Behavior of EWMA type control charts for small smoothing parameters

Taras Lazariv [a], Yarema Okhrin [b,*], Wolfgang Schmid [a]

[a] *Department of Statistics, European University Viadrina, PO Box 1786, 15207 Frankfurt (Oder), Germany*

[b] *Department of Statistics, University of Augsburg, D-86159 Augsburg, Germany*

1. Introduction

The exponentially weighted moving average (EWMA) chart of Roberts (1959) is one of the most popular approaches to statistical surveillance. Its main advantage, especially valued by practitioners, is its simple and intuitive implementation based on the exponential smoothing of the historical values. Similar to the CUSUM chart of Page (1954), the smoothing parameter regulates the impact of past observations. Setting it equal to one leads to the Shewhart chart, cf. Shewhart (1931). Smaller values of the smoothing parameter lead to an increasing influence of the preceding observations. The EWMA scheme is mainly applied for monitoring the mean of independent data, but the procedure can be adapted in a straightforward way to monitor any characteristics of the process. The stochastic structure of the data generating process can be rather general too. A chart for the standard deviation of an independent random process was introduced by Crowder and Hamilton (1992). The extension of the EWMA chart to monitor the mean of a stationary time series was given by Schmid (1997). EWMA charts for the variance of a stationary process were proposed in Schipper and Schmid (2001) while Rosołowski and Schmid (2003) discussed simultaneous EWMA schemes for the mean, the variances, and the autocovariances.

The choice of the optimal smoothing parameter is in general an unsolved problem. Several authors have given recommendations for specific setups of the monitoring problem. Montgomery (2009) states that values between 0.05 and 0.25 work well in practice. Lucas and Saccucci (1990) determine, via simulations, the optimal value of the smoothing parameter which minimizes the out-of-control average run length (ARL) for a given value of the expected shift in the

* Corresponding author. Tel.: +49 821 5984152; fax: +49 821 5984227.
E-mail address: yarema.okhrin@wiwi.uni-augsburg.de (Y. Okhrin).

target process and for a given in-control ARL. The authors rely on a grid search over the interval [0.03;1]. The non-zero lower boundary in the study is due to numerical instabilities arising when the smoothing parameter approaches zero (see, e.g., Brook and Evans, 1972, Crowder, 1987).

Recent technological developments allow showing that the out-of-control ARL is decreasing in the smoothing parameter, implying that the limiting chart with the smoothing parameter approaching zero in the limit should be analysed more thoroughly. This problem was described by Chan and Zhang (2000). Morais et al. (forthcoming) established several important limit results for the distribution of the run length of the one-sided EWMA chart for the mean, assuming that the target process is independent and normally distributed. They proved that the in-control ARL of the EWMA scheme based on the exact variance is a decreasing function in the smoothing parameter, and showed that the limiting chart as the smoothing parameter approaches zero is equivalent to the repeated significance test. Note that both papers focus on the mean chart and independent data.

In the present paper we consider a very general family of EWMA schemes which can be used to monitor an arbitrary real-valued parameter of a stationary process. Because of its generality, this approach covers most of the EWMA schemes discussed in the literature. Furthermore, the suggested approach allows us to consider one-sided and two-sided charts within a single framework. We distinguish between the charts based on the exact variance, i.e., the control limits change in time, and the charts based on the asymptotic variance, i.e., the control limits are fixed. Since in the former case the variance must be recalculated at each time point, most practitioners prefer to work with the asymptotic variance.

A popular approach in modern statistical process control is to use a head start, i.e., to set the starting value of the control statistics not equal to the target value of the parameter. If correctly selected, the head start helps to improve the performance of the charts. We show in Section 3 that the run lengths of the charts based on the exact variance converge in distribution to the run length of the repeated significance test if the head start is proportional to the smoothing parameter. If the head start is constant, as it usually is, then the control chart degenerates, i.e., it gives either always or never signals. In the most realistic case, the probability of a signal up to a fixed point in time converges to 1 as well in the in-control state as in the out-of-control state. This is a very unpleasant property. Moreover, it shows that the ARL converges to infinity if the smoothing parameter tends to zero. For the EWMA scheme based on the asymptotic variance, an additional normalization of the procedure is required. Section 4 contains illustrative examples, which explain the limiting performance and the damaging impact of a constant head start.

2. Modified EWMA charts for stationary processes

In the following, we consider a very general surveillance problem. We make use of an EWMA type control scheme. Suppose that observations $x_1, x_2, \ldots$ of the real-valued stochastic process $\{X_t\}$ defined on the probability space $(\Omega, \mathcal{F}, P)$ are sequentially taken. Suppose that the distribution of $(X_1, \ldots, X_t)$ depends on a real-valued parameter θ_t. For instance, θ_t could be the mean of the process, its variance, etc. Let $\theta_t = \theta$ for all $t = 1, \ldots, q-1$ and $\theta_t \neq \theta$ for $t \geq q$. Here θ is assumed to be a known value. The process $\{X_t\}$ is said to be in-control if $q = \infty$ and we refer to the corresponding process as the in-control process. If $q < \infty$, then $\{X_t\}$ is said to be out of control. Note that here we do not make any further assumption on $\{X_t\}$.

Our aim is to detect whether the parameter changes over time or whether it is constant. In order to quickly detect the change, we use a sequential procedure and after every new observation, analyse whether a change has occurred or not.

Suppose that T_t is a point estimator of θ_t based on the information set available at time point t, i.e. $T_t = f_t(X_1, \ldots, X_t)$ with f_t a measurable function. Assume that T_t is an unbiased estimator of θ in the in-control state, i.e. $E(T_t) = \theta$ for all $t \geq 1$ provided that $q = \infty$. If θ_t is the mean of $\{X_t\}$, then we can choose, e.g., $T_t = X_t$ (Schmid, 1997), $T_t = \sum_{v=1}^{k} X_{t+1-v}/k$ or $T_t = c_1 med\{X_{t+1-k}, \ldots, X_t\}$. The shift from the in-control to the out-of-control state is frequently described by the change-point model satisfying $E(X_t) = \mu + aI_{\{q,q+1,\ldots\}}(t)$ with $a \in I\mathbb{R} - \{0\}$ and μ being the in-control mean of $\{X_t\}$. In case θ_t is the variance of $\{X_t\}$, possible choices would be, e.g., $T_t = (X_t - \mu)^2$ (e.g., Schipper and Schmid, 2001) or $T_t = \sum_{v=1}^{k}(X_{t+1-v} - \mu)^2/k$. The corresponding change-point model for the variance is often chosen to guarantee $\mathrm{Var}(X_t) = \Delta\sigma^2$ for $t \geq q$ and $\Delta \in (0, \infty) - \{1\}$, with σ^2 being the in-control variance of the process.

The control statistic is obtained by applying the EWMA recursion to T_t. This leads to

$$Z_t = \begin{cases} Z_0, & t = 0 \\ (1-\lambda)Z_{t-1} + \lambda T_t, & t = 1, 2, \ldots, \end{cases} \tag{1}$$

with an initial value Z_0. Z_0 is assumed to be a deterministic quantity. The parameter λ is a smoothing parameter taking values within (0, 1]. It equals the weight of the most recent observed value. By recursive substitution, it can be shown that Z_t is a weighted sum of the observations available at time point $t \geq 1$:

$$Z_t = \lambda \sum_{i=0}^{t-1} (1-\lambda)^i T_{t-i} + (1-\lambda)^t Z_0 \tag{2}$$

$$= \theta + \lambda \sum_{i=0}^{t-1} (1-\lambda)^i (T_{t-i} - \theta) + (1-\lambda)^t (Z_0 - \theta). \tag{3}$$

This representation will be useful in the discussion below. Consequently a value of λ close to one leads to a short memory EWMA chart and for $\lambda = 1$ the Shewhart chart is obtained. Taking λ close to zero leads to EWMA charts that give little importance to the most recent observations. Because the weights do not decrease as fast as for larger values of the smoothing parameter, more distant observations have a stronger influence on the present value. Moreover, for smaller values of λ, the starting value Z_0 has a stronger impact and if λ tends to zero, Z_t converges to Z_0 provided that Z_0 is a constant.

Since T_t is an unbiased estimator of θ, we get in the in-control state that

$$E_\infty(Z_t) = \theta + (1-\lambda)^t[Z_0 - \theta].$$

The index "∞" means, throughout the remainder of this paper, that the quantity (an expectation, a variance, a covariance, a probability, etc.) is calculated with respect to the in-control situation.

To calculate the variance of the EWMA statistic based on $\{T_t\}$ we need stronger assumptions about the stochastic properties of $\{T_t\}$. We suppose that in the in-control state $\{T_t\}$ is a (weakly) stationary process with mean θ and autocovariance function $\{\gamma_{T,h}\}$. Then

$$\begin{aligned} \mathrm{Var}_\infty(Z_t) &= \lambda^2 \sum_{i,j=0}^{t-1} (1-\lambda)^{i+j} \gamma_{T,|i-j|} \\ &= \frac{\lambda}{2-\lambda}\left[[1-(1-\lambda)^{2t}]\gamma_{T,0} + 2\sum_{v=1}^{t-1}(1-\lambda)^v[1-(1-\lambda)^{2(t-v)}]\gamma_{T,v} \right]. \end{aligned} \tag{4}$$

Note that here we only need $\{T_t\}$ to be weakly stationary in the in-control state. In particular situations it is preferable to have direct assumptions on the observed process rather than on a transformed one. If, e.g., $\{X_t\}$ is strictly stationary in the in-control case, $T_t = f_t(X_1, \ldots, X_t)$ and f_t is a measurable function, and if the second moment of T_t in the in-control state exists, then it follows that $\{T_t\}$ is strictly as well as weakly stationary in the in-control state.

Because Z_0 is a deterministic value, it only influences the mean of Z_t but not its variance. In case $Z_0 = \theta$, the process starts in the in-control state and Z_t is an unbiased estimator of θ provided that the process is in control. If $Z_0 \neq \theta$, then the EWMA recursion has a head start. A head start permits a fast detection of an initial out-of-control situation and is useful after a process correction. Fast initial response features of EWMA charts for independent samples have been studied by, e.g., Steiner (1999) and Knoth (2005). Commonly Z_0 is taken as θ plus or minus a multiple of the asymptotic standard deviation of the EWMA statistic. Consequently we have the situation that Z_0 may depend on λ. This is taken into account in what follows.

The upper one-sided EWMA control chart is used to detect an upward shift. If

$$Z_t > \theta + c\sqrt{\mathrm{Var}_\infty(Z_t)},$$

for any $t \geq 1$, then it is concluded that the parameter θ has increased and the process is out-of-control. The quantity c is a fixed constant value and it determines the control limit. Note that we make use of the asymptotic mean θ instead of the exact one $E_\infty(Z_t)$.

If we are interested in detecting a change in any direction, then we apply two-sided EWMA charts. Then a signal is given if

$$|Z_t - \theta| > c\sqrt{\mathrm{Var}_\infty(Z_t)},$$

with $c > 0$.

All these schemes are based on the following decision rule. Let $\mathcal{R} \subset I\!\mathbb{R}$ be an arbitrary Borel set. Here it denotes the rejection area. Note that it does not depend on t. The EWMA chart gives a signal if

$$\frac{Z_t - \theta}{\sqrt{\mathrm{Var}_\infty(Z_t)}} \in \mathcal{R}.$$

We consider two different types of $\mathcal{R}$. For the upper one-sided, we set $\mathcal{R} = (c, \infty)$, while the two-sided case is obtained with $\mathcal{R} = (-\infty, -c) \cup (c, \infty)$. In general it is sufficient to assume that $\mathbb{R} - \mathcal{R}$ is an interval.

In the literature, most of the published papers on EWMA schemes make use of the asymptotic variance instead of the exact one. In that case the control limits are constant and do not depend on t. A justification of this procedure is provided by arguing that the exact variance quickly converges to the asymptotic one. However, this is only the case if λ is not small. The decision rule for the EWMA schemes based on the asymptotic variance gives a signal if

$$\frac{Z_t - \theta}{\sqrt{\lim_{t\to\infty} \mathrm{Var}_\infty(Z_t)}} \in \mathcal{R}.$$

Note that the limit of $\mathrm{Var}_\infty(Z_t)$ exists if $\sum_{v=1}^{\infty} |\gamma_{T,v}| < \infty$. Now,

$$\lim_{t\to\infty} \mathrm{Var}_\infty(Z_t) = \frac{\lambda}{2-\lambda}\left[\gamma_{T,0} + 2\sum_{v=1}^{\infty}(1-\lambda)^v \gamma_{T,v}\right].$$

In this paper we always assume that θ, $\mathrm{Var}_\infty(Z_t)$, and $\lim_{t\to\infty}\mathrm{Var}_\infty(Z_t)$ are known quantities. In practice, however, these quantities have to be estimated using a prerun. This influences the behavior of the run length, but goes beyond the scope of this paper (see, e.g., Alberts and Kallenberg, 2004, Jensen et al., 2006).

Example 1. (a) Monitoring the mean of a stationary process.
In this case we have $\theta = \mu$. Choosing $T_t = X_t$ we obtain the mean chart introduced by Schmid (1997). Assuming $\{X_t\}$ to be stationary, then in the in-control state, $\{T_t\}$ is stationary as well. For a causal ARMA process, the autocovariances of $\{X_t\}$ can be determined recursively by making use of the Yule–Walker equations (e.g., Brockwell and Davis, 1991, Chapter 3). For a stationary GARCH process $\{X_t\}$, it follows that

$$\mathrm{Var}_\infty(Z_t) = \frac{\lambda}{2-\lambda}[1-(1-\lambda)^{2t}]\gamma_{T,0}.$$

(b) Monitoring the variance of a stationary process.
Suppose that $\mu = 0, \theta = \gamma_0, T_t = X_t^2$ (e.g., Schipper and Schmid, 2001) and $X_t = \sum_{i=-\infty}^{\infty} a_i\varepsilon_{t-i}$, where $\{a_i\}$ is absolutely summable. Let $\{\varepsilon_t\}$ be independent and normally distributed with $E(\varepsilon_t) = 0$ and $\mathrm{Var}(\varepsilon_t) = \sigma_\varepsilon^2$. Then $\gamma_{T,h} = \gamma_0^2 + 2\gamma_h^2$ (cf. Brockwell and Davis, 1991, p. 227), where γ_h stands for the autocovariance function of $\{X_t\}$. Thus, we get

$$\mathrm{Var}_\infty(Z_t) = \frac{\lambda}{2-\lambda}\left[2\sum_{i=1}^{t-1}(1-\lambda)^i[1-(1-\lambda)^{2(t-i)}]\gamma_i^2 \right.$$
$$\left. + \gamma_0^2 \quad 3[1-(1-\lambda)^{2t}] + \frac{2(1-\lambda)}{\lambda}[1-(1-\lambda)^t][1-(1-\lambda)^{t-1}] \right].$$

For a stationary ARMA process, this quantity can be calculated recursively, as described in (a).

(c) Monitoring the autocovariances of a stationary process.
Let $h \geq 0$ be fixed and $\theta = \gamma_h = Cov(X_t, X_{t-h})$. Let $T_t = (X_t - \mu)(X_{t-h} - \mu)$. This chart is a special case of the control chart considered in Rosołowski and Schmid (2003).

In our paper we shall discuss the relation between the EWMA chart and the so-called repeated significance test. The repeated significance test relies on the decision rule

$$\frac{\bar{T}_t - \theta}{\sqrt{\mathrm{Var}_\infty(\bar{T}_t)}} \in \mathcal{R},$$

with $\bar{T}_t = \sum_{v=1}^{t} T_v/t$, i.e. the sequential application of the significance test with the rejection area $\mathcal{R}$. Thus, unlike the geometrically decaying weights in the EWMA scheme, the complete history is taken into account with equal weight. Note that

$$t\mathrm{Var}_\infty(\bar{T}_t) = \gamma_{T,0} + 2\sum_{v=1}^{t-1}(1-v/t)\gamma_{T,v}. \quad (5)$$

If the smoothing parameter tends to zero, then it can be shown that the variance of the EWMA control statistic converges to the variance of the repeated significance test statistic

$$\lim_{\lambda\to 0+}\frac{\mathrm{Var}_\infty(Z_t)}{\lambda^2} = t^2\mathrm{Var}_\infty(\bar{T}_t).$$

Moreover, provided that $\sum_{v=1}^{\infty}|\gamma_{T,v}| < \infty$,

$$\lim_{\lambda\to 0+}\frac{\lim_{t\to\infty}\mathrm{Var}_\infty(Z_t)}{\lambda} = \gamma_{T,0} + 2\sum_{v=1}^{\infty}\gamma_{T,v} = \lim_{t\to\infty} t\mathrm{Var}_\infty(\bar{T}_t).$$

3. Behavior for small smoothing parameter

In this section we analyse the performance of these control schemes if lambda tends to zero. A classical performance measure is the run length, and we first introduce the notation for these charts. Let $N(\lambda, \mathcal{R})$ denote the run length of the control chart based on the exact variance, i.e.

$$N(\lambda, \mathcal{R}) = \inf \quad t \in I\!N : \frac{Z_t - \theta}{\sqrt{\mathrm{Var}_\infty(Z_t)}} \in \mathcal{R} \quad ,$$

and let $N_{asymp}(\lambda, \mathcal{R})$ be the run length of the scheme with the asymptotic variance

$$N_{asymp}(\lambda, \mathcal{R}) = \inf \left\{ t \in I\mathbb{N} : \frac{Z_t - \theta}{\sqrt{\lim\limits_{t\to\infty} \mathrm{Var}_\infty(Z_t)}} \in \mathcal{R} \right\}.$$

The run length of the repeated significance test is denoted by

$$N(\mathcal{R}) = \inf \left\{ t \in I\mathbb{N} : \frac{\bar{T}_t - \theta}{\sqrt{\mathrm{Var}_\infty(\bar{T}_t)}} \in \mathcal{R} \right\}.$$

Moreover, let

$$N_{asymp}(\mathcal{R}) = \inf \left\{ t \in I\mathbb{N} : \frac{\sqrt{t}(\bar{T}_t - \theta)}{\sqrt{\lim\limits_{t\to\infty} t\mathrm{Var}_\infty(\bar{T}_t)}} \in \mathcal{R} \right\}.$$

3.1. The modified EWMA scheme based on the exact variance

Here we discuss the behavior of the general EWMA scheme introduced in Section 2 as the smoothing parameter tends to zero. We calculate the limit of the probability of getting no signal up to a fixed time point. It is very remarkable that the following theorems hold both in the in-control and in the out-of-control states. It shows how the limit depends on the choice of the head start. In principle, the relation between the distance of the head start from the target value, here $d(\lambda)$, and the smoothing parameter λ determines the asymptotic behavior. In practice we will either choose $d(\lambda)$ as a constant or proportional to the (asymptotic) standard deviation of Z_t.

In the following we use the symbol $\bar{\mathcal{R}}$ to denote the closure of $\mathcal{R}$. If, e.g., $\mathcal{R} = (c, \infty)$, then the closure is given by $\bar{\mathcal{R}} = [c, \infty]$ and for $\mathcal{R} = (-\infty, -c) \cup (c, \infty)$, it is equal to $\bar{\mathcal{R}} = [-\infty, -c] \cup [c, \infty]$.

Theorem 1. *Suppose that the distribution of $\{X_t\}$ does not depend on λ and that in the in-control state, $\{T_t\}$ is a weakly stationary process with mean θ. Let $Z_0 = Z_0(\lambda) = \theta + d(\lambda)$ and $\lim_{\lambda\to 0+} d(\lambda)/\lambda = d \in I\mathbb{R} \cup \{\infty, -\infty\}$. Let $I\mathbb{R} - \mathcal{R}$ be an interval and suppose that $\mathcal{R}$ does not depend on λ. Then for any fixed $k = 1, 2, \ldots,$*

$$\lim_{\lambda\to 0+} P[N(\lambda, \mathcal{R}) > k] = \begin{cases} 0 & \text{if} \quad d = \infty \quad \text{and} \quad \infty \in \bar{\mathcal{R}} \quad \text{or} \\ & \phantom{\text{if}} \quad d = -\infty \quad \text{and} \quad -\infty \in \bar{\mathcal{R}} \\ P[A] & \text{if} \quad d \in I\mathbb{R} \\ 1 & \text{if} \quad d = \infty \quad \text{and} \quad \infty \notin \bar{\mathcal{R}} \quad \text{or} \\ & \phantom{\text{if}} \quad d = -\infty \quad \text{and} \quad -\infty \notin \bar{\mathcal{R}} \end{cases}$$

with

$$A = \bigcap_{t=1}^{k} A_t = \bigcap_{t=1}^{k} \left\{ (x_1, \ldots, x_k) : \frac{\bar{T}_t(x_1, \ldots, x_t) - \theta + d/t}{\sqrt{\mathrm{Var}_\infty(\bar{T}_t)}} \notin \mathcal{R} \right\}.$$

Note that the above result is quite general and valid for any EWMA chart whose input statistics $\{T_t\}$ are governed by a weakly stationary process. This assumption is only needed to derive the in-control variance of the EWMA recursion. For proving Theorem 1 we make use of the fact that $P(A_t) \to 0$ as $t \to \infty$ if $A_1 \supset A_2 \supset \cdots \supset A_t \supset \ldots$ and $\cap_{t=1}^{\infty} A_t = \emptyset$. A detailed proof is given in the Appendix.

Corollary 1. *Suppose that the assumptions of Theorem 1 are satisfied.*

(a) *If $d = 0$, then*

$$\lim_{\lambda\to 0+} P[N(\lambda, \mathcal{R}) > k] = P[N(\mathcal{R}) > k]$$

and

$$\lim_{\lambda\to 0+} E[N(\lambda, \mathcal{R})] = E[N(\mathcal{R})].$$

(b) *If $d(\lambda)$ is constant but unequal to zero, say $d(\lambda) = \tau \in I\mathbb{R} - \{0\}$ for all λ, then*

$$\lim_{\lambda\to 0+} P[N(\lambda, \mathcal{R}) > k] = \begin{cases} 0 & \text{if} \quad \tau > 0 \quad \text{and} \quad \infty \in \bar{\mathcal{R}} \quad \text{or} \\ & \phantom{\text{if}} \quad \tau < 0 \quad \text{and} \quad -\infty \in \bar{\mathcal{R}} \\ 1 & \text{if} \quad \tau > 0 \quad \text{and} \quad \infty \notin \bar{\mathcal{R}} \quad \text{or} \\ & \phantom{\text{if}} \quad \tau < 0 \quad \text{and} \quad -\infty \notin \bar{\mathcal{R}} \quad . \end{cases}$$

This result is a direct consequence of Theorem 1. In (a), $E[N] = \sum_{k=0}^{\infty} P[N > k]$ is used.

The proofs of Theorem 1 and Corollary 1 show that a non-degenerate limit distribution is only obtained if the limit of $d(\lambda)/\lambda$ exists and is finite. This is the case if, e.g., we choose $d(\lambda)$ as a multiple of the asymptotic standard variance of Z_t, because

$$\lim_{t\to\infty} \mathrm{Var}_\infty(Z_t) = \frac{\lambda}{2-\lambda}\left[\gamma_{T,0} + 2\sum_{v=1}^{\infty}(1-\lambda)^v \gamma_{T,v}\right].$$

Note that the theorem suggests selecting the head start for small λ proportional to the variance of the asymptotic distribution and not proportional to the standard deviation, as is usually done.

The limit distribution is equal to the corresponding probability of the repeated significance test. Roughly speaking, the run length of the EWMA chart converges to that of the repeated significance test. For this reason, it is recommended to choose $d(\lambda)$ proportional to λ if a head start is taken into account.

If, however, $d(\lambda)$ is chosen as a constant $\tau \neq 0$, then we have a degenerate limit distribution. If τ is positive, then, in the upper one-sided case, $\infty \in \bar{\mathcal{R}}$ and thus $P[N(\lambda, \mathcal{R}) > k] \to 0$ as $\lambda \to 0+$. This is a very desirable property. However, since the result holds in the out-of-control case too, the probability of no signal in the out-of-control state converges to zero also. This defeats the purpose of the chart. Note that a negative choice of τ does not make sense for an upper one-sided scheme because the scheme would start with a delay, which contradicts the idea of a head start. In the two-sided case, the same conclusions hold because the limit is zero for $\tau \neq 0$.

Corollary 1(a) gives a statement about the behavior of the run length if the EWMA recursion has no head start. If λ tends to zero, then the ARL of the EWMA chart for monitoring θ converges to the ARL of the repeated significance test.

A detailed illustration of the results obtained in this section is given in Section 4.

3.2. *The modified EWMA scheme based on the asymptotic variance*

Next we analyse the run length of the scheme based on the asymptotic variance in the in-control state. Let $\sqrt{\lambda}\cdot\mathcal{R} = \{\sqrt{\lambda}\, r : r \in \mathcal{R}\}$.

Theorem 2. *Suppose that the assumptions of Theorem 1 are satisfied. Furthermore let $\sum_{v=1}^{\infty} |\gamma_{T,v}| < \infty$. Then for any fixed $k \in I\mathbb{N}$,*

$$\lim_{\lambda\to 0+} P[N_{asymp}(\lambda, \sqrt{\lambda}\cdot\mathcal{R}) > k] = \begin{cases} 0 & \text{if} \quad d = \infty \quad \text{and} \quad \infty \in \bar{\mathcal{R}} \quad \text{or} \\ & \phantom{\text{if}} \quad d = -\infty \quad \text{and} \quad -\infty \in \bar{\mathcal{R}} \\ P[A^*] & \text{if} \quad d \in I\mathbb{R} \\ 1 & \text{if} \quad d = \infty \quad \text{and} \quad \infty \notin \bar{\mathcal{R}} \quad \text{or} \\ & \phantom{\text{if}} \quad d = -\infty \quad \text{and} \quad -\infty \notin \bar{\mathcal{R}} \end{cases}$$

with

$$A^* = \bigcap_{t=1}^{k} \left\{ (x_1, \ldots, x_k) : \frac{t(\bar{T}_t(x_1, \ldots, x_t) - \theta) + d}{\sqrt{\gamma_{T,0} + 2\sum_{v=1}^{\infty}\gamma_{T,v}}} \notin \mathcal{R} \right\}.$$

Corollary 2. *Suppose that the assumptions of Theorem 2 are satisfied.*

(a) *If $d = 0$, then*

$$\lim_{\lambda\to 0+} P[N_{asymp}(\lambda, \sqrt{\lambda}\cdot\mathcal{R}) > k] = P[N_{asymp}(\mathcal{R}) > k]$$

and

$$\lim_{\lambda\to 0+} E[N_{asymp}(\lambda, \sqrt{\lambda}\cdot\mathcal{R})] = E[N_{asymp}(\mathcal{R})].$$

(b) *If $d(\lambda)$ is constant but unequal to zero, say $d(\lambda) = \tau \in I\mathbb{R} - \{0\}$ for all λ, then*

$$\lim_{\lambda\to 0+} P[N_{asymp}(\lambda, \sqrt{\lambda}\cdot\mathcal{R}) > k] = \begin{cases} 0 & \text{if} \quad \tau > 0 \quad \text{and} \quad \infty \in \bar{\mathcal{R}} \quad \text{or} \\ & \phantom{\text{if}} \quad \tau < 0 \quad \text{and} \quad -\infty \in \bar{\mathcal{R}} \\ 1 & \text{if} \quad \tau > 0 \quad \text{and} \quad \infty \notin \bar{\mathcal{R}} \quad \text{or} \\ & \phantom{\text{if}} \quad \tau < 0 \quad \text{and} \quad -\infty \notin \bar{\mathcal{R}} \quad . \end{cases}$$

Corollary 3. *Suppose that the assumptions of Theorem 2 are satisfied. Let $d(\lambda)$ be a constant, i.e. $d(\lambda) = \tau \in \mathbb{R}$ for all λ.*

(a) *For* $\mathcal{R} = (-\infty, -c) \cup (c, \infty)$ *with* $c > 0$,

$$\lim_{\lambda \to 0+} P[N_{asymp}(\lambda, \mathcal{R}) > k] = \begin{cases} 1 & \text{if } \tau = 0 \\ 0 & \text{if } \tau \neq 0. \end{cases}$$

(b) *For* $\mathcal{R} = (c, \infty)$ *with* $c > 0$,

$$\lim_{\lambda \to 0+} P[N_{asymp}(\lambda, \mathcal{R}) > k] = \begin{cases} 0 & \text{if } \tau > 0 \\ 1 & \text{if } \tau \leq 0. \end{cases}$$

(c) *For* $\mathcal{R} = (-\infty, -c)$ *with* $c > 0$,

$$\lim_{\lambda \to 0+} P[N_{asymp}(\lambda, \mathcal{R}) > k] = \begin{cases} 0 & \text{if } \tau < 0 \\ 1 & \text{if } \tau \geq 0. \end{cases}$$

Note that Corollary 3 describes the situation for the usually applied EWMA chart with asymptotic variance. If there is no head start, i.e. $\tau = 0$, the probability of a signal up to time k converges to 1. This result holds in the in-control as well as the out-of-control case. While it is desirable in the out-of-control case, it is very unpleasant in the in-control case. A consequence of this result is also that $\lim_{\lambda \to 0+} E[N_{asymp}(\lambda, \sqrt{\lambda} \cdot \mathcal{R})] = \infty$ and therefore the behavior of the chart cannot be assessed by the average run length. Recall that the in-control ARL is used to determine the control limits, but in the present case this cannot be done since this quantity does not exist if λ tends to zero. Consequently the control statistics have to be adapted, to avoid degeneration when λ tends to zero. How this has to be done is shown in Corollary 2(a).

In sum, if λ takes values close to zero, the control chart based on the exact variance must be favored: it has at least a reasonable limit behavior—that of the repeated significance test. For a better understanding of this limit behavior, it is necessary, however, to analyse the properties of the repeated significance test.

4. Illustration of the limit behavior

In this section we illustrate the theoretical results derived above. We focus on the detection of a change in the mean, and choose $T_t = X_t$. Moreover, we restrict ourselves to the two-sided problem, i.e. we choose $\mathcal{R} = (-\infty, -c) \cup (c, \infty)$. Then the limit chart, i.e. the repeated significance test, has the run length

$$N(c) = \inf \left\{ t \in I\!N : \frac{\sum_{i=1}^{t} (X_i - \mu)}{\sqrt{t[\gamma_0 + 2 \sum_{i=1}^{t-1} (1 - i/t)\gamma_i]}} > c \right\}.$$

In order to illustrate how the limit scheme may behave, we focus on the case of independent random variables. A detailed analysis of the limit behavior of the one-sided scheme for independent normally distributed random variables is given in Morais et al. (forthcoming). In that case, we have that

$$N(c) = \inf \left\{ t \in I\!N : \frac{\sum_{i=1}^{t} (X_i - \mu)}{\sqrt{t\gamma_0}} > c \right\}.$$

We make use of the change point model

$$E(X_t) = \begin{cases} \mu & \text{for } t < q \\ \mu + a\sqrt{\gamma_0} & \text{for } t \geq q. \end{cases} \tag{6}$$

The control limit c was chosen as a solution of $E_\infty[N(\lambda, c)] = \xi$, with $\xi > 1$; write $c = c(\lambda, \xi)$. We set $\xi = 50$ and the solution was found numerically using 10^7 paths of the process. The observations were generated by independent sampling from a Gaussian distribution. The results are shown in Fig. 1. Both for the exact and asymptotic charts, the control limits decrease with decreasing λ. Note that the control limit for the asymptotic charts tends to zero and is always smaller than the control limit for the exact chart.

The results for the out-of-control ARLs for the asymptotic and the exact variance are shown in Figs. 2 and 3, respectively. We consider three types of shifts, $a = 0.5, 1, 2$, and four different head starts. In particular, we consider no head start with $d(\lambda) = 0$; a constant head start with $d(\lambda) = 0.5$; and the head start asymptotically proportional to the smoothing parameter with $d(\lambda) = \frac{\lambda}{2-\lambda}$ and $d(\lambda) = 2 \cdot \frac{\lambda}{2-\lambda}$. The out-of-control ARLs of the chart with the asymptotic variance in Fig. 2 is not monotonic: they attain their minima for $\lambda > 0.1$. This is consistent with the usual recommendations in the literature for the choice of the optimal smoothing parameter. The head start which is proportional to the asymptotic variance of Z_t leads to finite and reduced out-of-control ARLs. This reflects the idea and the purpose of head starts. If, however, the head

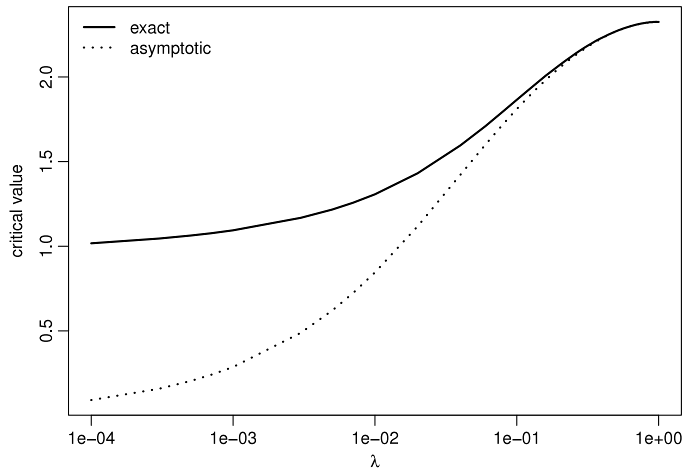

Fig. 1. The critical value c as a function of λ for an EWMA scheme with exact and asymptotic variance, for $\xi = 50$ and i.i.d. Gaussian observations. The results are based on a Monte Carlo study with 10^7 replications.

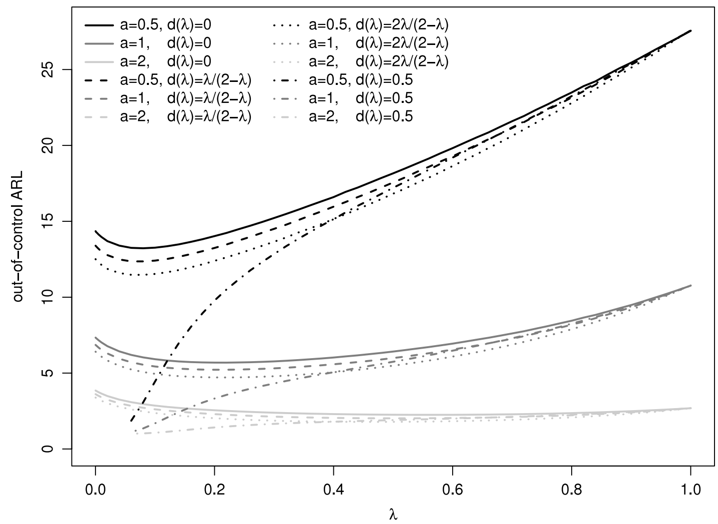

Fig. 2. The out-of-control ARL as a function of λ for the EWMA scheme with asymptotic variance and with different head starts $d(\lambda)$. The in-control ARL is set to $\xi = 50$. The results are based on a Monte Carlo study with 10^7 replications and i.i.d. Gaussian observations.

start is constant, it clearly leads to substantially lower out-of-control ARL for modest values of λ. However, the behavior of the ARL for small smoothing parameters becomes numerically unstable and cannot be determined by the simulation study. This illustrates the statement of Corollary 2(b).

The out-of-control ARL for the EWMA chart with exact variance in Fig. 3 attains its minimal value in the limit case as λ tends to zero. The limiting point corresponds to the repeated significance test with critical value equal to $c = 0.9768$ and the out-of-control ARLs for the considered shifts are given by 3.87, 2.04 and 1.18, while for $\lambda = 0.0001$, these equal 5.46, 2.62 and 1.35, respectively. Thus the ARL, as a function of λ, becomes extremely steep as the smoothing parameter approaches zero. The impact of the head start in this framework is similar to that in the situation with the asymptotic variance, i.e. the head start reduces the out-of-control ARL for modest smoothing parameters, while the latter cannot be computed for $\lambda < 0.1$. This is consistent with the theoretical results in Section 3.

5. Summary

This paper essentially provides a thorough study of the behavior of the run length of EWMA charts with exact and asymptotic control limits when λ tends to zero. We ought to stress that the results are quite general and apply to the control of any parameter of a stationary process. We proved that when the smoothing parameter λ tends to zero, the run length of EWMA charts based on the exact variance has the same behavior as the run length of a chart based on a repeated significance test. But the out-of-control run length of EWMA charts based on the asymptotic variance is infinite if the rejection area includes the origin.

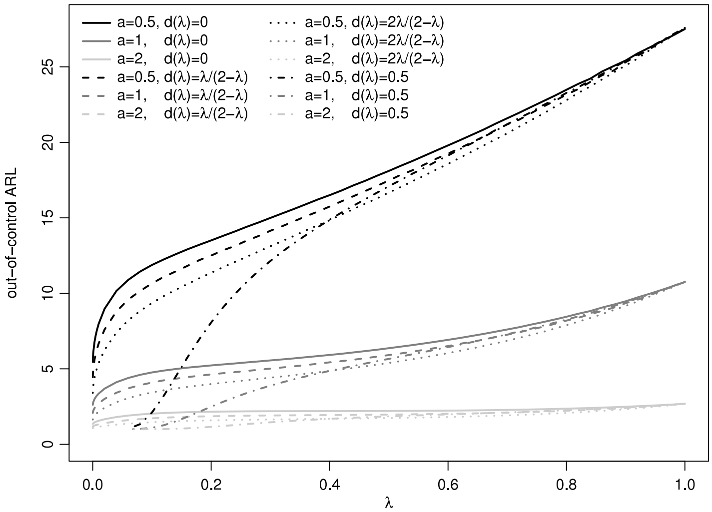

Fig. 3. The out-of-control ARL as a function of λ for the EWMA scheme with exact variance and with different head starts $d(\lambda)$. The in-control ARL is set to $\xi = 50$. The results are based on a Monte Carlo study with 10^7 replications and i.i.d. Gaussian observations.

Appendix

Proof of Theorem 1. The run length of the EWMA scheme with exact variance exceeds k with probability

$$\begin{aligned} P[N(\lambda, \mathcal{R}) > k] &= P\left(\frac{Z_t - \theta}{\sqrt{\mathrm{Var}_\infty(Z_t)}} \notin \mathcal{R}\ \forall\, t = 1, \ldots, k\right) \\ &= E[I_{A(\lambda)}(X_1, \ldots, X_k)] \end{aligned}$$

where $A(\lambda) = \cap_{t=1}^{k} A_t(\lambda)$ with

$$A_t(\lambda) = \left\{(x_1, \ldots, x_k) : \frac{Z_t - \theta}{\sqrt{\mathrm{Var}_\infty(Z_t)}} \notin \mathcal{R}\right\}$$

for $1 \le t \le k$. Here, $I_A(Z)$ denotes the indicator function, i.e. the function equal to 1 if $Z \in A$, but otherwise 0. Because

$$Z_t - \theta = \lambda \sum_{i=0}^{t-1} (1-\lambda)^i (T_{t-i} - \theta) + (1-\lambda)^t (Z_0 - \theta)$$

we get that

$$\frac{Z_t - \theta}{\sqrt{\mathrm{Var}_\infty(Z_t)}} = \frac{\sum_{i=0}^{t-1} (1-\lambda)^i (T_{t-i} - \theta)}{\sqrt{\mathrm{Var}_\infty(Z_t)}/\lambda} + \frac{d(\lambda)}{\lambda} \frac{(1-\lambda)^t}{\sqrt{\mathrm{Var}_\infty(Z_t)}/\lambda}. \tag{7}$$

Note that

$$\begin{aligned} \lim_{\lambda \to 0+} \frac{\mathrm{Var}_\infty(Z_t)}{\lambda^2} &= \frac{1}{2} \lim_{\lambda \to 0+} \frac{[1-(1-\lambda)^{2t}]\gamma_{T,0} + 2\sum_{v=1}^{t-1} (1-\lambda)^v [1-(1-\lambda)^{2(t-v)}]\gamma_{T,v}}{\lambda} \\ &= t^2 \mathrm{Var}_\infty(\bar{T}_t). \end{aligned}$$

Assume that $d(\lambda)/\lambda \to d \in I\mathbb{R}$. Then

$$\lim_{\lambda \to 0+} A_t(\lambda) = \left\{(x_1, \ldots, x_k) : \frac{\bar{T}_t(x_1, \ldots, x_t) - \theta}{\sqrt{\mathrm{Var}_\infty(\bar{T}_t)}} + \frac{d}{t\sqrt{\mathrm{Var}_\infty(\bar{T}_t)}} \notin \mathcal{R}\right\}$$

and

$$\lim_{\lambda \to 0+} \bigcap_{t=1}^{k} A_t(\lambda) = \bigcap_{t=1}^{k} \left\{(x_1, \ldots, x_k) : \frac{\bar{T}_t(x_1, \ldots, x_t) - \theta + d/t}{\sqrt{\mathrm{Var}_\infty(\bar{T}_t)}} \notin \mathcal{R}\right\} = A. \tag{8}$$

Consequently, it follows from the dominated convergence theorem that

$$\lim_{\lambda \to 0+} E[I_{A(\lambda)}(X_1, \ldots, X_k)] = E[I_A(X_1, \ldots, X_k)] = P[A],$$

thus proving the result.

The remaining parts of the theorem follow in the same way.

Proof. First note that

$$P\left[\frac{1}{\sqrt{\lambda}} \frac{Z_t - \theta}{\sqrt{\lim\limits_{t\to\infty} \text{Var}_\infty(Z_t)}} \notin \mathcal{R},\ \forall\, t = 1, \ldots, k\right] = E[I_{A_a(\lambda)}(X_1, \ldots, X_k)]$$

where $A_a(\lambda) = \cap_{t=1}^k A_{a,t}(\lambda)$ with

$$A_{a,t}(\lambda) = \left\{(x_1, \ldots, x_k) : \frac{1}{\sqrt{\lambda}} \frac{Z_t - \theta}{\sqrt{\lim\limits_{t\to\infty} \text{Var}_\infty(Z_t)}} \notin \mathcal{R}\right\}$$

for $1 \le t \le k$. Because

$$\frac{1}{\sqrt{\lambda}} \frac{Z_t - \theta}{\sqrt{\lim\limits_{t\to\infty} \text{Var}_\infty(Z_t)}} = \frac{\sum\limits_{i=0}^{t-1}(1-\lambda)^i (T_{t-i} - \theta)}{\sqrt{\lim\limits_{t\to\infty} \text{Var}_\infty(Z_t)/\lambda}} + \frac{d(\lambda)}{\lambda} \frac{(1-\lambda)^t}{\sqrt{\lim\limits_{t\to\infty} \text{Var}_\infty(Z_t)/\lambda}} \tag{9}$$

and

$$\begin{aligned}\lim_{\lambda\to 0+} \frac{\lim\limits_{t\to\infty} \text{Var}_\infty(Z_t)}{\lambda} &= \frac{1}{2} \lim_{\lambda \to 0+}\left[\gamma_{T,0} + 2\sum_{v=1}^{\infty}(1-\lambda)^v \gamma_{T,v}\right] \\ &= \gamma_{T,0} + 2\sum_{v=1}^{\infty} \gamma_{T,v},\end{aligned}$$

we get, in the case that $d(\lambda)/\lambda \to d \in I\mathbb{R}$,

$$\lim_{\lambda\to 0+} A_{a,t}(\lambda) = \left\{(x_1, \ldots, x_k) : \frac{t(\bar{T}_t - \theta) + d}{\sqrt{\gamma_{T,0} + 2\sum\limits_{v=1}^{\infty} \gamma_{T,v}}} \notin \mathcal{R}\right\},$$

$$\lim_{\lambda\to 0+} \bigcap_{t=1}^{k} A_{a,t}(\lambda) = A^*,$$

and

$$\lim_{\lambda\to 0+} E[I_{A_a(\lambda)}(X_1, \ldots, X_k)] = E[I_{A^*}(X_1, \ldots, X_k)] = P[A^*].$$

The other cases are derived by analogy.

Proof of Corollary 3. Following the argumentation in the proof of Theorem 2, we get that

$$P[N_{asymp}(\lambda, \mathcal{R}) > k] = E[I_{A_a^*(\lambda)}(X_1, \ldots, X_k)]$$

with

$$A_{a,t}^*(\lambda) = \left\{(x_1, \ldots, x_k) : \frac{Z_t - \theta}{\sqrt{\lambda}\sqrt{\gamma_{T,0} + 2\sum\limits_{v=1}^{\infty} \gamma_{T,v}}} \notin \mathcal{R}(\lambda)\right\},$$

$$A_a^*(\lambda) = \bigcap_{t=1}^{k} A_{a,t}^*,$$

$$\mathcal{R}(\lambda) = \left\{ \frac{r}{\sqrt{\lambda}} - \frac{\tau}{\lambda} \frac{(1-\lambda)^t}{\sqrt{\gamma_{T,0} + 2\sum_{v=1}^{\infty} \gamma_{T,v}}} : r \in \mathcal{R} \right\}.$$

Suppose that $\mathcal{R} = (-\infty, -c) \cup (c, \infty)$ with $c > 0$. Then $\lim_{\lambda \to 0+} \mathcal{R}(\lambda) = I\mathbb{R}$ for $\tau \neq 0$, while for $\tau = 0$, the limit set is the empty set $\emptyset$. The result follows using the same arguments as in Theorem 2.

The other parts are proved by analogy.

References

Alberts, W., Kallenberg, W.C.M., 2004. Are estimated control charts in control? Statistics 38, 67–79.
Brockwell, P.J., Davis, R.A., 1991. Time Series: Theory and Methods. Springer-Verlag, New York.
Brook, D., Evans, D.A., 1972. An approach to the probability distributions of CUSUM run length. Biometrika 3, 539–549.
Chan, L.K., Zhang, J., 2000. Some issues in the design of EWMA charts. Comm. Statist. Simulation Comput. 29, 207–217.
Crowder, S.V., 1987. A simple method for studying run-length distributions of exponentially weighted moving average charts. Technometrics 29, 401–407.
Crowder, S.V., Hamilton, M.D., 1992. An EWMA for monitoring a process standard deviation. J. Qual. Technol. 24, 12–21.
Jensen, W.A., Jones-Farmer, L.A., Champ, C.W., Woodall, W.H., 2006. Effects of parameter estimation on control chart properties: a literature review. J. Qual. Technol. 38, 349–364.
Knoth, S., 2005. Fast initial response features for EWMA Control Charts. Statist. Papers 46 (1), 47–64.
Lucas, J.M., Saccucci, M.S., 1990. Exponentially weighted moving average control schemes: properties and enhancements. Technometrics 32, 1–12.
Montgomery, D.C., 2009. Introduction to Statistical Quality Control, sixth ed. John Wiley & Sons, New York.
Morais, M.C., Okhrin, Y., Schmid, W., 2014. Quality surveillance with EWMA control charts based on exact control limits. Statist. Papers, http://dx.doi.org/10.1007/s00362-014-0612-8. forthcoming.
Page, E.S., 1954. Continuous inspection schemes. Biometrika 41, 239–250.
Roberts, S.W., 1959. Control charts tests based on geometric moving averages. Technometrics 1, 239–250.
Rosołowski, M., Schmid, W., 2003. EWMA charts for monitoring the mean and the autocovariances of stationary Gaussian processes. Sequential Anal. 22, 257–285.
Schipper, S., Schmid, W., 2001. Sequential methods for detecting changes in the variance of economic time series. Sequential Anal. 20, 235–262.
Schmid, W., 1997. On EWMA charts for time series. In: Lenz, H.J., Wilrich, P.-Th. (Eds.), Frontiers of Statistical Quality Control, Vol. 5. Physica-Verlag, Heidelberg, pp. 115–137.
Shewhart, W.A., 1931. Economic Control of Quality of Manufactured Product. Van Nostrand, Princeton.
Steiner, S.H., 1999. EWMA control charts with time-varying control limits and fast initial response. J. Qual. Technol. 31 (1), 75–86.