International Statistical Review (2003), 71, 2, 403–434, Printed in The Netherlands
© International Statistical Institute

Statistical Surveillance. Optimality and Methods

Marianne Frisén

Department of Statistics, Göteborg University, Box 660, SE-40530 Göteborg, Sweden

Summary

Different criteria of optimality are used in different subcultures of statistical surveillance. One aim with this review is to bridge the gap between the different areas. The shortcomings of some criteria of optimality are demonstrated by their implications. Some commonly used methods are examined in detail, with respect to optimality. The examination is made for a standard situation in order to focus on the inferential principles. A uniform presentation of methods, by expressions of likelihood ratios, facilitates the comparisons between methods. The correspondences between criteria of optimality and methods are examined. The situations and parameter values for which some commonly used methods have optimality properties are thus determined. A linear approximation of the full likelihood ratio method, which satisfies several criteria of optimality, is presented. This linear approximation is used to examine when linear methods are approximately optimal. Methods for complicated situations are reviewed with respect to optimality and robustness.

Key words: Change-point; Control chart; CUSUM; EWMA; Likelihood ratio; Monitoring; Quality control; Repeated decisions; Shewhart; Statistical process control; Stopping rule.

1 Introduction

There is often a need for 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 methods with this goal. An example is the surveillance of the foetal heart during labor described by Fris én (1992). An abnormality, caused by e.g. a lack of oxygen due to the umbilical cord around the neck of the foetus might happen at any time. Detection has to be made as soon as possible after the event has occurred to ensure that a rescuing action, such as a Cesarean section, is of value.

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 are described in e.g. the special issue (no. 3, 1989) of "Statistics in Medicine". Monitoring for detection of changes in public health is described by e.g. Williamson & Hudson (1999) and Sonesson & Bock (2003). Methods for post marketing surveillance of adverse effects of drugs are described by e.g. Lao *et al.* (1998). Needs for environmental control are described in the book edited by Barnett & Turkman (1993). Surveillance technique is used for environmental monitoring by Pettersson (1998b). 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". Comments on the role of statistical quality control in industry are given in the paper by Banks (1993) and the connected discussion. Yashchin (1993) discusses the relation between "Engineering Process Control" where the corrective formula is important and "Statistical Process Control" where the detection of the abrupt change is the main

aim.

In applied work a single optimality criterion is not always enough but evaluations of different properties might be necessary (Frisén, 1992). However, optimality plays an important role both in applied work and for theory. There are many papers which claim to give the optimal method of surveillance. However, the suggested optimality criteria differ in important aspects. Most of the 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 methods can all be described by different likelihood ratios. The explanations will be given in the text. The focus of the paper is on inferential matters. A complete review of the area of statistical surveillance cannot be made within one paper.

In some applications the whole process will be stopped as soon as an alarm occurs. An example is the surveillance of the foetal heart during labor mentioned above. This is called active surveillance in contrast to passive surveillance, where our actions at an earlier time point do not affect the process (Frisén & de Maré, 1991). This can be the case in flood warning systems when alarms do not affect the level of the water. Most of the discussion in this paper concerns active surveillance, but the differences with respect to stochastic properties between the active and passive surveillance will be pointed out.

The timeliness and also the simplicity of procedures is considered 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. In cases where several changes may follow after each other, the process might be characterized as a hidden Markov chain and the posterior probability for a certain state determined (e.g. Harrison & Stevens (1976) and Hamilton (1989)). Estimation of the time of change (e.g. Hinkley (1970) and Gombay (2000)) is not discussed here.

The theory of statistical surveillance has developed independently in different statistical subcultures. Thus, the terminology is diverse. Different names are used to refer to "statistical surveillance" as described here. However, there are some differences in how the words are used. "Optimal stopping" is most often used in probability theory, especially in connection with financial problems, and does not always include the statistical inference from observations to the model. Literature on "Change point problems" does not always treat the case of continual observations but often the case of a retrospective analysis of a fixed number of observations. The term "early warning system" is most often used in economical literature. "Monitoring" is most often used in medical literature and as a nonspecific term. "Statistical process control" and "quality control" are used in the literature on industrial production and are also used to include nonstatistical aspects.

Some broad surveys and bibliographies on methods for statistical surveillance are found in e.g. Zacks (1983), Vardeman & Cornell (1987), Basseville & Nikiforov (1993), 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 further developed. Also the book by Brodsky & Darkhovsky (1993) on nonparametric methods for change-point problems is in the same spirit. A collection of papers on change-point problems is edited by Carlstein *et al.* (1994). A survey of statistical process control (SPC) is given by Woodall & Montgomery (1999). In this survey it is pointed out that a cross-fertilization between SPC and the mathematical statistical literature on e.g. change-point analysis would be fruitful. In Crowder *et al.* (1997) it is stated: "There are few areas of statistical application with a wider gap between methodological development and application than is seen in SPC." In a short review on SPC, Stoumbos *et al.* (2000) state a need for a greater synthesis of the theoretical change point and applied SPS literature. The bibliographies mentioned above treat both the case of a fixed period and the case of sequential observation. The survey by James *et al.* (1987) only treats the fixed period case. In the following detection of a single change will be discussed and only the case of sequential observations will be treated.

In Section 2 the notation is described. Also, a description is made of the most commonly discussed case in the literature, that of a shift in the mean of a normal distribution. This case is used to derive the connections between methods and optimality. In Section 3 some general criteria of optimality are described and analyzed. In Section 4 general methods are described and compared with those derived from the optimality criteria. Thus, the commonly used methods are characterized by their optimality properties. We start in Section 4.1 with the full likelihood ratio method, LR, which corresponds to the use of the posterior distribution and fulfils important optimality criteria. This method, which relies on generally accepted principles of inference, can then be a benchmark for other methods. In Section 4.1.2 linear approximations of the LR method are derived. The approximations are used in Section 4.2 to determine the approximate optimality of the exponential moving average method, EWMA, and also to discuss for which situation EWMA will be a suitable method. Different variants of CUSUM methods are analyzed in Sections 4.3 and 4.4 with respect to their optimality. In Section 5 there is a description of methods and a discussion of optimality for some more complicated situations, like multivariate surveillance, non-parametric surveillance, more complicated models and more complicated changes. Section 6 contains some concluding remarks.

2 Notations and Specifications

The process under surveillance is denoted by $X = \{X(t) : t = 1, 2, \ldots\}$, 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 denoted by $\mu(t), t = 1, 2, \ldots$. To demonstrate some features, a simple specific situation is used in most of Sections 3 and 4, while more complicated situations are treated in Section 5. This standard case will now be described.

As in most literature on quality control, the standard 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. 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 $t = 1, \ldots, \tau - 1$ and $\mu(t) = \mu^1$ for $t = \tau, \tau + 1, \ldots$. For each decision time $s, s = 1, 2, \ldots$ we want to discriminate between the two events $C(s) = \{\tau \le s\}$ 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 $X_s = \{X(t) : t \le s\}$ is a subset of $A(s)$, there is an indication that the event $C(s)$ has occurred. The time of the first alarm is $t_A = \min\{s : X_s \in A(s)\}$.

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 random variable with the probabilities $\pi_t = P(\tau = t)$. These probabilities can also be regarded as priors. The intensity, ν_t, of a change is

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

The aim is to discriminate between the states of the system at each decision time $s, s = 1, 2, \ldots$ by the set of observations $X_s = \{X(s) : t \le s\}$ under the assumption that $X(1) - \mu(1), X(2) - \mu(2), \ldots$ are independent normally distributed random variables with mean zero and with the same known standard deviation σ. For clarity, standardization to $\mu^0 = 0$ and $\sigma = 1$ is used and the size of the shift after standardization is denoted by μ. The case $\mu > 0$ is described here. The case $\mu < 0$ is treated in the same way. Two-sided procedures will be discussed in Section 5.1.1.

3 Optimality Criteria

In this section general criteria are discussed and illustrated by the standard case of Section 2 with specified values of μ^0 and μ^1. In Section 5 the special problems with optimality for multidimensional

and other complicated situations are discussed.

The performance of a method for surveillance depends on the time τ of the change. Alarm probabilities will in general 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 by Frisén (1992) and Frisén & Wessman (1999). 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. Averaging measures with respect to the distribution of τ can be used when assumption on this distribution is available. Error probabilities are described in Section 3.1 and expectations and utilities are described in Section 3.2.
2. The worst possible value of τ is used by the minimax criterion (Section 3.3).
3. The situation when $\tau = 1$ is often studied in the literature on quality control. This is the situation when the change occurred at the same time as the surveillance started. The approach is discussed in Section 3.5 on ARL.
4. Sometimes it is assumed that the surveillance has been started a very long time before a possible change (e.g. Lindgren (1985), Pollak & Siegmund (1991), Srivastava & Wu (1993)). In that case the asymptotic results when τ tends to infinity are relevant. Since many interesting problems on the influence of the distribution of τ are avoided, this approach is not much discussed below.

From the perspective of optimal decisions costs are given to the different errors which can be made and a utility function is maximized. This is the motivation for the criterion of expected delay treated in Section 3.2.

3.1 *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 $C(s) = \{\tau \leq s\}$ and $D(s) = \{\tau > s\}$. 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 (1977) and Frisén & de Maré (1991).

A constant probability of exceeding the alarm limit at each s means that we have a system of repeated significance tests. This might work well also as a system of surveillance and is often used. The Shewhart method of Section 4.6 has this property. This is also the motivation for using the limits with the exact variance in the EWMA method of Section 4.2 and a variant of the CUSUM method of Section 4.4 given by Brown *et al.* (1975). However, the probability of exceeding the alarm limit conditionally on no earlier alarm is not constant for these methods. Evaluation by the significance level and power of the (repeated) test is often used, especially in the econometric literature, even when the aim obviously is on-line detection of a shift in sequentially obtained data.

Chu *et al.* (1996) advocate monitoring methods which have a fixed (asymptotic) probability of any false alarm during an infinitely long surveillance period. For some applications this might be important because a strict significance test is the goal. In that case, ordinary statements for hypotheses testing about e.g. size and power can be made. However, the price for this additional feature is high as the expected delay of the detection of a change will be very large as pointed out by Pollak & Siegmund (1975).

3.2 *Expected Delay*

From the perspective of optimal decision a utility function with specified costs is of interest. Then, the expected delay of a desired alarm plays an important role.

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

$$\mathrm{ED}(i) = E[\max(0, t_A - i)|\tau = i]$$

The $\mathrm{ED}(i)$ will typically tend to zero as i increases. The conditional expected delay

$$\mathrm{CED}(i) = E[t_A - i|\tau = i, t_A \geq i] = \mathrm{ED}(i)/P(t_A \geq i)$$

on the other hand, will for most methods converge to a constant value. This value is sometimes named the "steady state ARL". The summarized expected delay is

$$\mathrm{ED} = E[\mathrm{ED}(\tau)],$$

where the expectation is with regard to the distribution of τ.

An important specification of utility is that of Girshick & Rubin (1952), who motivate the utility function from incomes and costs at industrial production, and Shiryaev (1963). They treat the case of constant intensity of a change where the gain of an alarm is a linear function of the expected value of the delay, $t_A - \tau$. The loss associated with a false alarm is a function of the same difference. This utility can be expressed as $U = E\{u(\tau, t_A)\}$, where

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

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

$$U = b\,P(\tau > t_A) + a_1\mathrm{ED} + a_2.$$

Thus, we would have a maximal utility if we have a minimal (a_1 is typically negative) expected delay from the change-point for a fixed false alarm rate. The criterion will be named "the criterion of minimal expected delay" or "the ED criterion", for short. The full likelihood ratio method LR (Section 4.1) satisfies this criterion.

3.3 *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 delay, for the worst value of τ. The solution is a randomized procedure. The start of the procedure is made in such a way that it avoids the properties being 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 4.4, 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 (1963) but depends on τ and t_A besides $t_A - \tau$. 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.

Asymptotic minimax optimality is the optimality criterion in much of the theoretical literature on stopping rules as in e.g. Lai (1995), Lai (1998) and Lai & Shan (1999). Results on the order of the convergence of the minimax value to its lower bound have been given for some methods by Yakir *et al.* (1999).

3.4 Other Evaluations

Optimality criteria are useful, but sometimes a single criterion is not enough for the evaluation. Margavio *et al.* (1995), Woodall & Montgomery (1999) and Carlyle *et al.* (2000) state that the use of the ARL criterion is usually recommended despite the knowledge that the run length distribution is poorly reflected by this measure. Margavio *et al.* (1995) suggest that the whole distribution of the alarm time should be used. The time dependent alarm limit should be utilized to give the desired distribution. Then special properties such as fast initial alarms could be designed. However, distributions for each value of τ would be necessary to get all information. Some examples of derived evaluation functions will be given below.

3.4.1 Descriptions of the delay of an alarm

Differences in shapes of CED(t) curves, for different methods, as illustrated by Frisén & Wessman (1999) motivate descriptions of those curves.

When the distribution of τ is geometrical with the intensity ν, it is sometimes useful to express the expected delay for a method as a function of ν as in Frisén & Wessman (1999).

In some applications, such as intensive medical care (Frisén, 1992) there is a limited time available for rescue actions. Then, the expected value of the difference $\tau - t_A$ is not of main interest. Instead of using the expected value as in Section 3.2 and 3.3, the probability that the difference does not exceed a preassigned limit is used. The limit, say d, is the time available for successful detection. The probability of successful detection

$$\mathrm{PSD}(\tau, d) = P(t_A - \tau \le d | t_A \ge \tau),$$

was used by Frisén (1992) and Frisén & Wessman (1999). Bojdecki (1979) considers the supremum (with respect to τ) of

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

Symmetrical measures around τ can be relevant, e.g. when the aim is to make an alarm as close as possible to a turning point in an economical index (Andersson, 2002).

3.4.2 Predictive value of an alarm

The predictive value $\mathrm{PV}(s) = P(C(s)|A(s))$ 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 anywhere in the past is, when we have an alarm. This is informative about the ongoing process and gives a prediction of what will happen if no action is taken. Thus, it gives important information about which action 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 probability $\mathrm{PD}(s) = P(C(s)|X_s)$ 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 & West (1983) and Harrison & Veerapen (1994).

PROPOSITION 1. *At passive surveillance the method based on the posterior distribution with the alarm set* ${}_pA(s) = \{X_s : PD(s) > c\}$ *implies* $PV(s) > c$.

At passive surveillance the predictive value increases to one as time s increases, for common methods, since $P(C(s)) = P(\tau \le s)$ tends to one. As an example, the predictive value for the Shewhart method, when τ has a geometric distribution with intensity ν, will be given. For the Shewhart method, the alarm probabilities $\alpha = P(t_A = t | t_A \ge t,\ D(t))$ and $(1-\beta) = P(t_A = t | t_A \ge t,\ C(t))$ do not depend on time which simplifies formulas. In this case we have

$$\begin{aligned}\mathrm{PV}(s) &= P(C(s)|A(s)) = P(C(s) \cap A(s))/P(A(s)) = \\ &= \sum_{\tau=1}^{s}(1-\nu)^{\tau-1}\nu(1-\beta) \Big/ \left[\sum_{\tau=1}^{s}(1-\nu)^{\tau-1}\nu(1-\beta) + (1-\nu)^s\alpha\right] \\ &= \left[(1-\beta)(1-(1-\nu)^s)\right] \Big/ \left[(1-\beta)(1-(1-\nu)^s) - \alpha(1-\nu)^s\right]\end{aligned}$$

which tends to one when s tends to ∞.

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

At active surveillance the predictive value usually has an asymptote less than one, since the probability that the first alarm occurs at time 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 & Wessman (1999). The predictive value is not monotonically increasing for all methods.

There is an important 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, are of interest.

3.5 ARL

A measure that is often used in quality control is the average of the run length until the first alarm (see e.g. Page (1954) and Wetherill & Brown (1991)). 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. To connect with the Section 3.2, it can be noted that $\mathrm{ARL}^1 = \mathrm{ED}(1) + 1$. For some situations and methods the properties are about the same, regardless of when the change occurred, but this is not always true as is illustrated by Frisén & Wessman (1999) and Frisén & Sonesson (2002). The run length distributions are often very skewed and the skewness depends on important parameters. Sometimes, it is suggested that the whole run length distribution should be reported. Instead of the average, Gan (1993) advocates that the median run length should be used on the grounds that it might be more easily interpreted. However, the main problem is that only the case $\tau = 1$ is considered.

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 commonly used in the literature on quality control and is often used also in more general statistical literature. Its dominating position will now be questioned. Consequences of this criterion, which make it unsuitable for many applications, will be demonstrated in Propositions 2, 3 and 4. Some might consider the consequences self-evident, but since the conclusions are in contradiction with much of the literature and practice, detailed proofs are given in Appendix 1. First we demonstrate, for the standard case specified in Section 2, that a linear combination with equal weights to all observations fulfils the criterion.

PROPOSITION 2. *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.

Thus, methods which give equal weights to all observations can satisfy the optimality criterion of minimal ARL^1 for fixed ARL^0. This is confirmed and exemplified in Sections 4.2.2 and 4.3. where also the drawbacks with such methods are discussed. Methods with equal weights for old and recent observations are seldom appropriate. The Proposition 2 thus demonstrates that the optimality criterion could be questioned as a formal criterion.

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

We construct a method, the Two-Point method, which fulfils the criterion of a minimal ARL^1 for a fixed ARL^0. Denote the fixed desired value of ARL^0 with A and the standard normal distribution function with Φ. The method has the alarm limits $c_1 = L$, $c_i = \infty$ for $i = 2, 3, \ldots k-1$ and $c_k = -\infty$, where $k = [A - \Phi(-L)]/\Phi(L)$ and L is restricted to those values which make k an integer.

PROPOSITION 3. *The Two-Point method fulfils "the criterion of minimal ARL^1" by having ARL^1 arbitrarily close to the minimal value, one, for a fixed ARL^0.*

The Two-Point method of the proposition above will have very bad properties. Whether an alarm is given at time 1 or at time k will depend only on the first observation. The above demonstration of the possibility to fulfil 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 the shortcomings of the criterion.

Now we give similar results for a more reasonable method, here named the LCUSUM method (Section 4.3) which minimizes the ARL^1 for a fixed false alarm probability.

PROPOSITION 4. *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 false alarm probability $P(t_A < \tau)$.

As pointed out by Pollak & Siegmund (1985) the maximal value of CED(i) is equal to CED(1) for many methods and with a minimax perspective this can be a motivation for the use of ARL^1 since CED(1) = ARL^1 − 1. However, this argument is not relevant for all methods. In particular it is demonstrated by Frisén & Sonesson (2002) that the maximal value is not CED(1) for the EWMA method of Section 4.2. For this method, there is no similarity between the solution to the ARL-criterion and the minimax-criterion, while the solutions to the criterion of expected delay and the minimax-criterion agree well.

From the point of optimal decision theory it is hard to motivate a cost function with no cost for a delay of the alarm when $\tau > 1$.

4 General Methods and their Optimality Properties

First, some general methods are described for the simple situation specified in Section 2 and their optimality properties are determined. Then, in Section 5, special methods for some more complicated situations will be described.

In Figure 1 the alarm sets 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.

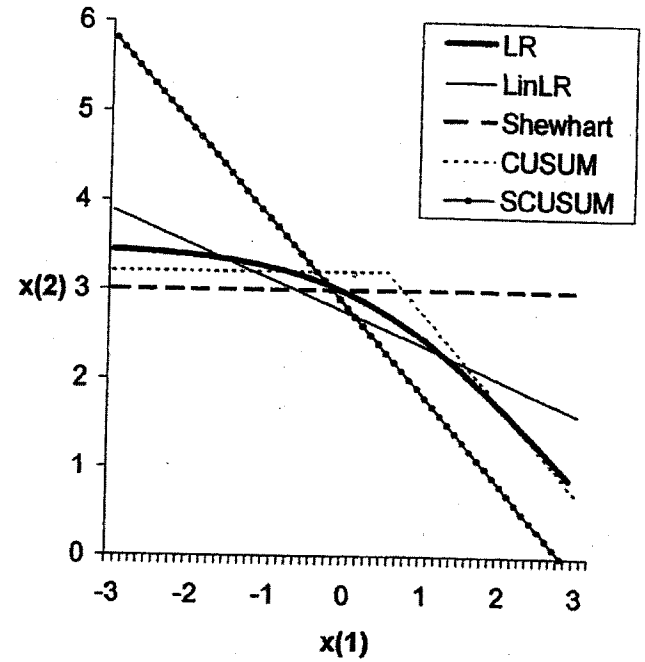

Figure 1. *Boundaries of the alarm sets at decision time $s = 2$ for some methods described in the text and in Table 1. The values $\nu = 0.01$ and $\mu = 1$ were used for those methods which can be optimized.*

In Table 1 some main characterizations of some methods are schematically described. Some methods are very flexible with several parameters, while others are fixed. The parameters can be chosen to make the method optimal for the specific conditions (e.g. the size of the change or the intensity of changes) of the application. The number of parameters which can be used to optimize for different situations is one important aspect to consider and is also given in the table. Many methods for surveillance are based in one way or another on likelihood ratios. For the comparison, expressions in terms of the partial likelihood ratios are also given in Table 1.

4.1 *The Likelihood Ratio Method*

All methods are described by likelihood ratios in Table 1. However, the full likelihood ratio method and its relation to the posterior probability has a special motivation. It will first be described and then approximations and other methods will be described.

4.1.1 *The LR method*

A method constructed by Frisén & 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 partial

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

Method		Formula number	Alarmfunction of L(t)	No. of parameters in the alarmfunction	Optimality
LR(full likelihood ratio)		(1)	$\sum_{t=1}^{s} w(s,t)L(s,t)$	2	min $E(t_A - \tau \mid t_a \geq \tau)$ for fixed $P(t_A < \tau)$ and max $P(A(s)\mid C)$ for fixed $P(A(s)\mid D)$ when $C = \{\tau \leq s\}$ and $D = \{\tau > s\}$
Shiryayev–Roberts		(1) with $\nu \to 0$	$\sum_{t=1}^{s} L(s,t)$	1	As for LR if $\nu \to 0$
LinLR (linearization of the LR method)		(2)		2	approximation of that for LR
EWMA	with λ_{EWLR}	(3)		1	approximation of that for LR
	with small λ				approximation of that for SCUSUM
SCUSUM		(4)	$L(s,1)$	0	max $P(A(s)\mid C)$ for fixed $P(A(s)\mid D)$ when $C = \{\tau = 1\}$ and $D = \{\tau > s\}$
LCUSUM		(5)	$L(s,1)$	0	min ARL^1 for fixed $P(A(s)\mid D)$
CUSUM		(6)	$\max_t L(s,t)$	1	best min max $E(t_A - \tau \mid t_A \geq \tau)$ for fixed $P(t_A < \tau)$
Shewhart		(7)	$L(s,s)$	0	min $E(t_A - \tau \mid t_A \geq \tau)$ for fixed $P(t_A < \tau)$ asymptotically for large μ and max $P(A(s)\mid C)$ for fixed $P(A(s)\mid D)$ when $C = \{\tau = s\}$ and $D = \{\tau > s\}$

likelihood ratios. Although methods based on likelihood ratios have been suggested earlier their use in practice is (still) rare. The likelihood ratio method will be used here 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_s for which the likelihood ratio exceeds a limit:

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

For the above definitions of C and D this can be expressed as

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

where $w(s,t) = P(\tau = t)/P(\tau \leq s)$ and the partial likelihood $L(s,t)$ is defined by

$$L(s,t) = f_{X_s}(x_s|\tau = t)/f_{X_s}(x_s|D).$$

For the case of normal distribution, 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

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

is a nonlinear function of the observations.

In order to achieve the maximum detection probability described in Section 3.1, an alarm should be given as soon as $p(x_s) > G_s$.

In the case of a geometric distribution of τ the condition of "minimal expected delay", as described in Section 3.2, is achieved if an alarm is made as soon as the posterior distribution exceeds a fixed limit (Shiryaev, 1963). To see this we have

$$PD(s) = P(\tau \leq s|X_s = x_s) > K \Leftrightarrow p(x_s) > \frac{P(\tau > s)}{P(\tau \leq s)} \frac{K}{1-K},$$

where K is a constant. Thus, the optimality is achieved by the likelihood ratio method when the alarm is specified by

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

The method that thus gives alarm for the first s for which

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

will here be called the LR method. A common 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 then

given for the first s for which

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

When ν tends to zero, the weights $w(s,t)$ do not depend on t. The limit G_s of the LR method does not depend on s since the ratios G_{s+1}/G_s tend to 1 when ν tends to zero. 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, such that

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

where G is a constant. The method has an approximately constant predictive value (Frisén & Wessman, 1999), which allows the same interpretation of early and late alarms.

The posterior distribution $PD(s) = P(C(s)|X_s)$ has been suggested as an alarm criterion by e.g. Smith & West (1983). When there are only two states, C and D, this criterion leads to the LR method (Frisén & de Maré, 1991). Sometimes the use of the likelihood ratio or equivalently the use of the posterior distribution is named "the Bayes' method". It depends on the situation whether the distribution of τ is considered as "a prior", an observed frequency-distribution (e.g. from earlier cases in intensive medical care) or just reflects which situation optimality is desired.

4.1.2 *Linear approximation of the likelihood ratio method*

A linear approximation of the LR method is of interest for two reasons. One is to obtain a method which is easier to use and analyze, but has similar good properties as the LR method. Another is to get a tool for the analysis of approximate optimality of other methods. Different approximations might be of interest for different situations. Here we will study three variants.The formulas and the derivations of the approximations are given in Appendix 2.

The first approximation, which is denoted by LinLR is achieved by a Taylor approximation of the alarm function. With the notation of Appendix 2, it will give an alarm for the first s for which

$$\sum_{t=1}^{s}w_{\text{LinLR}}(s,t)x(t) > L_{\text{LinLR}}. \quad (2)$$

The weights of the LinLR method can be approximated by exponential weights, and then we have the EWLR method. This corresponds to using an exponentially weighted moving average statistic (Section 4.2) with the value of λ set to

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

A third approximation, EWlnLR, is achieved by a Taylor approximation of the logarithm of the alarm function of the LR method and further approximation of the weights by use of exponential weights with $\lambda = \lambda^*$.

A large scale simulation study by Frisén & Sonesson (2002) demonstrates that all the approximations are satisfactory and that the properties approximate those of the LR method well.

4.2 *Exponentially Weighted Moving Average*

A method for surveillance based on exponentially weighted moving averages, EWMA, was described by Roberts (1959). Positive reports of the quality of the method are given by, e.g. Robinson & Ho (1978), Crowder (1987), Ng & Case (1989), Lucas & Saccucci (1990) and Domangue & Patch (1991).

The alarm statistic is $Z_s = (1-\lambda)Z_{s-1} + \lambda x(s)$, $s = 1, 2, \ldots$, where $0 < \lambda \le 1$. In the standard version of the method, Z_0 is the target value μ^0, which is here chosen to be zero. The statistic is sometimes referred to as a geometric moving average since it can equivalently be written as

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

where $b = 1/(1-\lambda)$ is a constant> 1. An 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 of Z_s. When we standardize with weights $w_E(s,t) = \lambda(1-\lambda)^{s-t}/[1-(1-\lambda)^s]$, which sum to one, this method will give an alarm for the first s for which

$$\sum_{t=1}^{s}w_E(s,t)x(t) > L_{EA} \quad (3)$$

where $L_{EA} = L\sigma_Z/[1-(1-\lambda)^s]$.

The EWMA statistic 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 4.6. If λ is near zero, all observations have approximately the same weight. Since the EWMA method has two parameters, λ and L, these can be chosen to make the EWMA method equal to any other linear method when $s = 2$, as in Figure 1. The EWMA method is thus not included separately in that figure. When $s > 2$ differences appear.

Many variants of EWMA with low alarm limits for early time points are suggested. One such suggestion is the use of the exact variance (Roberts, 1959) instead of the asymptotic. Steiner (1999) suggests the FIR (fast initial response) which starts with $Z_0 > 0$. This implies that the unavoidable false alarms have a higher probability to come early instead of late.

4.2.1 *Approximate ED-optimality*

Differences and similarities with the linearizations of the LR method will now be examined in order to find conditions for approximate optimality of the EWMA method. All proofs are given in Appendix 1.

PROPOSITION 5. *There does not exist any λ which makes the EWMA exactly optimal with respect to the "maximum detection probability" or the "minimal expected delay".*

PROPOSITION 6. *For late observations approximate identification with the weights of the LinLR method is achieved with $\lambda = \lambda^*$.*

Thus, approximation of the good properties of the LR method according to "the maximum detection probability criterion" of Section 3.1 can be expected.

A comparison between the weights of the observations by LinLR method and the weights in the EWMA method with $\lambda = \lambda^*$ is made in Figure 2. In the beginning of the surveillance the EWMA gives more weight to the older observations than the LinLR method. However, already for decision time $s = 10$, the differences between the two methods are without importance for the case in the figure. For $s = 15$ it is not possible to see any difference in the scale of the figure. The approximation deteriorates as λ^* decreases.

For a full evaluation of optimality, it is necessary also to consider how the limits for alarm depend on s.

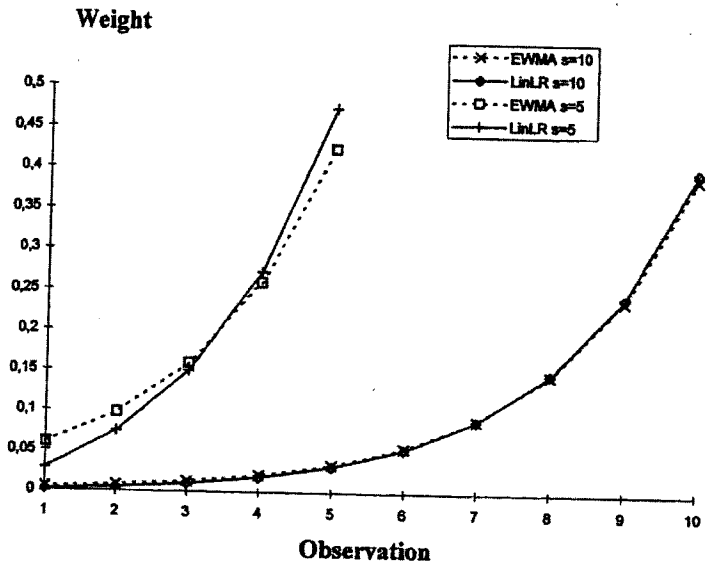

Figure 2. *Connections with straight lines of the weights $w(s,t)$ of the observations $x(t)$. The weights of the EWMA method are calculated for $\lambda = 1-\exp(\mu^2/2)/(1-\nu)$. The LinLR method is optimized for the case when the change τ has a geometric distribution with intensity $\nu = 0.01$ and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

PROPOSITION 7. *For late decisions and $\lambda = \lambda^*$, the alarm limit of the EWMA approximates those of the LinLR, EWLR and EWlnLR methods.*

Thus, the EWMA method approximates the approximations of the LR method. When both these approximations are good, EWMA will in turn approximate the LR method. Thus, the EWMA method could be expected to approximately fulfil also the optimality condition of Section 3.2 of a minimal expected delay. The simulation study by Frisén & Sonesson (2002) demonstrates that for large values of μ, the EWMA method has much worse expected delay than the LR method and the approximations. However, for small and moderate values the choice $\lambda = \lambda^*$ makes EWMA nearly as good as the LR method.

4.2.2 *The value of the parameter which gives ARL-optimality*

According to Proposition 2, 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. This is confirmed by simulations by Chan & Zhang (2000) and Frisén & Sonesson (2002) in studies of different values of the parameter of the EWMA method. The ARL^1 decreases when λ approaches

zero. There are a large 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. When λ approaches zero, the standard EWMA approaches the SCUSUM method of Section 4.3, while the ARL^1 for the commonly recommended value of λ (for two-sided procedures) corresponds to a local minimum for the one-sided specific case studied by Frisén & Sonesson (2002). This should not be interpreted as a disadvantage of the commonly used values of λ but as a warning against uncritical use of the ARL criterion.

4.3 *Simple Cumulative Sums*

The methods in this section will be demonstrated to be closely related to the ARL criterion. Sometimes CUSUM is used as a unifying term 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, here $\mu^0 = 0$, exceeds a fixed limit

$$C_s = \sum_{t=1}^{s} x(t) > L, \tag{4}$$

where L is a constant. This method is sometimes called "the simple CUSUM". It will here be denoted as SCUSUM. The similarity with the EWMA method when λ tends to zero is illustrated by Frisén & Sonesson (2002). The SCUSUM method gives optimal error probabilities for $\tau = 1$ in the case specified in Section 2. However, when $\tau > 1$, SCUSUM cannot compete with other methods. As is seen in Figure 1 the shape of the alarm set is quite different from the ED-optimal one.

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 4 we have an alarm as soon as

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

This method, which gives an alarm as soon as C_s exceeds a linear function of s, is here called the LCUSUM method. By choosing L small enough in this method, the finite value of ARL^1 can be made arbitrarily close to one. Still, for this L the ARL^0 will not be finite and thus greater than any fixed value. 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 defined in Section 4.1, 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 set 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.

4.4 *CUSUM*

This method will be demonstrated to be closely related to the minimax criterion. The variant of cusum tests, which is most often advocated, is named "the CUSUM method" Page (1954). There is

an alarm for the first s for which

$$C_s - C_{s-i} > h + ki \quad \text{for some } i = 1, 2, \ldots, s, \tag{6}$$

where $C_0 = 0$ and h and k are chosen constants. By the CUSUM method (in contrast to the simple variants of Section 4.3) the information from earlier observations is handled quite differently depending on the position in the time series. Sometimes (e.g. Lorden (1971) and Siegmund (1985)) the CUSUM method is presented in a general way by likelihood ratios (which in the normal case reduce to $C_s - C_{s-i}$). Yashchin (1993) and Hawkins & Olwell (1998) give thorough reviews of the CUSUM method.

The CUSUM method is the result of a natural combination of methods. Each of these is optimal, with respect to the expected delay, 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

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

where G is a constant. The method is sometimes called the likelihood ratio method, but this combination of likelihood ratios should not be confused with the full likelihood ratio method, LR, of Section 4.1. In Figure 1 the boundary of the alarm set of the CUSUM method is seen to be a two-phase linear approximation of the nonlinear limit of the LR method.

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 & Kemp (1960), where it is concluded from a nomogram that this value seems to be good. 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 $\mu/2$. This slope is optimal, with respect to the expected delay, in each step. However it does not prove that it is ED-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 3.3. Different variants and generalizations are discussed in the theoretical literature on minimax optimal methods e.g. Lai (1995), Lai (1998) and Lai & Shan (1999).

4.5 *Moving Average*

The moving average method gives an alarm as soon as

$$C_s - C_{s-d} > L,$$

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

$$L(s, s - d) > G$$

where G is a constant. It will thus have the optimal error probabilities of the LR method with $C = \{\tau = s - d\}$.

4.6 *Shewhart*

This method, which was suggested by Shewhart (1931) is much used in quality control. 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

$$x(s) > G. \tag{7}$$

The limit G for a fixed ARL^0, is calculated by the relation: $P(X(s) > G|\mu(s) = \mu^0) = 1/\text{ARL}^0$. For an illustration of the alarm set at decision time $s = 2$ see Figure 1. More detailed descriptions are found in many textbooks like Wetherill & Brown (1991).

The alarm statistic of the LR method

$$f_{X_s}(x_s|C)/f_{X_s}(x_s|D)$$

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

$$L(s, s) > G$$

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

For large shifts, Frisén & Wessman (1999) showed that the LR method and the CUSUM method converge to the Shewhart method.

5 Methods and Optimality for Complicated Situations

Much research has been done on construction of methods for special situations. The panel discussion edited by Montgomery & Woodall (1997) contains many references. In complicated problems it is not always easy to achieve, or even define, optimality and this is seldom done. When the states between which the change occurs are completely specified the full likelihood ratio with its good optimality properties can be used. Pollak & Siegmund (1985) point out that the martingale property (for continuous time) of the Shiryaev–Roberts method makes it more suitable than the CUSUM method, (which does not have this property) for adaption to complicated problems. On the other hand Lai (1995), Lai (1998) and Lai & Shan (1999) point out that the good minimax properties of generalizations of the CUSUM method make the CUSUM suitable for complicated problems. In this section different inferential approaches, and corresponding optimality properties, to some complicated problems are described.

5.1 *Special Kinds of Changes*

5.1.1 *Two-sided alternatives*

In the earlier sections, one-sided procedures were discussed in order to get some sharp results on optimality. However, in many applications two-sided procedures are motivated. A common approach is to use two parallel one-sided surveillance procedures and signal an alarm as soon as either of the procedures gives a signal. This is a special case of the union intersection method discussed in Section 5.5 for multivariate surveillance. Another common approach is to use symmetric limits for the alarm statistic.

For CUSUM the two approaches give the same result, and the properties are easily related to those of a one-sided procedure as is demonstrated by Kemp (1961) and van Dobben de Bruyn (1968). This is due to the maximization in the statistic. The same result for both approaches is not achieved in general and not for the SCUSUM or the EWMA methods. The properties for the one- and two-sided versions are not easily related to each other. It has been suggested by Champ *et al.* (1991), Gan (1995) and Gan (1998) to use a barrier for each of the one-sided EWMA procedures and thus get a simple relation between properties for two-sided and one-sided methods for surveillance and at the same time avoid "worst possible" effects with respect to the history before the change. A slight

modification of the barriers used in the papers above is used by Morais & Pacheco (1998) to achieve more easily approximated ARL properties. Comparisons between one- and two-sided EWMA with and without barriers with respect to ARL and ED are reported by Sonesson (2001). For the one-sided case the ARL-optimal value of λ is zero, as expected by the characteristics of the ARL criterion discussed in Section 3.5. However, this is not true for the two-sided case. This means that for the two-sided case the order of the observations (which for $\tau = 1$ should be an ancillary statistic) influences the ARL-optimal alarm statistic. This conflict between the ARL-criterion and the ancillary principle is explained by the deficiency of the ARL-criterion.

Pollak & Siegmund (1985) suggest that a two-sided version of the Shiryaev–Roberts method should be constructed by a weighted average of the statistics for the two one-sided variants. With known probabilities for the two alternatives the full likelihood ratio method could be used. Thus, we have the "maximum detection probability" and the "minimal expected delay".

5.1.2 *Gradual shifts*

Most of the literature on surveillance treats the case of an abrupt change. However, in applications it is not uncommon to have a gradual change which starts at an unknown time. One example is the recording of radioactivity when a radioactive cloud is brought with the wind from a site with a nuclear incident (Järpe, 2000a). Another case is the post marketing surveillance of adverse drug effects, where a start of a gradual increase of cases in the population is expected if a released drug turns out to be harmful (Sveréus, 1995). In both these papers it is demonstrated that the methods in current use in Sweden, which in both cases are based on differences between moving windows, are inefficient for detection of gradual changes. Comparisons between methods for the case of a linear change are made by Aerne *et al.* (1991) and Gan (1992). Arteaga & Ledolter (1997) compare several procedures with respect to ARL properties for several different monotonic changes. One of the suggested methods in that paper is based on the likelihood ratio, isotonic regression and a window. Yashchin (1993) discusses generalizations of the CUSUM and the EWMA methods which could detect both sudden and gradual changes.

5.1.3 *Turning points*

Sometimes the timely detection of a change in monotonicity is important. This is the case in natural family planning, where a change from increasing to decreasing (or vice versa) values of some indicators of hormone production are markers of start or end of the fertile phase (Royston, 1991). Another situation where timely detection of turning points is important is for governments and business, when the turns in leading business indicators are used to predict a future turn in the business cycle. A third example is for financial decisions, where the selling of an asset is desired at the time of maximum price (or function of it). Hidden Markov Models are natural for the switches between the up- and down phase and are used both for business cycles and finance (Dewachter, 2001). A piecewise linear curve is often assumed for suggested methods for business cycles and finance. When the assumptions on prior knowledge are the same, it is demonstrated by Andersson *et al.* (2001) that the HMM method is identical to the LR method.

When the knowledge on the shape of the curve is uncertain, a non-parametric method is of interest. A maximum likelihood ratio method was constructed by Frisén (2000) with the likelihood statistic in the LR method of Section 4.1 replaced by the maximum likelihood ratio

$$p(x_s) = \sum_{k=1}^{s} \frac{\pi_k}{\Pr(\tau \le s)} \exp \frac{1}{2}\{Q(0) - Q(k)\}$$

where $Q(k)$ is the (standardized) quadratic deviation from the model with a turn at time k which has the least deviation. $Q(0)$ is the deviation from the best fitting model without a turn in the specified time period. The deviations are based on the observations available at each decision time, s. These deviations can be calculated by the methods given by Frisén (1980) and Frisén (1986) for unimodal regression. The expected delay by this method is investigated by Andersson (2002). Evaluations and comparisons with currently used methods in Sweden for detection of turns in business cycles are made by Andersson *et al.* (2001).

5.1.4 *Change in variance or interaction*

The variance is the most commonly studied distributional parameter, except the mean. Monitoring a transformation of the variance by standard charts is often suggested. Nelson (1990), Acosta-Mejia (1998) and Ncube & Li (1999) suggest that the (subgroup) range is used in combination with standard methods for surveillance. von Collani & Sheil (1989) use the standard deviation. Chang & Gan (1995), Srivastava (1997) and Morais & Pacheco (1998) use the logarithm of the variance. The statistics might be affected not only by a change in the variance but also by a change in the level, which also might be relevant. Thus, much of the discussion on change in the variance is a discussion on change in variance and/or mean. This multivariate problem will be discussed in Section 5.5.1. Robustness with respect to non-normality and serially correlated observations of the CUSUM method for monitoring of the variance is examined by Chang & Gan (1995) and the method is compared with the EWMA method. Comparisons with respect to ARL between different ways to monitor the variation are done by e.g. Acosta-Mejia *et al.* (1999).

In connection with spatial surveillance, Järpe (1999) suggests a method for detection of a change in the parameter for spatial interaction in a generalized linear model.

5.1.5 *Change between unknown parameter values*

The method by Bell *et al.* (1994) to detect a change to a stochastically larger distribution (nonparametric but geared to the exponential distribution), is applied in their paper to the detection of a change of the parameter in a Bernoulli process to an unknown but larger value. Asymptotic efficiency is reported. Gordon & Pollak (1997) use invariant statistics combined by the Shiryaev–Roberts method to handle the case of an unknown pre-change distribution in regard to a nuisance parameter, e.g. the pre-change mean of a normal distribution, and evaluate the methods by ARL.

Lai (1998) suggests the GLR method, where G stands for generalization and LR for the CUSUM-combination of partial likelihood ratios. A prior distribution for the value after the change is used. To avoid cumbersome computation the suggestion is to use a window so that only recent observations are used. The method fulfils an asymptotic minimax criterion.

5.2 *Change in a Non-Normal Distribution*

5.2.1 *Methods designed for other specified distributions than the normal*

Surveillance of the frequency of rare events is an important example of change of a parameter in a distribution which is not Gaussian. Usually, as in Gan (1998), the methods are based on the distances in time between successive events. Rossi *et al.* (1999) suggest that the Poisson distribution should be approximated by the normal one. A bibliography of control charts based on attribute (or count) data is given by Woodall (1997). Padgett & Spurrier (1990) and Ramalhoto & Morais (1999) construct Shewhart type methods suitable for the Weibull distributions. Padgett & Spurrier (1990) give the Shewhart limits for the lognormal distribution.

5.2.2 *Non-parametric methods and robustness with respect to distribution*

An overview is given by the book on non-parametric change-point problems by Brodsky & Darkhovsky (1993). In Bell *et al.* (1994) a non-parametric method based on the Shiryaev–Roberts method and geared to the exponential distribution is suggested for the surveillance of a change in distribution to a stochastically larger distribution. Very high asymptotic relative efficiency is reported. In Gordon & Pollak (1997) invariant statistics are used for a similar setting and the ARL properties are given. Ranks are used for modified EWMA (Hackl & Ledolter, 1991) and modified Shewhart and CUSUM (Liu, 1995). Liu & Tang (1996) construct a completely non-parametric generalization of the Shewhart method based on the bootstrap technique. Jones & Woodall (1998) compare the ARL properties of some methods based on the bootstrap technique. Chakraborti *et al.* (2001) critically examine several methods which are claimed to be distribution free, especially those based on the Hodges–Lehmann estimates.

The robustness with respect to non-normality of the CUSUM method and some modified methods is examined by Lucas & Crosier (1982). Chang & Gan (1995) examine the effect of non-normality of the CUSUM and EWMA methods for surveillance of the variance. Robustness of the ARL ^{0} for skewed and heavy tailed distributions is examined for the Shewhart method and EWMA-variants by Borror *et al.* (1999). Robustness of different methods against deviations from the normal distribution and also lack of independence is examined by Stoumbos & Reynolds (2000).

5.3 *Dependent Observations*

The three most common ways to treat the case of dependent observations are 1) to use the ordinary method and to study the robustness, 2) to use the ordinary method but with wider alarm limits based on the correct variance or 3) to use the residuals from a time series model.

The robustness of the ordinary EWMA and CUSUM methods when the data are generated by an AR(1) process is investigated by VanBrackle & Reynolds (1997), who also suggest modifications. Properties of the EWMA method in the presence of autocorrelation are derived by Schmid & Sch öne (1997).

Schmid (1997) uses limits based on the correct variance, given the autocorrelation, and compares this approach with the residual-based versions with respect to ARL properties. Liu & Tang (1996) suggest a nonparametric bootstrap-based generalization of the Shewhart method which does not require independent observations.

VanBrackle & Williamson (1999) examine the ARL properties of several general methods and several types of shifts when one-step ahead forecasts are used. In Cardinal *et al.* (1999) integer-valued counts of diseases are monitored for public health purposes by monitoring of the forecasts by a model suitable for this kind of data. In Lu & Reynolds Jr. (1999) and Lu & Reynolds Jr. (2001) the EWMA and CUSUM methods, respectively are applied to the original observations and to the residuals. In their paper on EWMA the ARL is used and in the paper on CUSUM the "steady state ARL" (the asymptotic value of CED) is used.

The three approaches mentioned above are compared by Pettersson (1998a) with respect to several measures, such as the predicted value and the expected delay, besides the usual ARL. It is also demonstrated in that paper that the residual statistic can be seen as a rough approximation of the full likelihood ratio statistic (with some terms deleted). Lai (1998) gives a generalization of the theorem by Lorden (1971), on asymptotic minimax optimality for the CUSUM method, for the case of dependent variables, by applying the method to the forecasts.

5.4 *Complicated Regression Models*

If the nuisance parameters of the model are known, the residuals might have simple properties which can be used for surveillance of possible changes from the model. When the parameters have to be estimated from the data the situation is more complicated. However, in many cases (Brown *et al.* (1975) and Frisén (1992)) the recursive residuals from an estimated regression model have simple properties and can easily be used by some general technique for surveillance. Brown *et al.* (1975) suggest the use of the CUSUM statistic of Section 4.4 but with different alarm limits. These limits are constructed to give a system of repeated significance tests as discussed in Section 3.1.

In McLaren *et al.* (2000) hierarchical multiple regression modeling is used as the basis for surveillance of changes in the pattern of individual patients laboratory data. In Yashchin (1995) a "regenerative likelihood ratio method" (named LR) of CUSUM type, which allows periodic discarding of data and thus is possible to compute also for complicated problems, is proposed for the monitoring of parameters of a nested random effect model. It is evaluated by the ARL ^{1}.

When the only assumption of pre-change regression is that it is monotonic, the (non-parametric) maximum likelihood estimator is suitable to use with the LR method, as discussed in Section 5.1.3.

5.5 *Multivariate Surveillance*

Reviews of multivariate surveillance are given by, e.g. Lowry & Montgomery (1995) and Woodall & Montgomery (1999). Also, the book by Basseville & Nikiforov (1993) contains much discussion on multivariate problems.

One common way to deal with multivarate problems is to construct an omnibus statistic which is supposed to take care of the important aspects of the multivariate problem. A survey of different omnibus methods is given by Kourti & MacGregor (1996). Commonly used statistics are the χ^2 and the T^2 statistics already suggested by Hotelling (1947) for surveillance. He used the Shewhart method and Crosier (1988) the CUSUM method to monitor the omnibus statistic.

Multivariate versions of the EWMA and CUSUM methods, named MEWMA (Lowry *et al.*, 1992) and MCUSUM (Crosier, 1988), are constructed by T^2 statistics based on the EWMA, respectively CUSUM, vectors. Crosier (1988) compares his two ways to combine T^2 and CUSUM and concludes that MCUSUM has better ARL properties. Runger & Prabhu (1996) give a numerical procedure based on Markov chains for the computation of the ARL of the MEWMA.

Projection methods, such as principal component analysis or partial least squares technique, to reduce the dimension of a multivariate surveillance problem, are recommended by Kourti & MacGregor (1996) and Scranton *et al.* (1996).

The union intersection principle can be used to handle parallel surveillance for each variable by signaling an alarm at the first time one of the procedures gives alarm. Different suggestions of the use of parallel procedures are given by Woodall & Ncube (1985), Hayter & Tsui (1994) and Timm (1996).

Hawkins (1991) suggests that the scaled residuals from the regression of each variable on the others are used. He notes that this is equivalent to basing the surveillance on the likelihood ratios for change in each direction. He suggests that the components are monitored by parallel Shewhart or CUSUM methods. The full likelihood ratio method can be applied as soon as the event to be detected is specified. This is done by Wessman (1998) for the surveillance when all the variables change at the same time and by Wessman (1999) for different change-points.

If the alternatives are completely specified general techniques as suggested by e.g. Lai (1995). Otherwise, optimality in multidimensional problems is hard to specify. In the literature on quality control, the ARL properties for different alternatives are discussed. In Tsui & Woodall (1993) the components of the combined statistic are weighted by the components of a loss function for shifts in different directions and then evaluated by ARL. Sometimes the Bonferroni method is used to

control an error when conclusions are made about several variables. In Wessman (1999) there is a comparison with respect to ARL and PSD between the T^2, the union intersection method, the full likelihood ratio statistic and the method with component likelihoods by Hawkins (1991) when the statistics are monitored by the Shewhart method.

Important examples of multivariate surveillance are the simultaneous monitoring of the mean and variance and also spatial statistics. These two areas will now be described.

5.5.1 *Methods for detection of change in mean and variance*

The different general approaches mentioned in Section 5.5 are of course applicable also for the mean and variance. However, this problem has drawn a special interest. An overview is given by Flury *et al.* (1995).

Use of parallel charts for the mean and the variance (or a function of the variance) is suggested by e.g. Saniga (1989) and Morais & Pacheco (2000). In the latter the probability of signaling in the wrong chart is evaluated. Monitoring of the maximum and the minimum in samples for detection of change in the mean and/or the variance is suggested by Amin *et al.* (1999).

One example of an omnibus statistic is the capability index, which according to Woodall & Montgomery (1999) is widely used in industry. Domangue & Patch (1991) compare several omnibus statistics when monitored by the EWMA method.

Comparisons between several variants of omnibus and marginal methods for the mean and variance by EWMA methods are made by Gan (1995) with respect to the ARL. The conclusion by Gan for the situations examined is that the omnibus methods have several drawbacks. It makes a great difference if the aim is to detect a simultaneous change in mean and variance or if the most interesting case is that only one will change but one doesn't know in advance which one.

5.5.2 *Methods for spatial surveillance*

In areas such as monitoring of geographical disease patterns (see e.g. Lawson *et al.*, 1999) and control of environmental risks (see e.g. Barnett & Turkman, 1993), and technical pattern recognition, it is often necessary with models to include both a spatial and a temporal structure. Rogerson (2001) monitors a spatial scan statistic (Kulldorff, 1997) with the CUSUM method. Järpe (1999) constructs a surveillance system for the simplest nontrivial spatial model, the Ising model. Surveillance problems related to the detection of the geographically spreading increased radiation level, in case of a nuclear incident, are treated by Järpe (2000b). Rogerson (2001) and Lawson (2001) point out that surveillance approaches to spatial statistics are still rare.

6 Concluding Remarks

The performance of a system of surveillance depends on the time, τ, of the change. To get an optimality criterion, either a summarizing measure over the distribution of τ, or evaluation for a specific value of τ, can be used. Evaluation for the value which gives the maximal expected delay is one interesting way. Evaluation of the case in which τ tends to infinity is common in theoretical literature, but is not enough for all applications. The other extreme is to study the case where $\tau = 1$. This is the commonly used procedure and will be discussed in more detail.

Often the criterion is stated as minimal ARL^1 for a fixed ARL^0. The frequent use of this criterion is an indication of that it is useful for many purposes. However, it might be dangerous to use it without caution for other purposes. As was noted in Proposition 2, this criterion agrees with methods where all observations are given the same weight. The shortcomings of such methods were pointed out in Section 4.3 and such methods are not often recommended. Instead, methods which have all

the 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 perform well if the change occurs later. If the problem is to discriminate between the hypotheses $\mu(t) = 0$ for all t and the hypothesis $\mu(t) = \mu$ for all t, then sequential methods for tests of fixed hypotheses (such as the power one SPRT method of Proposition 4) are appropriate. Only the situations where a change is expected to happen after an unknown time, τ, require the special methods for surveillance.

An argument for the use of the ARL criterion has been that it agrees with the minimax criterion. However, this is true only for some methods and not at all for others.

A summarizing optimality criterion is the expected delay with respect to the distribution of τ. Exact information about the distribution might be lacking. However, the drawbacks of the criteria based on ARL demonstrate the importance of any information on the distribution of τ. The robustness is important. The true values of the shift μ or the intensity ν might not be known. Instead, the corresponding parameters are set to the values M and V. The method will then be optimal for these values but not for the true ones. Srivastava & Wu (1993) studied the asymptotic effect of different values of the μ for a fixed value M of the parameter. Järpe & Wessman (2000) studied the same effect for small samples. Frisén & Wessman (1999) studied the small sample properties for different values of M for a fixed value of μ to examine the robustness to the choice of parameter value M. The theorems and the figures in that paper 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 like 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 the choice of method is not very important when M is large. No great differences between methods could be seen in the simulation study by Frisén & Wessman (1999) for M larger than 2 for $\mu = 1$. This confirms the results by Mevorach & Pollak (1991) that the Shiryaev–Roberts method and the CUSUM method have similar properties for the cases $M = 5$ and $M = 7$ for $\mu = 1$. The study by Frisén & Wessman (1999) also confirms the conjecture by Roberts (1966) about the robustness with respect to differences between the assumed and true intensities V and ν.

Criteria based on the posterior distribution have an intricate relation both to the expected delay and to the predictive value of an alarm. These relations were analyzed in Section 3.4.2 for passive and active surveillance.

Results on the optimality of different methods are summarized in Table 1. The LR method, which is the solution to the criterion of minimal expected delay, has a nonlinear alarm function with respect to the data. Commonly used methods are equivalent to the LR method only in extreme cases where the non-linearity disappears. Linear approximations are here used mainly for the comparison with other linear methods and to establish for which situations the methods have (approximate) optimality. The EWMA method has continuously decreasing weights for older observations. The CUSUM method has a discrete adaptive way of including old observations. This explains the good minimax properties for the CUSUM method. A good thing would be to have continuous adaptive weights. That is actually what the LR method has. The simple cumulative sum methods SCUSUM and LCUSUM satisfy optimality conditions for $C = \{\tau = 1\}$. 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 alarm sets 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 when different methods are applied. Thus, the area under the curves cannot be interpreted. However, the shapes of the boundaries demonstrate some characteristics. The linear methods LinLR, EWLR, EWlnLR and EWMA (with two and one adjustable parameter respectively) can approximate the nonlinear LR method rather well. The CUSUM method, which has one adjustable parameter and for $s = 2$ is two-phase linear

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 aim has been to give the links between optimality and methods. The situations and parameter values for which some commonly used methods have different optimality properties are determined. The suggested linear approximations of the full likelihood ratio method turned out to be a useful tool for this. The commonly used ARL optimality is demonstrated to lead to unwanted consequences when used as a formal criterion and should thus be used with caution. The concept of optimality is often hard to specify in complicated multidimensional cases. Uniform optimality can seldom be achieved. Instead, the most commonly seen evaluation is a description of the ARL properties for a set of situations. Much work remains to be done in this field which involves interesting inferential matters and which might be of great importance in many applications.

Appendix 1: Proofs

Proof of Proposition 1.

$$\mathrm{PV}(s) = P(C(s)|A(s)) = P(C(s)|X_s : P(C(s)|X_s) > c) > c.$$

Proof of Proposition 2.

First, some properties of surveillance systems based on

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

are derived. In this proof, let $C(s) = \{\tau = 1\}$ and $D(s) = \{\tau = \infty\}$ with the notation that $\tau = \infty$ is the event that no change ever happens. As a technical tool, passive surveillance with the alarm set denoted by ${}_pA(\cdot)$, is used to start with. Then, with the specifications in Section 2, the likelihood ratio method (Section 4.1) has the alarm set

$$\begin{aligned} {}_pA(s) &= \{X_s : f_{X_s}(x_s|C)/f_{X_s}(x_s|D) > a_s\} \\ &= \{X_s : \exp\left\{\frac{1}{2}\mu^2\right\}\exp\{\mu\sum_{t=1}^{s} X(t)\} > b_s\} = \{X_s : \sum_{t=1}^{s} X(t) > c_s\} \end{aligned}$$

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 & de Maré (1991) that

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

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

$$\begin{aligned} {}_aA(s) &= \{X_s : \sum_{t=1}^{s} X(t) > c_s\} \bigcap \{X_s : \sum_{t=1}^{r} X(t) \le c_r, \quad r = 1, \ldots, s-1\} \\ &= \{X_s : s = \min\{i : \sum_{t=1}^{i} X(t) > c_i\}\}. \end{aligned}$$

Thus, the monitoring system in the proposition is identical to that of a certain known likelihood-based one. 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 $P(A(s)|D(s))$ according to de Maré (1980) and Frisén & de Maré (1991).

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 $\mu(t)$ 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 $\tau = \infty$, with our notation.

$$\begin{aligned} \mathrm{ARL}^0 &= E(t_A|\mu(t) \equiv 0) = \\ &= \sum_{t=1}^{\infty} tP(t_A = t|\tau = \infty) = \sum_{t=1}^{\infty} tP({}_aA(t)|D(t)). \end{aligned}$$

$$\begin{aligned} \mathrm{ARL}^1 &= E(t_A|\mu(t) = \mu) = \\ &= \sum_{t=1}^{\infty} tP(t_A = t|\tau = 1) = \sum_{t=1}^{\infty} tP({}_aA(t)|C(t)). \end{aligned}$$

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.

Proof of Proposition 3.

The two-point method has $\mathrm{ARL}^0 = 1 - \Phi(L-0) + k\Phi(L-0) = A$ and $\mathrm{ARL}^1 = 1 - \Phi(L - \mu) + k\Phi(L-\mu) = 1 - \Phi(L-\mu)[\Phi(L) + A - \Phi(-L)]/\Phi(L)$, which has the limit one when L tends to minus infinity, since $\Phi(L-\mu)/\Phi(L)$ has the limit zero.

Proof of Proposition 4.

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

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

where G and L are constants. The expected delay, which is minimal, is equal to $\mathrm{ARL}^1 - 1$, since $t = 1$. Thus, also ARL^1 is minimal.

Proof of Proposition 5.

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.

Proof of Proposition 6.

At constant intensity ν

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

The weights, $m(t)$ of the LinLR method are found in Appendix 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 LinLR method as they are for the EWMA method. However, for large values of u the relative weight tends to b when $b > 1$. Then, $m(t+1)/m(t) = 1/(1-\lambda) = b = (1-\nu)\exp(\mu^2/2)$ and thus $\lambda = 1 - \exp(-\mu^2/2)/(1-\nu)$.

Proof of Proposition 7.

The alarm limit for the EWMA method depends on the decision time s as $1/[1-(1-\lambda)^s] = b^s/(b^s-1)$ when $\lambda = \lambda^*$. This is also the case for the LinLR, EWLR and EWlnLR methods as can be seen from the results in Appendix 2.

Appendix 2: Linearizations of the LR Method

Approximation by Taylor expansion of the alarm function at $X(i) = 0$ and with $a = \exp(\mu^2/2)$ results in the following linear approximation of the alarm function:

$$\begin{aligned} p_s(x_s) \approx p_s^*(X_s) &= p_s(0) + \sum_{i=1}^{s} x(i)\frac{\delta p_s}{\delta x(i)} \\ &= \sum_{i=1}^{s} \pi_i a^i + \mu \sum_{i=1}^{s} \pi_i a^i \sum_{t=i}^{s} x(t) = \\ &= m(s) + \mu \sum_{t=1}^{s} x(t) m(t), \end{aligned}$$

where the weights for the observations are

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

An alarm is given as soon as

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

exceeds the limit given by the LR method in Section 4.1

$$\begin{aligned} &[G_s/g(s) - m(s)]/\mu = \\ &= [a^{s+1} P(\tau > s)\frac{K}{1-K} - m(s)]/\mu. \end{aligned}$$

If the intensity is constant, then 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 for $b \neq 1$

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

and

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

If $b = 1, m(t) = t\nu/(1-\nu)$ and the relative weights will tend to one when t tends to infinity. Also for $b < 1$, the relative weights will tend to one when t tends to infinity. For $b > 1$ the relative weight tends to b and we have exponential weights.

For $b \neq 1$, with standardization to make the sum of the weights equal to one, we have

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

and the alarm is triggered if

$$\sum_{t=1}^{s} w_{\text{LinLR}}(s,t) x(t) > L_{\text{LinLR}}$$

where

$$L_{\text{LinLR}} = \left[\frac{b^s}{b(b^s-1) - s(b-1)}\left\{\frac{(b-1)^2 K}{\nu(1-K)} - (b-1)\right\} + \frac{b-1}{b(b^s-1) - s(b-1)}\right] \Big/ \mu.$$

This linear approximation of the LR method is here denoted as the **LinLR** method. For large s and $b > 1$, the alarm limit tends to

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

which is proportional to $b^s/(b^s-1)$.

When the weights, which are proportional to $(b^s - 1)$, are approximated by exponential weights the method will be named the **EWLR** method.

Another approximation is achieved when the Taylor expansion is made for the logarithm of the alarm function

$$\begin{aligned} \ln p_s(x_s) \approx \ln p_s^*(x_s) &= \ln p_s(0) + \sum_{i=1}^{s} x(i)\frac{\delta \ln p_s}{\delta x(i)} \\ &= \ln m(s) + \mu \sum_{t=1}^{s} x(t) m(t)/m(s). \end{aligned}$$

An alarm is given as soon as

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

exceeds the limit

$$\begin{aligned} &\ln\{G_s/g(s)\} - \ln m(s)]/\mu\} = \\ &= \left[\ln\left[a^{s+1} P(\tau > s)\frac{K}{1-K}\right] - \ln(m(s))\right] m(s)/\mu. \end{aligned}$$

For $b \neq 1$ we have the alarm limit

$$\begin{aligned} &= \left[\ln\left(\exp(\mu^2/2) b^s \frac{K}{1-K}\right) - \ln(m(s))\right] m(s)/\mu = \\ &= \left[\mu^2/2 + s\ln b + \ln\frac{K}{1-K} - \mu^2/2 - \ln\frac{\nu(b^s-1)}{(b-1)}\right] m(s)/\mu \\ &= \left[s\ln b + \ln\frac{K}{1-K} - \ln\frac{\nu(b^s-1)}{(b-1)}\right] m(s)/\mu \\ &= \left[s\ln b + \ln\frac{K}{1-K} - \ln\frac{\nu(b^s-1)}{(b-1)}\right] \frac{b\nu(b^s-1)}{(b-1)(1-\nu)\mu}. \end{aligned}$$

With weights standardized to have the sum 1 we have the alarm limit

$$\left[s\ln b+\ln\frac{K}{1-K}-\ln\frac{\nu(b^s-1)}{(b-1)}\right]\frac{b\nu(b^s-1)}{(b-1)(1-\nu)\mu}\frac{(b-1)(1-\nu)}{b\nu}\frac{b-1}{b(b^s-1)-s(b-1)}=$$

$$=\left[s\ln b+\ln\frac{K}{1-K}-\ln\frac{\nu(b^s-1)}{(b-1)}\right]\frac{(b^s-1)}{\mu}\frac{b-1}{b(b^s-1)-s(b-1)}=$$

$$=\left[\ln\frac{K}{1-K}-\ln\frac{\nu}{(b-1)}+s\ln b-\ln(b^s-1)\right]\frac{(b^s-1)}{\mu}\frac{b-1}{b(b^s-1)-s(b-1)}=$$

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

where the constant L, to be determined by false alarm properties, is

$$L=\ln\frac{K}{1-K}-\ln\frac{\nu}{(b-1)}=\ln\frac{k}{1-K}-\ln\nu+\ln\frac{\lambda}{1-\lambda}.$$

This method, with exponential weights is named the **EWlnLR** method.

Acknowledgements

This work was supported by the Swedish Council for Research in the Humanities and Social Sciences. The author thanks Samad Hedayat, Hans van Houwelingen, Christian Sonesson and Muni Srivastava for their interest and helpful comments. The editor has given constructive instructions.

References

Acosta-Mejia, C.A. (1998). Monitoring reduction in variability with the range. *IIE Transactions*, **30**, 515–523.
Acosta-Mejia, C.A., Pignatello, J.J.J. & Rao, B.V. (1999). A comparison of control charting procedures for monitoring process dispersion. *IIE Transactions*, **31**, 569–579.
Aerne, L.A., Champ, C.W. & Rigdon, S.E. (1991). Evaluation of Control Charts Under Linear Trend. *Communications in Statistics - Theory and Methods*, **20**, 3341–3349.
Amin, R.W., Wolff, H., Besenfelder, W. & Baxley, R. (1999). EWMA control charts for the smallest and largest observations. *Journal of Quality Technology*, **31**, 189–206.
Andersson, E. (2002). Monitoring cyclical processes. A nonparametric approach. *Journal of Applied Statistics*, **29**, 973–990.
Andersson, E., Bock, D. & Frisén, M. (2001). Likelihood based methods for detection of turning points in business cycles. A comparative study. Research Report, 2001:5 Department of Statistics, Göteborg University.
Arteaga, C. & Ledolter, J. (1997). Control charts based on order-restricted tests. *Statistics & Probability Letters*, **32**, 1–10.
Banks, D. (1993). Is Industrial Statistics Out of Control? *Statistical Science*, **8**, 356–409.
Barnett, V. & Turkman, K.F. (1993). *Statistics for the Environment*. Wiley.
Basseville, M. & Nikiforov, I. (1993). *Detection of Abrupt changes—Theory and Application*. Prentice Hall.
Bell, C., Gordon, L. & Pollak, M. (1994). An Efficient Nonparametric Detection Scheme and Its Application to Surveillance of a Bernoulli Process with Unknown Baseline. In *Change-point Problems*, Eds. E. Carlstein, H.-G. Miller and D. Siegmund, pp. 7–27. IMS Lecture Notes, Hayward.
Bojdecki, T. (1979). Probability maximizing approach to optimal stopping and its application to a disorder problem. *Stochastics*, **3**, 61–71.
Borror, C.M., Montgomery, D.C. & Runger, G.C. (1999). Robustness of the EWMA control chart to non-normality. *Journal of Quality Technology*, **31**, 309–316.
Brodsky, B.E. & Darkhovsky, B.S. (1993). *Nonparametric methods in change point problems*. Dordrecht: Kluwer Academic Publishers.
Brown, R.L., Durbin, J. & Evans, J.M. (1975). Techniques for Testing the Constancy of Regression Relationships over Time. *Journal of the Royal Statistical Society B*, **37**, 149–192.
Cardinal, M., Roy, R. & Lambert, J. (1999). On the application of integer-valued time series models for the analysis of disease incidence. *Statistics in Medicine*, **18**, 2025–2039.
Carlstein, E., Müller, H.G. & Siegmund, D. (1994). *Change-point problems*. Inst. of Mathematical Statistics, California.
Carlyle, W.M., Montgomery, D.C. & Runger, G.C. (2000). Optimization Problems and Methods in Quality Control and Improvement. *Journal of Quality Technology*, **32**, 1–19.
Chakraborti, S., van der Laan, P. & Bakir, S.T. (2001). Nonparametric Control Charts: An Overview and Some Results. *Journal of Quality Technology*, **33**, 304–315.
Champ, C.W., Woodall, W.H. & Mohsen, H. (1991). A generalized quality control procedure. *Statistics & Probability Letters*, **11**, 211–218.

Chan, L.K. & Zhang, J. (2000). Some issues in the design of EWMA charts. *Communications in Statistics - Simulations and Computations*, **29**, 207–217.
Chang, T.C. & Gan, F.F. (1995). A Cumulative Sum Control Chart For Monitoring Process Variance. *Journal of Quality Technology*, **27**, 109–119.
Chu, C.-S.J., Stinchcombe, M. & White, H. (1996). Monitoring structural change. *Econometrica*, **64**, 1045–1065.
Crosier, R.B. (1988). Multivariate Generalizations of Cumulative Sum Quality-Control Schemes. *Technometrics*, **30**, 291–303.
Crowder, S. (1987). A simple method for studying run-length distributions of exponentially weighted moving average charts. *Technometrics*, **29**, 401–407.
Crowder, S., Hawkins, D.M., Reynolds Jr., M.R. & Yashchin, E. (1997). Process Control and statistical Inference. *Journal of Quality Technology*, **29**, 134–139.
de Maré, J. (1980). Optimal Prediction of Catastrophes with Application to Gaussian Processes. *Annals of Probability*, **8**, 841–850.
Dewachter, H. (2001). Can Markov switching models replicate chartist profits in the foreign exchange market? *Journal of International Money and Finance*, **20**, 25–41.
Domangue, R. & 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**, 363–380.
Flury, B.D., Nel, D.G. & Pienaar, I. (1995). Simultaneous Detection of Shift in Means and Variances. *Journal of the American Statistical Association*, **90**, 1474–1481.
Frisén, M. (1980). U-shaped regression. In *Compstat, Proceedings in computational statistics*, pp. 304–307.
Frisén, M. (1986). Unimodal regression. *The Statistician*, **35**, 479–485.
Frisén, M. (1992). Evaluations of Methods for Statistical Surveillance. *Statistics in Medicine*, **11**, 1489–1502.
Frisén, M. (2000). Statistical Surveillance of Business Cycles. Research Report, 1994:3 Revised, Department of Statistics, Göteborg University.
Frisén, M. & de Maré, J. (1991). Optimal Surveillance. *Biometrika*, **78**, 271–280.
Frisén, M. & Sonesson, C. (2002). Optimal surveillance based on exponentially weighted moving averages. Research Report, 2002:1 Department of Statistics, Göteborg University.
Frisén, M. & Wessman, P. (1999). Evaluations of likelihood ratio methods for surveillance. Differences and robustness. *Communications in Statistics - Simulations and Computations*, **28**, 597–622.
Gan, F. F. (1992). Cusum Control Charts Under Linear Drift. *Statistician*, **41**, 71–84.
Gan, F. F. (1993). An optimal design of EWMA control charts based on median run-length. *Journal of Statistical Computation and Simulation*, **45**, 169–184.
Gan, F. F. (1995). Joint Monitoring of Process Mean and Variance Using Exponentially Weighted Moving Average Control Charts. *Technometrics*, **37**, 446–453.
Gan, F. F. (1998). Designs of one- and two-sided exponential EWMA charts. *Journal of Quality Technology*, **30**, 55–69.
Girshick, M. A. & Rubin, H. (1952). A Bayes approach to a quality control model. *Annals of Mathematical Statistics*, **23**, 114–125.
Gombay, E. (2000). Sequential change-point detection with likelihood ratios. *Statistics & Probability Letters*, **49**, 195–204.
Gordon, L. & Pollak, M. (1997). Average run length to false alarm for surveillance schemes designed with partially specified pre-change distribution. *Annals of Statistics*, **25**, 1284–1310.
Hackl, P. & Ledolter, J. (1991). A Control Chart Based On Ranks. *Journal of Quality Technology*, **23**, 117–124.
Hamilton, J.D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, **57**, 357–384.
Harrison, P.J. & Stevens, C.F. (1976). Bayesian forecasting, with discussion. *Journal of the Royal Statistical Society B*, **38**, 205–247.
Harrison, P.J. & Veerapen, P.P. (1994). A Bayesian Decision Approach to Model Monitoring and Cusums. *Journal of Forecasting*, **13**, 29–36.
Hawkins, D.M. (1991). Multivariate Quality Control Based on Regression-Adjusted Variables. *Technometrics*, **33**, 61–75.
Hawkins, D.M. & Olwell, D.H. (1998). *Cumulative Sum Charts and Charting for Quality Improvement*. New York, NY: Springer.
Hayter, A.J. & Tsui, K.L. (1994). Identification and Quantification in multivariate quality control problems. *Journal of Quality Technology*, **26**, 197–208.
Hinkley, D.V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika*, **57**, 1–17.
Hotelling, H. (1947). Multivariate Quality Control. In *Techniques of statistical analysis*, Eds. C. Eisenhart, M.W. Hastay and W.A. Wallis. New York: McGraw-Hill.
James, B., James, K.L. & Siegmund, D. (1987). Tests for a change-point. *Biometrika*, **74**, 71–83.
Jones, L.A. & Woodall, W.H. (1998). The Performance of Bootstrap Control Charts. *Journal of Quality Technology*, **30**, 362–375.
Järpe, E. (1999). Surveillance of the Interaction Parameter in the Ising Model. *Communications in Statistics - Theory and Methods*, **28**, 3009–3025.
Järpe, E. (2000a). Detection of environmental catastrophes. Research Report, 2000:6 Department of Statistics, Göteborg University.
Järpe, E. (2000b). On univariate and spatial surveillance. Ph.D Thesis, Department of Statistics Göteborg University.
Järpe, E. & Wessman, P. (2000). Some power aspects of methods for detecting shifts in the mean. *Communications in Statistics*

- *Simulations and Computations*, **29**, 633–646.
Kemp, K.W. (1961). The average run length of the cumulative sum chart when a V-mask is used. *Journal of the Royal Statistical Society B*, **23**, 149–153.
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*, pp. 1–21.
Kourti, T. & MacGregor, J.F. (1996). Multivariate SPC methods for process and product monitoring. *Journal of Quality Technology*, **28**, 409–428.
Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics - Theory and Methods*, **26**, 1481–1496.
Lai, T.L. (1995). Sequential Changepoint Detection in Quality-Control and Dynamical-Systems. *Journal of the Royal Statistical Society B*, **57**, 613–658.
Lai, T.L. (1998). Information Bounds and Quick Detection of Parameters in Stochastic Systems. *IEEE Transactions on Information Theory*, **44**, 2917–2929.
Lai, T.L. & Shan, Z. (1999). Efficient Recursive Algorithms for Detection of Abrupt Changes in Signals and Control Systems. *IEEE Transactions on Automatic Control*, **44**, 952–966.
Lao, C.S., Kessler, L.G. & Gross, T.P. (1998). Proposed statistical methods for signal detection of adverse medical device events. *Drug Information Journal*, **32**, 183–191.
Lawson, A., Böhning, D., Lesaffre, E., Biggeri, A., Viel, J.-F. & Bertollini, R. (1999). *Disease mapping and risk assessment for public health*. Wiley.
Lawson, A. B. (2001). Comments on the papers by Williams *et al.*, Kulldorf, Knorr-Held, Best & Rogerson. *Journal of the Royal Statistical Society A*, **164**, 97–99.
Lindgren, G. (1985). Optimal prediction of level crossings in Gaussian processes and sequences. *Annals of Probability*, **13**, 804–824.
Liu, R.Y. (1995). Control charts for multivariate processes. *Journal of the American Statistical Association*, **90**, 1380–1388.
Liu, R.Y. & Tang, J. (1996). Control charts for dependent and independent measurements based on bootstrap methods. *Journal of the American Statistical Association*, **91**, 1694–1707.
Lorden, G. (1971). Procedures for reacting to a change in distribution. *Annals of Mathematical Statistics*, **42**, 1897–1908.
Lowry, C.A. & Montgomery, D.C. (1995). A Review of Multivariate Control Charts. *IIE Transactions*, **27**, 800–810.
Lowry, C.A., Woodall, W.H., Champ, C.W. & Rigdon, S.E. (1992). A multivariate exponentially weighted moving average control chart. *Technometrics*, **34**, 46–53.
Lu, C.W. & Reynolds Jr., M.R. (1999). EWMA control charts for monitoring the mean of autocorrelated processes. *Journal of Quality Technology*, **31**, 166–188.
Lu, C.W. & Reynolds Jr., M.R. (2001). Cusum Chart For Monitoring An Autocorrelated Process. *Journal of Quality Technology*, **33**, 316–334.
Lucas, J.M. & Crosier, R.B. (1982). Robust CUSUM: A robustness study for cusum quality control schemes. *Communications in Statistics - Theory and Methods*, **11**, 2669–2687.
Lucas, J.M. & Saccucci, M.S. (1990). Exponentially weighted moving average control schemes: properties and enhancements. *Technometrics*, **32**, 1–12.
Margavio, T.M., Conerly, M.D., Woodall, W.H. & Drake, L.G. (1995). Alarm Rates For Quality-Control Charts. *Statistics & Probability Letters*, **24**, 219–224.
McLaren, C.E., Kambour, E.L., McLachlan, G.J., Lukaski, H.C., Li, X., Brittenham, G.M. & McLaren, G.D. (2000). Patient-specific analysis of sequential haematological data by multiple linear regression and mixture distribution modelling. *Statistics in Medicine*, **19**, 83–98.
Mevorach, Y. & Pollak, M. (1991). A small sample size comparison of the CUSUM and Shiryaev–Roberts approaches to changepoint detection. *American Journal of Mathematics and Management*, **11**, 277–298.
Montgomery, D.C. & Woodall, W.H. (1997). A Discussion on Statistically-Based Process Monitoring and Control. *Journal of Quality Technology*, **29**, 121–122.
Morais, M.C. & Pacheco, A. (1998). Two stochastic properties of one-sided exponentially weighted moving average control charts. *Communications in Statistics - Simulations and Computations*, **27**, 937–952.
Morais, M.C. & Pacheco, A. (2000). On the performance of combined EWMA schemes for mu and sigma: A Markovian approach. *Communications in Statistics - Simulations and Computations*, **29**, 153–174.
Moustakides, G.V. (1986). Optimal stopping times for detecting changes in distributions. *The Annals of Statistics*, **14**, 1379–1387.
Neube, M. & Li, K. (1999). An EWMA-CUSCORE quality control procedure for process variability. *Mathematical and Computer Modelling*, **29**, 73–79.
Nelson, L.S. (1990). Monitoring Reduction in Variation with a Range Chart. *Journal of Quality Technology*, **22**, 163–165.
Ng, C.H. & Case, K.E. (1989). Development and Evaluation of Control Charts Using Exponentially Weighted Moving Averages. *Journal of Quality Technology*, **21**, 242–250.
Padgett, W.J. & Spurrier, J.D. (1990). Shewhart-Type Charts for Percentiles of Strength Distributions. *Journal of Quality Technology*, **22**, 283–288.
Page, E.S. (1954). Continuous inspection schemes. *Biometrika*, **41**, 100–114.
Pettersson, M. (1998a). Evaluation of some methods for statistical surveillance of an autoregressive process. Research Report, 1998:4 Department of Statistics, Göteborg University.
Pettersson, M. (1998b). Monitoring a freshwater fish population: Statistical surveillance of biodiversity. *Environmetrics*, **9**, 139–150.
Pollak, M. (1985). Optimal detection of a change in distribution. *Annals of Mathematical Statistics*, **13**, 206–227.
Pollak, M. & Siegmund, D. (1975). Approximations to the Expected Sample Size of Certain Sequential Tests. *Annals of Statistics*, **3**, 1267–1282.

Pollak, M. & Siegmund, D. (1985). A diffusion process and its applications to detecting a change in the drift of Brownian motion. *Biometrika*, **72**(2), 267–280.
Pollak, M. & Siegmund, D. (1991). Sequential detection of a change in a normal mean when the initial value is unknown. *Annals of Statistics*, **19**, 394–416.
Ramalhoto, M.F. & Morais, M. (1999). Shewhart control charts for the scale parameter of a Weibull control variable with fixed and variable sampling intervals. *Journal of Applied Statistics*, **26**, 129–160.
Ritov, Y. (1990). Decision theoretic 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**, 411–430.
Robinson, P.B. & Ho, T.Y. (1978). Average Run Lengths of Geometric Moving Average Charts by Numerical Methods. *Technometrics*, **20**, 85–93.
Rogerson, P.A. (2001). Monitoring point patterns for the development of space-time clusters. *Journal of the Royal Statistical Society A*, **164**, 87–96.
Rossi, G., Lampugnani, L. & Marchi, M. (1999). An approximate CUSUM procedure for surveillance of health events. *Statistics in Medicine*, **18**, 2111–2122.
Royston, P. (1991). Identifying the fertile phase of the human menstrual cycle. *Statistics in Medicine*, **10**, 221–240.
Runger, G.C. & Prabhu, S.S. (1996). A Markov chain model for the multivariate exponentially weighted moving averages control chart. *Journal of the American Statistical Association*, **91**, 1701–1706.
Saniga, E.M. (1989). Economic Statistical control Chart Designs With an Application to X and R Charts. *Technometrics*, **31**, 313–320.
Schmid, W. (1997). CUSUM control schemes for Gaussian processes. *Statistical Papers, Berlin*, **38**, 191–217.
Schmid, W. & Schöne, A. (1997). Some Properties of the EWMA Control Chart in the Presence of Autocorrelation. *The Annals of Statistics*, **25**, 1277–1283.
Scranton, R., Runger, G.C., Keats, J.B. & Montgomery, D.C. (1996). Efficient shift detection using multivariate exponentially-weighted moving average control charts and principal components. *Quality and Reliability Engineering International*, **12**, 165–171.
Shewhart, W.A. (1931). *Economic Control of Quality of Manufactured Product*. London: MacMillan and Co.
Shiryaev, A.N. (1963). On optimum methods in quickest detection problems. *Theory of Probability and its Applications*, **8**, 22–46.
Siegmund, D. (1985). *Sequential analysis. Tests and confidence Intervals*. Springer.
Smith, A.F. & West, M. (1983). Monitoring Renal Transplants: An Application of the Multiprocess Kalman Filter. *Biometrics*, **39**, 867–878.
Sonesson, C. (2001). Evaluations of some exponentially weighted moving average methods. Research Report, 2001:6 Department of Statistics, Göteborg University.
Sonesson, C. & Bock, D. (2003). A review and discussion of prospective statistical surveillance in public health. *Journal of the Royal Statistical Society A*, **166**, 5–21.
Srivastava, M.S. (1997). Cusum procedures for monitoring variability. *Communications in Statistics - Theory and Methods*, **26**, 2905–2926.
Srivastava, M.S. & Wu, Y. (1993). Comparison of EWMA, CUSUM and Shiryayev–Roberts Procedures for Detecting a Shift in the Mean. *Annals of Statistics*, 21, 645–670.
Steiner, S.H. (1999). EWMA control charts with time-varying control limits and fast initial response. *Journal of Quality Technology*, **31**, 75–86.
Stoumbos, Z.G. & Reynolds, M.R. (2000). Robustness to non-normality and autocorrelation of individuals control charts. *Journal of Statistical Computation and Simulation*, **66**, 145–187.
Stoumbos, Z.G., Reynolds, M.R., Ryan, T.P. & Woodall, W.H. (2000). The state of statistical process control as we proceed into the 21st century. *Journal of the American Statistical Association*, **95**, 992–998.
Sveréus, A. (1995). Detection of successive changes. Statistical methods in postmarketing surveillance. Research Report, 1995:2 Department of Statistics, Göteborg University.
Telksnys, L. (1986). *Detection of changes in random processes*. New York: Springer.
Timm, N.H. (1996). Multivariate quality control using finite intersection tests. *Journal of Quality Technology*, **28**, 233–243.
Tsui, K.L. & Woodall, W.H. (1993). Multivariate Control Charts Based on Loss Functions. *Sequential Analysis*, **12**, 79–92.
van Dobben de Bruyn, C. S. (1968). *Cumulative sum Tests: Theory and Practice*. Griffin.
VanBrackle, L. & Williamson, G. (1999). A study of the average run length characteristics of the National Notifiable Diseases Surveillance System. *Statistics in Medicine*, **18**, 3309–3319.
VanBrackle, L.M. & Reynolds, M.R. (1997). EWMA and Cusum Control Charts in the Presence of Correlation. *Communications in Statistics - Simulations and Computations*, **26**, 979–1008.
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.
von Collani, E. & Sheil, J. (1989). An Approach to Controlling Process Variability. *Journal of Quality Technology*, **21**, 87–96.
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.
Wessman, P. (1999). The surveillance of several processes with different change points. Research Report, 1999:2 Department of Statistics, Göteborg University.
Wetherill, G.B. & Brown, D.W. (1991). *Statistical process control*. Chapman and Hall.
Williamson, G. & Hudson, G. (1999). A monitoring system for detecting aberrations in public health surveillance reports.

Statistics in Medicine, **18**, 3283–3298.
Woodall, W.H. (1997). Control Charts Based on Attribute Data: Bibliography and Review. *Journal of Quality Technology*, **29**, 172–183.
Woodall, W.H. & Montgomery, D.C. (1999). Research Issues and Ideas in Statistical Process Control. *Journal of Quality Technology*, **31**, 376–386.
Woodall, W.H. & Ncube, M.M. (1985). Multivariate Cusum Quality Control Procedures. *Technometrics*, **27**, 285–292.
Yakir, B., Krieger, A.M. & Pollak, M. (1999). Detecting a change in regression: First-order optimality. *Annals of Statistics*, **27**, 1896–1913.
Yashchin, E. (1993). Statistical Control Schemes—Methods, Applications and Generalizations. *International Statistical Review*, **61**, 41–66.
Yashchin, E. (1995). Likelihood ratio methods for monitoring parameters of a nested random effect model. *Journal of the American Statistical Association*, **90**, 729–738.
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*, pp. 245–269.

Résumé

Des différents critères d'optimalité sont utilisés dans différentes subcultures de la surveillance statistique. Un des objectifs de cette étude est celui d'établir un rapprochement entre les différentes disciplines. Les faults de quelques uns des critères d'optimalité sont montrés par leurs implications. Quelques méthodes fréquemment utilisées sont examinées en détail quant à leur optimalité. Cet analyse est fait pour une situation standard, se concentrant sur les principes d'inférence. Une présentation uniforme des méthodes, par expressions de rapports de vraisemblance, facilite la comparaison entre les méthodes. On examine les correspondances entre les critères d'optimalité et les méthodes. On présente une approximation linéaire de la méthode du rapport de vraisemblance totale, qui satisfait plusieurs critères de optimalité. Cette approximation linéaire est utilisée pour examiner quand les méthodes linéaires sont approximativent optimales. Des méthodes pour des situations compliquées sont étudiées quant à leur optimalité est robustesse.

[*Received May 2002, accepted January 2003*]

International Statistical Review (2003), 71, 2, 435–459, Printed in The Netherlands
© International Statistical Institute

Bootstrap Methods for Time Series

Wolfgang Härdle[1], Joel Horowitz[2] and Jens-Peter Kreiss[3]

[1]CASE - Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Berlin, Germany. [2]Department of Economics, Northwestern University, Evanston, IL, USA. [3]Institute for Mathematical Stochastics, Technical University of Braunschweig, Braunschweig, Germany

Summary

The bootstrap is a method for estimating the distribution of an estimator or test statistic by resampling one's data or a model estimated from the data. The methods that are available for implementing the bootstrap and the accuracy of bootstrap estimates depend on whether the data are an independent random sample or a time series. This paper is concerned with the application of the bootstrap to time-series data when one does not have a finite-dimensional parametric model that reduces the data generation process to independent random sampling. We review the methods that have been proposed for implementing the bootstrap in this situation and discuss the accuracy of these methods relative to that of first-order asymptotic approximations. We argue that methods for implementing the bootstrap with time-series data are not as well understood as methods for data that are independent random samples. Although promising bootstrap methods for time series are available, there is a considerable need for further research in the application of the bootstrap to time series. We describe some of the important unsolved problems.

Key words: Resampling; Block bootstrap; Asymptotic approximation confidence interval.

1 Introduction

The bootstrap is a method for estimating the distribution of an estimator or test statistic by resampling one's data or a model estimated from the data. Under conditions that hold in a wide variety of applications, the bootstrap provides approximations to distributions of statistics, coverage probabilities of confidence intervals, and rejection probabilities of tests that are at least as accurate as the approximations of first-order asymptotic distribution theory. Often, the bootstrap provides approximations that are more accurate than those of first-order asymptotic theory. This is important in applied research, because first-order asymptotic approximations are often inaccurate with the sample sizes found in applications.

The methods that are available for implementing the bootstrap and the improvements in accuracy that it achieves relative to first-order asymptotic approximations depend on whether the data are an independently and identically distributed (*iid*) random sample or a time series. If the data are *iid*, the bootstrap can be implemented by sampling the data randomly with replacement or by sampling a parametric model of the distribution of the data. The distribution of a statistic is estimated by its empirical distribution under sampling from the data or parametric model. Beran & Ducharme (1991), Hall (1992), Efron & Tibshirani (1993), and Davison & Hinkley (1997) provide detailed discussions of bootstrap methods and their properties for *iid* data.

The situation is more complicated when the data are a time series because bootstrap sampling must be carried out in a way that suitably captures the dependence structure of the data generation process (DGP). This is not difficult if one has a finite-dimensional parametric model (e.g., a finite-order ARMA model) that reduces the DGP to independent random sampling. In this case and under