UNIVERSITY OF GÖTEBORG

Department of Statistics

RESEARCH REPORT 1994:2
ISSN 0349-8034

CHARACTERIZATION OF METHODS FOR SURVEILLANCE BY OPTIMALITY

by

Marianne Frisén

Statistiska institutionen
Göteborgs Universitet
Viktoriagatan 13
S-411 25 Göteborg
Sweden

CHARACTERIZATION
OF
METHODS FOR SURVEILLANCE
BY
OPTIMALITY.

Marianne Frisén
Department of Statistics, University of Göteborg, S-41125
Göteborg, Sweden

Different criteria of optimality are discussed. The shortcomings of some earlier criteria of optimality are demonstrated by their implications. The correspondences between some criteria of optimality and some methods are examined. The situations under which some commonly used methods have a certain optimality are thus illuminated. Linear approximations of the LR (likelihood ratio) method, which satisfies several criteria of optimality, are presented. These linear approximations are used for comparisons with other linear methods, especially the EWMA (exponentially weighted moving average) method. These comparisons specify the situations for which the linear methods can be regarded as approximations of the LR method.

KEY WORDS: Quality control; Warning system; Control chart; Utility; Performance; EWMA; CUSUM; Moving average.

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. The timeliness of decisions is taken into account in the vast literature on quality control charts where it is often important with simplicity. Also the literature on stopping rules is relevant. 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), Vardeman and Cornell (1987) and Frisén (1994b). 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 in 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) only treats the former case.

In recent years there has 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 were described in i.e the special issue (no. 3, 1989) of "Statistics in Medicine" and by Frisén (1992). Applications in economics and especially the surveillance of business cycles were treated in i.e. the special issue (no. 3/4, 1993) of "Journal of Forecasting" and by Frisén (1994a).

In Section 1 some notations are given and the case studied is specified. In Section 2 some criteria of optimality are described and analyzed. In Section 3 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 some concluding remarks are given.

1. NOTATIONS AND SPECIFICATIONS

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

The critical event of interest at decision time s is denoted C(s). As in most literature on quality control, the case of shift in the mean of Gaussian random variables 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,...,τ-1 and $\mu_t = \mu^1$ for t= τ, τ+1,.... We want to discriminate between

$$C(s) = \{\tau \leq s\} = \{\mu(s)=\mu^1\} \text{ and } D(s) = \{\tau > s\} = \{\mu(s)=\mu^0\}.$$

We will consider different ways to construct alarm sets A(s) with the property that, when X_s belongs to A(s), there is an indication that C(s) occurs.

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 random variable with the density

$$\pi_t = \text{pr}(\tau = t)$$

and $\Sigma\pi_t = 1-\pi_\infty$. The intensity q_t of a change is

$$q_t = pr(\tau = t \mid \tau \geq t)$$

The aim is to discriminate between the states of the system at each decision time s, s=1,2,... by the observation $X_s = \{X(s): t \leq s\}$ under the assumption that $X(1) - \mu_1$, $X(2) - \mu_2$,... are independent normally distributed random variables with mean zero and with the same known standard deviation (say $\sigma=1$). In some calculations below, where no confusion is possible, μ^1 is denoted μ and $\mu^0=0$ and $\sigma=1$ for clarity.

2. OPTIMALITY CRITERIA

The performance of methods for surveillance is dependent on the time τ between the start of the surveillance and the time of the change. Sometimes it is appropriate to express the performance as a function of τ, as in Frisén (1992), Frisén and Åkermo (1993) and Frisén and Cassel (1994). Sometimes, however, a single criterion of optimality is needed. In order to get an index, which is independent of τ, several approaches have been used:

1. In the literature of quality control it is often assumed that the surveillance started at the same time as the change occurred, that is $\tau=0$. See the section on ARL below.

2. Sometimes it is assumed that the surveillance has been started a very long time before a possible change, that is $\tau=\infty$ (Lindgren 1985, Pollak and Siegmund 1991, Srivastava and Wu 1993).

3. A probability distribution of τ is considered and summarizing measures over this distribution are used. See the Sections 2.2 and 2.3.

4. A minimax criterion for the worst possible value of τ is used. See Section 2.4 below.

2.1 ARL

A measure which is often used in quality control is the average run length (ARL) until an alarm. See e.g. Wetherill and Brown (1990). It was suggested already by Page (1954). The average run length, ARL^0, is the average number of runs until an alarm when there is no change in the system under surveillance. The average run length under the alternative hypothesis, ARL^1, is the mean number of decisions that must be taken to detect a true level change (that occurred at the same time as the inspection started). The part of the definition in the parenthesis is seldom spelled out but seems to be generally used in the literature on quality control.

In quality control optimality is often stated as minimal ARL^1 for fixed ARL^0.

Statement 2.1.1. The alarm statistic

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

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

Proof. Both ARL^1 and ARL^0 are expected values under the condition that $\tau=0$. Under this condition and under the specifications in Section 1, the LR method described in Section 3.1 has the alarm statistic in the statement. The LR method has the property of Section 2.2, that for each decision time s it gives the maximal probability of alarm for fixed false alarm probability. The constants c_s can be chosen to match any given false alarm probabilities and thus any given ARL^0. For this fixed value the alarm statistic in the statement gives maximal detection probabilities for all times and thus the minimal ARL^1. □

Thus, methods based on equal weight of all observations satisfy the optimality criterion above. Such methods are not very often used in quality control. Examples of such methods are the simple CUSUM variants described in Section 3.4, where also the drawbacks with these methods are discussed.

Sometimes optimality is defined as minimal ARL^1/ARL^0. The skewness of the run length distributions (especially under the alternative) and other facts makes it easy to construct situations where obviously inferior methods satisfy this criterion. Below the shortcoming of this criterion is illustrated by an example.

Statement 2.1.2 The optimality criterion of minimal ARL^1/ARL^0 has unwanted consequences.

Proof. The often used Shewhart method has the alarm set $A(s) = \{X_s>g\}$. The method has $ARL=1/(1-\Phi(g-\mu))$ and thus a ratio ARL^1/ARL^0 which is monotonically decreasing with g. This consequence is not reasonable. □

2.2 Error probabilities

The problem of finding the method which maximizes the detection probability for a fixed false alarm probability and a fixed decision time was treated by de Maré (1980) and Frisén and de Maré (1991). The LR method of Section 3.1 is the solution to this criterion.

2.3 Utilities

Different kinds of utility functions were discussed by Frisén and de Maré (1991). An important specification of utility is that of Girshick and Rubin (1952) and Shiryaev (1963). They treat the case where the gain of an alarm is a linear function of the difference τ-t_A 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. Their solution to the maximisation of the expected utility is identical to the LR method (with constant limit) of Section 3.1 for the situation specified in Section 1.

2.4 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 problem of minimizing the 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 which would hardly be used in practice. The start of the procedure is made in a way that avoids the properties to be dependent on τ. For most 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 fade off with time.

Moustakides (1986) uses a still more pessimistic criterion by using not only the worst time τ but also the worst possible outcome $X_{\tau-1}$ before the change occurs. The CUSUM method below is (except for the first time point) the solution to the criterion posed by Moustakides.

Ritov (1990) considers a loss function which is not identical to that of Shiryaev but depends on τ and t_A in addition to the dependency on τ - t_A. The worst possible distribution

$$\Pr(\tau=s+1 \mid \tau>s;X_s)$$

is assumed for each time s. With this assumption of a worst possible distribution (based on earlier observations) CUSUM minimizes the loss function.

2.5 Successful detection within a time limit

In some applications there is a limited time available for rescuing 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 2.3 and 2.4,

the probability that the difference does not exceed a fixed limit is used. The fixed limit, say d, is the time available for successful detection. This probability (as a function of τ) was suggested by Frisén (1992) as a measure, PSD, of the performance. Bojdecki (1979) considered a criterion which is equivalent to the maximum of the minimum (with respect to τ) of

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

See Section 3.6 for discussion of consequences of this optimality criterion.

2.6 Predictive value and posterior distribution

The posterior distribution PD(s) = pr(C(s) | X_s) has been suggested as an alarm criterion by e.g. Smith et al (1983). Frisén and de Maré (1991) demonstrated that, when there are only two states C and D, this criterion leads to the LR method of Section 3.1.

The predictive value PV(s) = pr(C(s) | A(s)) has been used as a criterion of evaluation by Frisén (1992), Frisén and Åkermo (1993) and Frisén and Cassel (1994).

The relation between the PV and the PD functions will now be analyzed.

Statement 2.6.1. At passive surveillance, that is when our actions at an earlier time point do not affect the distributions, we have :
A method based on PD, that is A(s)= [X_s; PD(s)>c] implies
PV(s) > c. Typically PV increases to one when s increases. □

Statement 2.6.2. At active surveillance , when the whole process will be stopped as soon as an alarm occurs, none of Statement 2.6.1 holds: Typically PV has an asymptote below one. PV is not monotonically increasing for all methods (not for CUSUM). □

At active surveillance (contrary to passive) it is desirable (for many applications) to be able to take the same action whenever an alarm occurs. In those cases a constant PV would be a good property. Another distinction is that between a single decision and a sequence of decisions. At a single decision, alarm for PD>c or (when there is no prior) significance at an ordinary test is natural. For a sequence of decisions, characteristics of the sequence (such as constant PV or the expected waiting time to an alarm) become interesting.

3. METHODS

3.1 The likelihood ratio method

A method constructed by Frisén and de Maré (1991) to meet several optimality criteria, i.e. those of Section 2.2 and 2.3, will first be presented. The general method uses combinations of likelihood ratios. Even though 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 "bench-mark". Commonly used methods are compared to it in order to clarify their optimality properties.

Here, the method of Frisén and de Maré (1991) is applied to the shift case specified in Section 1. The "catastrophe" to be detected at decision time s is $C = \{ \tau \leq s\}$ and the alternative is $D = \{ \tau > s\}$. The method for this case will here be called the likelihood ratio method or shorter the LR method.

The LR 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 normal distribution specified in Section 1 we have

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

where

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

and

$$p_s(x_s) = \sum_{k=1}^{s} \pi_k \exp\left\{\frac{1}{2} k(\mu^1)^2\right\} \exp\{\mu^1 \sum_{u=k}^{s} x(u)\}$$

which is a nonlinear function of the observations.

In order to achieve the optimal error probabilities described in Section 2.2 an alarm should be given as soon as $p(x_s) > G_s$.

In order also to achieve maximization of the utilities mentioned in Section 2.3 it is required that $G_s \equiv G$ and we must also consider the function g(s).

In Figures 1 and 6 the LR method is illustrated for s=2.

3.2 Linear approximation of the likelihood ratio method

To get a method which is easier to use, and also to clarify the connection with other methods, a linear approximation of p_s is of interest. The exponential functions of the x-values will be approximated by linear functions. The aim is to get a good approximation of the limit for alarm. Thus Taylor expressions around

values which might cause an alarm are used. Such values will approximately satisfy

$$\frac{\sum_{u=k}^{s} x(u)}{\sqrt{\sigma^2(s-k+1)}} = z$$

where σ here is set to one. By using

$$\exp\left\{\mu[\sum_{u=k}^{s} x(u) - z\sqrt{s-k+1}]\right\} \approx 1 + \mu[\sum_{u=k}^{s} x(u) - z\sqrt{s-k+1}]$$

the following linear approximation is achieved:

$$p_s(x_s) \approx p_s^*(x_s) = c + \sum_{k=1}^{s} \pi_k a(k) \mu \sum_{u=k}^{s} x(u) = c + \mu \sum_{u=1}^{s} x(u) m(u),$$

where c does not depend on the data,

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

and

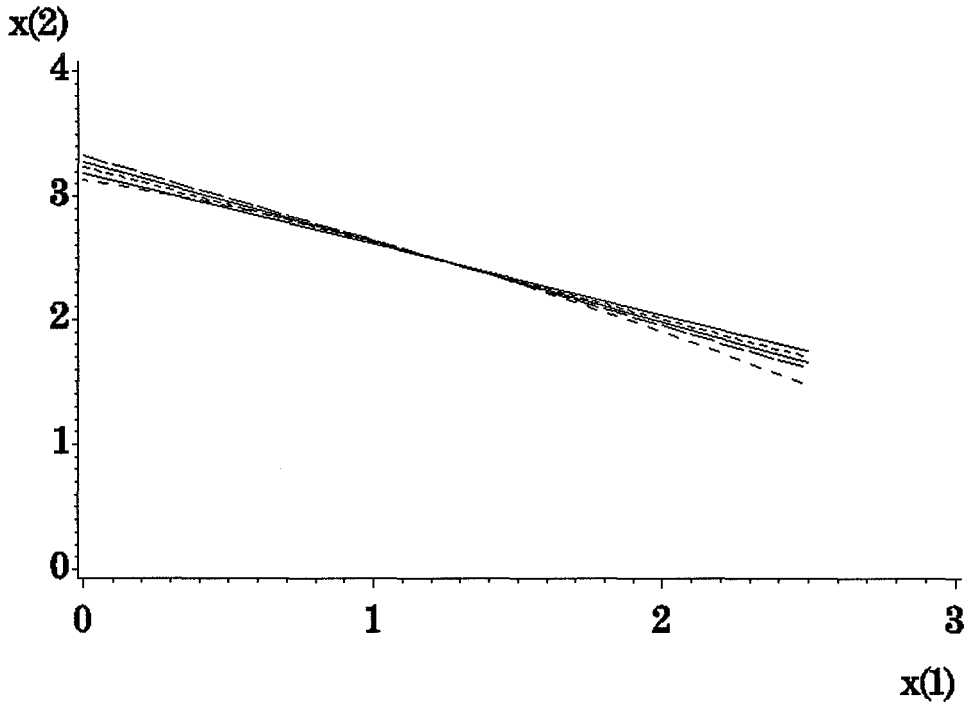
$$a(i) = \exp\{i\mu^2/2 + \mu z\sqrt{s-i+1}\}.$$

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

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

exceeds a limit.The value of z which gives the best approximation depends on how tight the limit for alarm is. The approximation is illustrated in Figure 1, for different values of z, for a case of a rather wide alarm limit. As can be seen, the approximation is not very sensitive to the value of z. In all illustrations below $\mu^0=0$, $\mu^1=\mu=1$ and $q_t=q=0.01$.

Figure 1. Linear approximations of the LR limits - - -.
The approximations are made with z=3 ---, z=3.5 —— and z=4 — —.



If the intensity is constant τ has a geometric distribution $\pi_k=(1\text{-}q)^{k-1}q$. Then, with

$$b=(1-q)\exp\{\mu^2/2\}$$

the weights m(u) are

$$m(u)=q/(1-q)\sum_{i=1}^{u} b^i e^{\mu z\sqrt{s-i+1}}$$

The weight of x(u) is thus increasing with u. Later observations thus have a greater weight than older ones.

Figure 2. The weights m(u) of the linear approximation LLR with z=3.5 —— and of the EWMA method - - - at the decision time s=10.

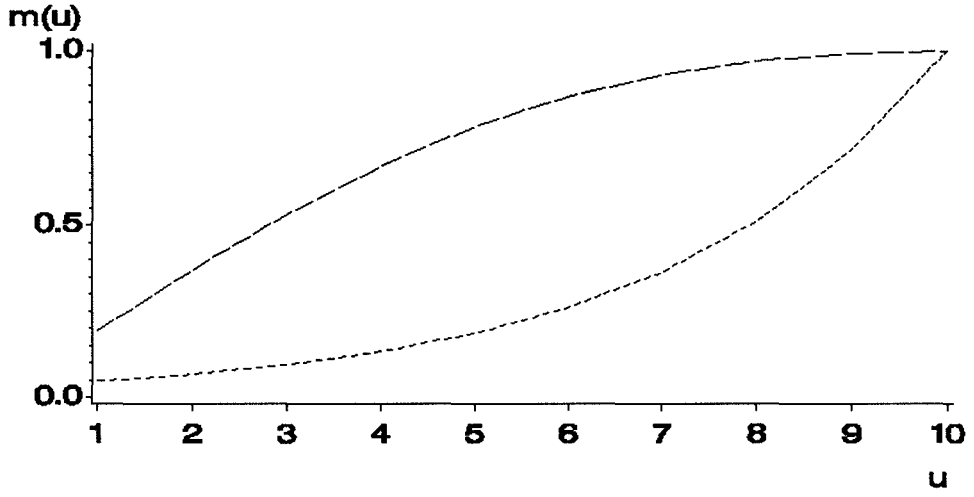


Figure 3. As Figure 2 but s=30.

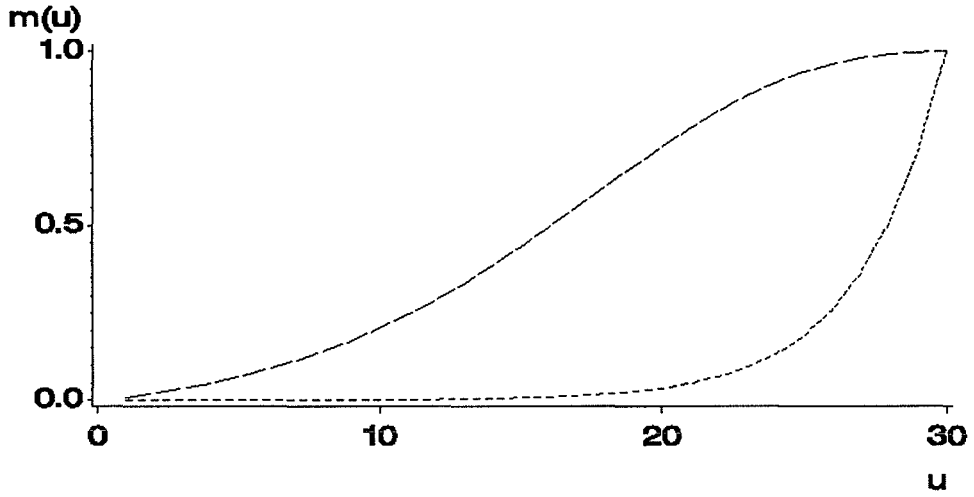
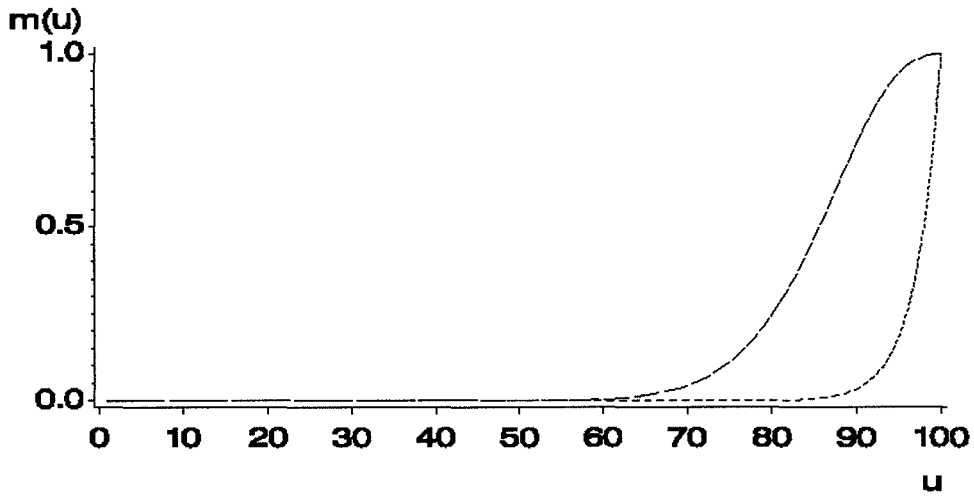


Figure 4. As Figure 2 but s=100.



With the approximation above, the relative weights depend on the decision time s, as was illustrated by Figures 2, 3 and 4. Some commonly used methods are linear but with weights which are independent of s. Thus, a further approximation of the LR method is made to get weights which are independent of s. In the figures above the case of wide limits for alarm was illustrated. If tight limits (which imply short run lengths) are used it might be reasonable to use z=0, that is

$$\exp\left\{\mu \sum_{u=k}^{s} x(u)\right\} \approx 1 + \mu \sum_{u=k}^{s} x(u).$$

For this LLR(0) method we have the weights

$$m(u) = \sum_{i=1}^{u} a^i \pi_i,$$

where

$$a = \exp\{\mu^2/2\}.$$

If the intensity is constant

$$m(u) = q/(1-q) \sum_{i=1}^{u} b^i = qb(b^u - 1)/(b-1)(q-1) \doteq b^u - 1.$$

3.3 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 it has got much attention. This may be due to papers by Robinson and Ho (1978), Crowder (1987), Lucas and Saccucci (1990), Ng and Chase (1989) and Domangue and Patch (1991) in which positive reports of the quality of the method are given. Also the paper by Hunter (1986), where simple interpretations of the method by its relation to forecastings are given, has drawn the attention to EWMA.

The statistic is

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

where $0<\lambda<1$ and in the standard version of the method $Z_0 = \mu^0$. The statistic is sometimes referred to as a geometric moving average since it can equivalently be written as

$$Z_s = \lambda\sum_{j=0}^{s-1}(1-\lambda)^j x(s-j) + (1-\lambda)^s Z_0 = \lambda(1-\lambda)^s\sum_{u=1}^{s}(1-\lambda)^{-u}x(u) \propto \sum_{u=1}^{s}k^u x(u)$$

The weights are thus k^u, 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 of the standard deviation.

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 test is a Shewhart test. If λ is near zero all observations have approximately the same weight.

Also other variants of EWMA have been proposed. See Frisén and Åkermo (1993) for a discussion of some variants and for a comparison with CUSUM. In the present study only the standard variant described above will be discussed.

Statement 3.3.1 There does thus not exist any λ or L which makes the EWMA exactly optimal in the sense of Sections 2.2 or 2.3.

Proof. The likelihood method 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. □

In Figure 6 some methods are illustrated for the first two observations. Since the EWMA has two parameters, λ and L, these can be chosen to equal any other linear method when only two

observations are illustrated. When more than two steps are considered this is not true.

Statement 3.3.2 The weights of the EWMA cannot be identified with the weights of the LLR(z) method.

Proof. The weights are dependent on s for the LLR(z) method but not for the EWMA. □

Statement 3.3.3 The weights of the EWMA cannot be identified with the weights of the LLR(0) method for the case of constant intensity.

Proof. At constant intensity q

$$\pi_i = (1-q)^{i-1}q \quad i=1,2,..$$

The weights m(u) of the LLR method are found in Section 3.2. The relative weights are

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

The relative weights are thus not constant for the LLR method as they are for the EWMA method. □

In the more general case one might ask which series of intensities would make an identification between the EWMA and the LLR(0) possible.

Statement 3.3.4 For each combination of μ^1, q_1 and π_∞ there is one and only one series of intensities that makes identification between EWMA and LLR possible.

Proof The identification implies that

$$m(u+1) = km(u)$$

where $k=1/(1-\lambda)>1$ is the constant which determines the weights for the EWMA. At u=1

$$m(2) = km(1)$$

$$a\pi_1 + a^2\pi_2 = ka\pi_1$$

$$(1) \quad \pi_2 = \pi_1(k-1)/a$$

At u>1

$$m(u+1) = m(u)$$

$$m(u) + a^{u+1}\pi_{u+1} = k[m(u-1) + a^u\pi_u]$$

$$(2) \quad \pi_{u+1} = c^{u-1}\pi_2$$

where

$$c=k/a.$$

The requirements (1) and (2) determine the series of intensities for each k. The value of k is uniquely determined by

$$\sum_{i=1}^{\infty}\pi_i = 1-\pi_\infty = \pi_1 + \pi_1(k-1)/a + \sum_{i=3}^{\infty}c^{i-2}\pi_1(k-1)/a = \pi_1(a-1)/(a-k)$$

which gives

$$k = a - \pi_1(a-1)/(1-\pi_\infty)$$

which in turn implies

$$\pi_2 = \pi_1(a-1)[1 - \pi_1/(1-\pi_\infty)]/a.$$

It follows that k>1 and the series π_t satisfies the requirements of probabilities. □

Corollary. In the case of $\pi_\infty = 0$ it follows that

$$k = a(1-\pi_1) + \pi_1$$

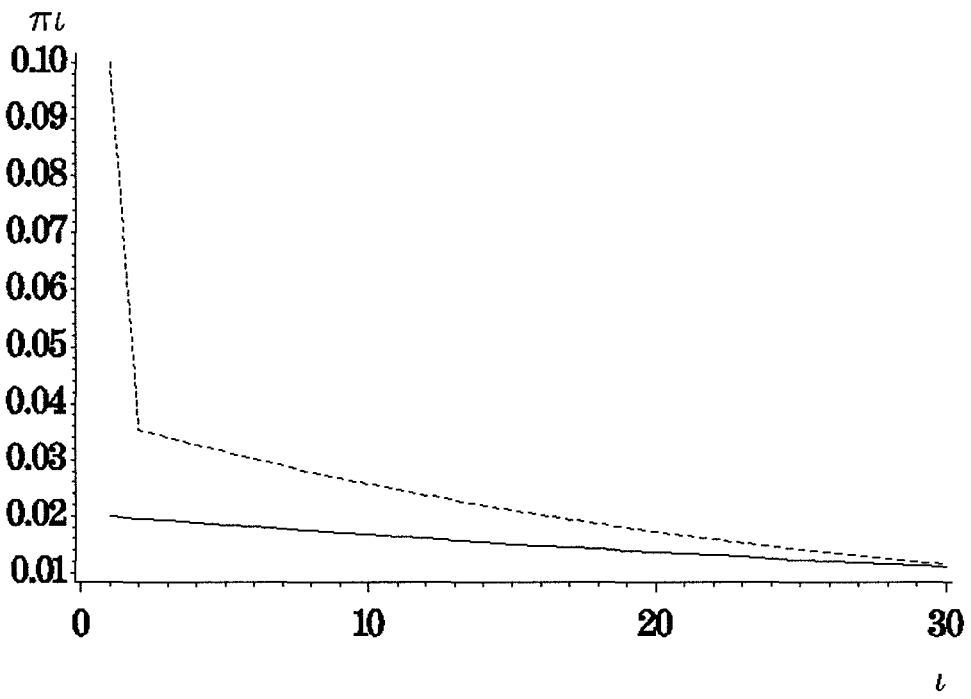
and the series of intensities is determined by

$$\pi_2 = \pi_1(1 - 1/a)(1 - \pi_1),$$

formula (2) above and

$c = 1-\pi_1 + \pi_1/a$. □

Figure 5. An example of values of π_i which makes identification between the EWMA and LLR(0) possible. As comparison the solid line is given. It represents a case of constant intensity.



The LLR methods are approximations of the LR method which has the optimality of Section 2.2. If, in addition, a constant limit (not depending on s) for the alarm statistic is used also the optimality of Section 2.3 is satisfied.

Statement 3.3.4 The EWMA method can never be identified with the LLR(0) method, with a limit which does not depend on s, as required for the optimality of Section 2.3.

Proof. The weights of the EWMA do not depend on s. For the usual version studied here, also the limit of the linear expression for alarm is independent of s. For the LR method with constant limit we have an alarm when

$$p(x_s)=g(s)p_s(x_s)>G$$

where

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

When p_s is approximated by LLR(0) the weights m(u) are independent of s but the limit is G/g(s). This limit is decreasing with s, as g(s) is increasing with s. □

3.4 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

$$C_t=\sum_{i=1}^{t}(X_i-\mu^0)$$

exceeds a fixed limit. This method is sometimes called the simple CUSUM. It will here be denoted as SCUSUM. For each t the likelihood ratio is a function of C_t only. As was demonstrated by Frisén and de Maré (1991), the SCUSUM is optimal in the sense of Section 2.2 for $\tau=0$ in the normal cased specified in Section 1. By Statement 2.1.1 it was seen that the SCUSUM minimizes the ARL^1 for fixed ARL^0. However, when $\tau>0$ it was demonstrated by Frisén (1992) that SCUSUM does not compare with other methods with the same ARL. The probability of successful detection within a short time is lower. Also, the predicted value of an alarm is strongly decreasing with the time of the alarm.

Another simple method based on cumulative sums gives an alarm as soon as C_t exceeds a linear function of t. This method is here called the LCUSUM method This method is identical with the method which gives an alarm when the likelihood ratio for $\tau=0$ exceeds a fixed constant. It is a sequential probability ratio test without the limit for acceptance. In Figure 6, 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.

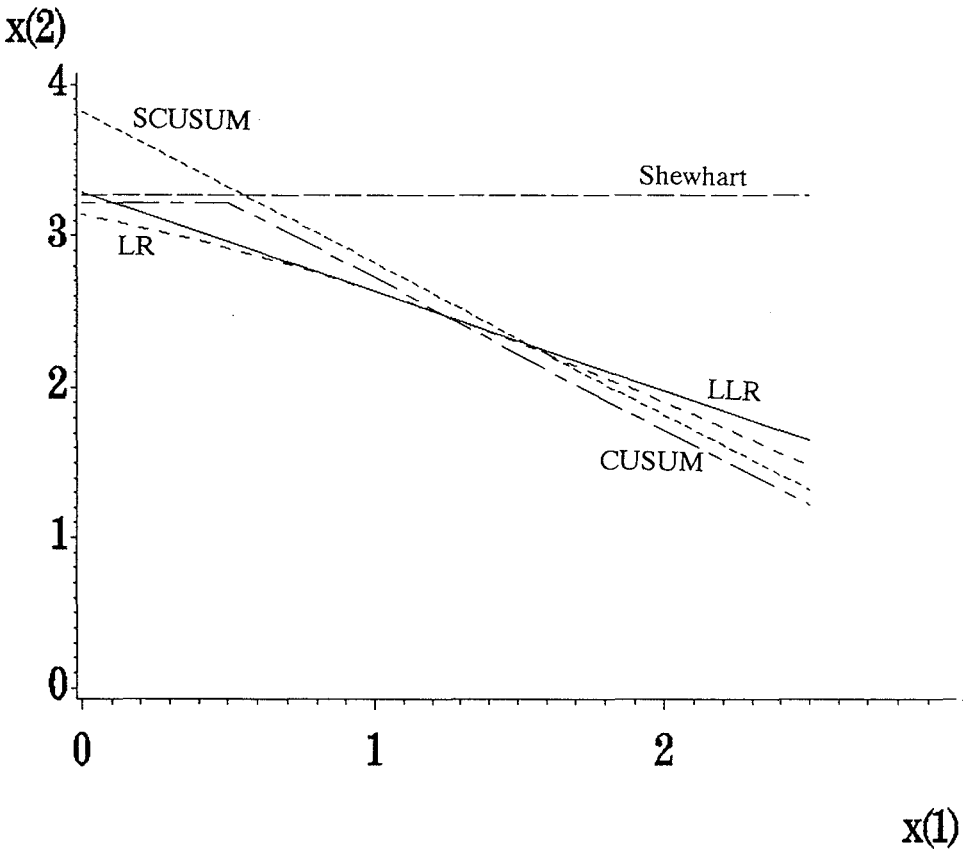
In both the SCUSUM and the LCUSUM, the data from all earlier points in the time series have the same weights as the last one. For most applications this is not considered rational. Anyhow, as soon as only $\tau=0$ is considered (as in the criterion that minimizes the ARL^1 for fixed ARL^0) these weights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

3.5 CUSUM

The variant of cusum tests which is most often advocated is 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 t for which $| C_t-C_{t-i} | > h + ki$ (for some i=1,2,... t), where $C_0=0$ and h and k are chosen constants. The parameter k determines the slopes of the legs in the "V" 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" 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 observations have a heavy weight and there are similarities to the Shewhart test. In this test the information from

earlier observations is handled quite differently depending on the position in the time series.

Figure 6. Alarm limits at decision time s=2.



Sometimes (see 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. It was demonstrated by Frisén and de Maré (1991) that the CUSUM is the result of a natural (but not optimal) combination of methods, where each of them is optimal to detect a change that occurs at a specific time point.

It is often stated that the choice of $k=(\mu^0+\mu^1)/2$ is optimal. The chain of references (if any) usually ends with Ewan and Kemp (1960). In that paper they conclude from nomograms that this value seems to be about the best. The optimal likelihood ratio method for $\tau=i$ and with constant limit $G_s=G$ gives alarm for

$$\sum_{t=i}^{s} X_t > c + i(\mu^0+\mu^1)/2.$$

Thus also here we have the slope $(\mu^0+\mu^1)/2$. 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 satisfies certain minimax conditions (Moustakides 1986 and Ritov 1990) as was discussed in Section 2.4 above. In Basseville and Benveniste (1986 p 18) it is stated that the CUSUM method have the optimality property of Section 2.3. However, this is true only under specific conditions. See Section 2.4.

3.6 Moving average

The moving average C_t-C_{t-d} for fixed window width d is compared with a fixed alarm limit. It can be shown to be a special case of the solution of Bojdecki (1979) to a maximization of

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

where t_A is the time of alarm. See Section 2.5 for discussion on this optimality criterion.

4. CONCLUDING REMARKS

The performance depends on the time of the change τ, as was demonstrated by the evaluations by Frisén (1992). To get 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 $\tau=0$, $\tau=\infty$ or $\tau=$ "worst possible value". In Roberts (1959 and 1966) the value $\tau=8$ was used, but that was because of technical

reasons. In quality control, optimality criteria based on ARL, which implies $\tau=0$, is the common choice. Sometimes the criterion is expressed as the ratio ARL^1/ARL^0. As was noted in Statement 2.1.2, this has unreasonable implications, such as "the greater limit for the Shewhart method the better". More often the criterion is stated as minimal ARL^1 for a fixed ARL^0. As was noted in Statement 2.1.1 this criterion implies methods where all observations have the same weight. The shortcomings of such methods were pointed out in Section 3.4 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. The solution to an optimal criterion based on $\tau=$ "worst possible value" is a randomized procedure. Later suggestions are to make the minimax criterion still more pessimistic by also assuming the worst possible outcome.

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 for special values of τ, demonstrate the importance of any information on the distribution of τ. Several criteria of this type result in the LR method. The error probabilities in each step are optimal for any limits. To achieve a minimum expected delay until an alarm, it is also necessary that the limits are independent of time.

Criteria based on the posterior distribution have an intricate relation both to the LR method and to the predicted value of an alarm. These relations were analyzed in Section 2.6 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(z) has relative weights which are dependent on the decision time s. Thus the linear method EWMA, which lacks this dependency, cannot be identified with the LLR(z) method.

For a further approximation to LLR(0) the identification is possible. For a specified choice of parameters, EWMA will be approximately optimal with respect to error probabilities in each step. However, this is possible only for a decreasing series of intensities. Especially the intensity in the first point must be great. The result that the EWMA method has good properties, only if the probability of a change is greatest in the beginning, is in accordance with the results in Frisén and Åkermo (1993) based on the predicted value.

Identification between the EWMA and the LLR(0) with constant limit, which is the requirement for minimum expected waiting time until an alarm, is not possible. The EWMA is too generous with alarms in the beginning. The suggestion in the literature of a variant of EWMA which is intended to give a fast initial response (FIR EWMA) by closer limits in the beginning would do this worse.

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 $\tau=0$. They are linear, but with equal weight of all observations in contrast to the linear approximations of the LR method which give more weight to later observations. .

ACKNOWLEDGEMENT

This work has been supported by the Swedish Council for Research in the Humanities and Social Sciences.

REFERENCES

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. & 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.

Domangue, R. and Patch, S. C. (1991), "Some omnibus exponentially weighted moving average statistical process monitoring schemes," *Technometrics,* 33, 299-313.

Ewan W. D. & Kemp K. W. (1960) "Sampling Inspection of Continous Processes with no Autocorrelation between Successive Result." *Biometrika*, 47, 363-.

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

Frisén M. (1994a) "Statistical Surveillance of Business Cycles." Research report, 1994:1,

Frisén M. (1994b) "A classified bibliography on statistical surveillance." Research report, 1994,

Frisén M. and Cassel C. (1994) "Visual evaluations of statistical surveillance." Research report, 1994:3,

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

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

Hunter, J. (1986), "The Exponentially Weighted Moving Average," *Journal of Quality Technology*, 18, 203-210.

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

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

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.

Moustakides, G. V. (1986), "Optimal stopping times for detecting changes in distributions," *Annals of Statistics* 14, 1379-87.

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.

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

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.

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.

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*, 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.

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

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

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.

1992:1	Teräsvirta, T., Tjøstheim, D. & Granger, C.W.J.	Aspects of Modelling Nonlinear Time Series
1992:2	Palaszewski, B.	A conditional stepwise test for deviating parameters
1992::3	Guilbaud, O.	Exact Semiparametric Inference About the Within--Subject Variability in 2 x 2 Crossover Trails
1992:4	Svensson, E. & Holm, S.	Separation of systematic and random errors in ordinal rating scales
1993:1	Frisén, M & Åkermo, G.	Comparison between two methods of surveillance: exponentially weighted moving average vs cusum
1993:2	Jonsson, R.	Exact properties of McNemar's test in small samples.
1993:3	Gellerstedt, M.	Resampling procedures in linear models.
1994:1	Frisén, M.	Statistical surveillance of business cycles.