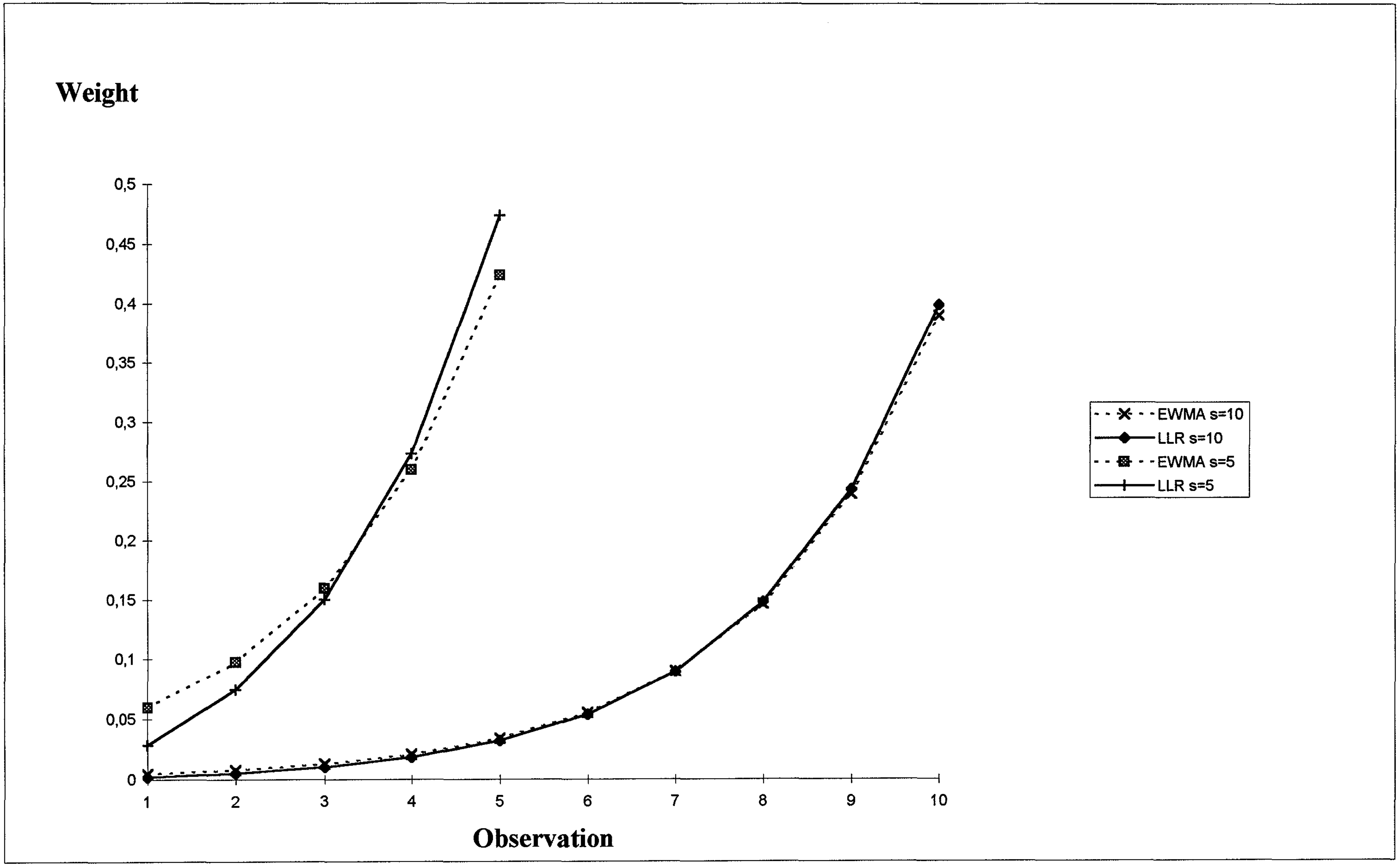
Method		Formula number	Alarmfunction of L(t)	No of parameters except the limit	Optimality
LR (full likelihood ratio)		(1)	$\sum w(t)L(t)$	2	min $E(t_A - \tau \| t_A \geq \tau)$ for fixed $P(t_A < \tau)$ and max $P(A(s) \| C)$ for fixed $P(A(s) \| D)$ when $C = \{ \tau \leq s \}$ and $D = \{ \tau > s \}$
Shiryaev Roberts		(1) with $\nu \to 0$	$\sum_{t=1}^{s} L(t)$	1	As for LR if $\nu \to 0$
LLR (linearization of the LR method)		(2)		2	approximation of that for LR
EWMA	with λ_{LLR}	(3)		1	approximation of that for LR
	with small λ				approximation of that for SCUSUM
SCUSUM		(4)	L(1)	0	max $P(A(s) \| C)$ for fixed $P(A(s) \| D)$ when $C = \{ \tau = 1 \}$ and $D = \{ \tau > s \}$
LCUSUM		(5)	L(1)	0	min ARL^1 for fixed total false alarm probability
CUSUM		(6)	maxL(t)	1	best min max $E(t_A - \tau \| t_A \geq \tau)$ for fixed $P(t_A < \tau)$
Moving average		(7)	L(s-d)	1	min $P(\|\tau - t_A\| \geq d)$
Shewhart		(8)	L(s)	0	min $E(t_A - \tau \| t_A \geq \tau)$ for fixed $P(t_A < \tau)$ asymptotically for large μ and max $P(A(s) \| C)$ for fixed $P(A(s) \| D)$ when $C = \{ \tau = s \}$ and $D = \{ \tau > s \}$

Figure 1



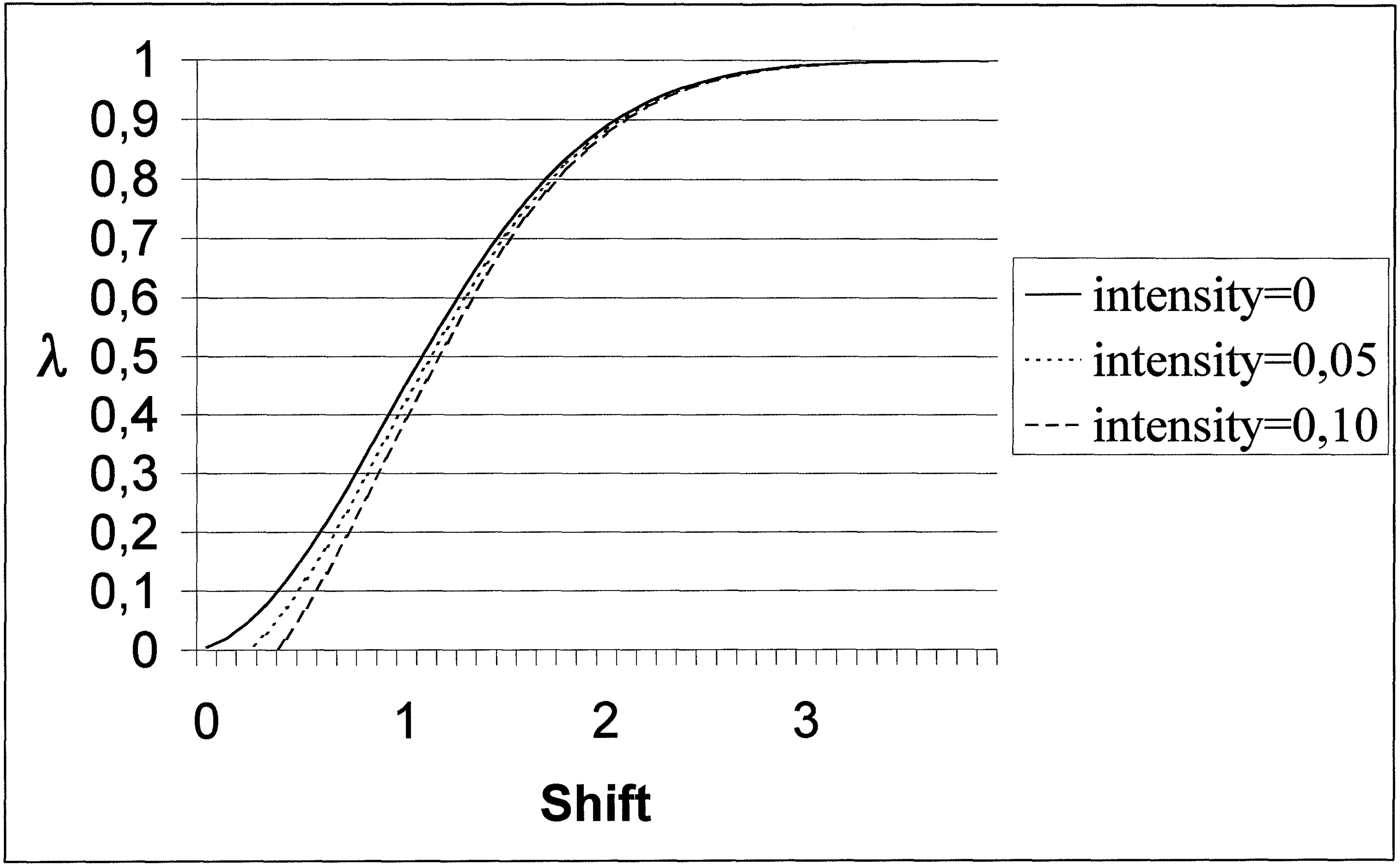
opt2d.xls 1999-08-08

Figure 2



optvikt.xls 1999-08-21

Figure 3



optlambda.xls 1999-08-19

Research Report

1999:1	Andersson, E.:	On monotonicity and early warnings with applications in economics.
1999:2	Wessman, P.:	The surveillance of several processes with different change points.
1999:3	Andersson, E.:	Monotonicity aspects on seasonal adjustment.
1999:4	Andersson, E.:	Monotonicity restrictions used in a system of early warnings applied to monthly economic data.
1999:5	Mantalos. P. & Shukur, G.:	Testing for cointegrating relations- A bootstrap approach.
1999:6	Shukur, G.:	The effect of non-normal error terms on the properties of systemwise RESET test.
1999:7	Järpe, E. & Wessman, P.:	Some power aspects of methods for detecting different shifts in the mean.
1999:8	Johnsson, T.:	On statistics and scientific thinking.
1999:9	Afsarinejad, K.:	Trend-free repeated measurement designs.
1999:10	Carlquist, A. m.fl.	The impact of season and climate on growth during early childhood in different socio--economic groups in Lahore, Pakistan.
1999:11	Carlquist, A, Erling, V. & Frisén, M.:	Longitudinal methods for analysis of the influence of breastfeeding on early child in Pakistan.
1999:12	Carlquist, A.:	Longitudinal methods for analysis of early child health in Pakistan.