Exponentially Weighted Moving Average Control Charts with Time-Varying Control Limits and Fast Inital Response

Stefan Steiner, University of Waterloo

RR-97-03
June 1997

Exponentially Weighted Moving Average Control Charts with Time-Varying Control Limits and Fast Initial Response

Stefan H. Steiner

Dept. of Statistics and Actuarial Sciences
University of Waterloo
Waterloo, Ontario N2L 3G1
Canada

Abstract

Based on the classical formulation the control limits of an exponentially weighted moving average (EWMA) chart should vary with time, approaching asymptotic limits as time increases. However, previous analytic analyses of EWMA charts consider only asymptotic control limits. In this article, the run length properties of EWMAs with time-varying control limits are approximated using non-homogeneous Markov chains. Comparing the average run lengths of EWMA with time-varying control limits and results previously obtained for asymptotic EWMA charts shows that using time-varying control limits is akin to the fast initial response (FIR) feature suggested for Cumulative Sum (CUSUM) charts. The ARL of the EWMA scheme with time-varying limits is substantially more sensitive to early process shifts especially when the EWMA weight is small. An additional improvement in FIR performance can be achieved by further narrowing the control limits for the first 20 observations. The methodology is illustrated assuming a normal process with known standard deviation where we wish to detect shifts in the mean.

Keywords: Average run length; Cumulative Sum; CUSUM; Fast Initial Response (FIR); EWMA; Non homogenous Markov Chain.

1. Introduction

In the quality control context, exponentially weighted moving average (EWMA) control charts are used to monitor process quality. EWMA charts, and other sequential approaches like Cumulative Sum (CUSUM) charts, are an alternative to Shewhart control charts especially effective in detecting small persistent process shifts (Montgomery, 1991). First introduced by Roberts (1959), EWMA charts have a fairly long history, but only recently have its properties been evaluated analytically (Crowder 1987; Lucas and Saccucci 1990). The EWMA also is known to have optimal properties in some forecasting and control applications (Box, Jenkins, and MacGregor, 1974). In this article we focus on the quality monitoring applications.

For monitoring the process mean, the EWMA control chart consist of plotting:

$$z_t = \lambda \bar{x}_t + (1-\lambda) z_{t-1}, \qquad 0 < \lambda \leq 1, \qquad (1)$$

versus time t, where λ is a constant and the starting value z_0 is set equal to an estimate of the process mean, often given as $\bar{\bar{x}}$ calculated from previous data. In this definition $\bar{x}_t$ is the sample mean from time period t, z_t is the plotted test statistic and λ is the weight assigned to the current observation. The definition of the EWMA test statistic given in (1) can be easily adapted to monitor any process parameter of interest. For example, by replacing $\bar{x}_t$ in (1) by the sample standard deviation, and changing the starting value to our best guess at the in-control standard deviation, the EWMA will monitor the process dispersion.

By writing out the recursion in (1) the EWMA test statistic is shown to be an exponentially weighted average of all previous observations. In quality monitoring applications, typical values for the weight λ are between 0.05 and 0.25, although larger values may be used in forecasting and control applications. In the limiting case, with $\lambda = 1$, the EWMA chart is the same as a Shewhart $\bar{X}$ control chart. Using an EWMA chart, the process is considered out-of-control whenever the test statistic z_t falls outside the range of the control limits. EWMA control limits are discussed in detail in the next section.

As shown in Montogomery (1991), the control limits for EWMA charts should be time-varying since the variance of the test statistic z_t depends on *t*, because the effect of the starting constant z_0 decreases as *t* increases. However, all past analytic study of the properties of the EWMA chart have used fixed (asymptotic) control limits to make the analysis easier. This article presents methodology for determining the expected value and standard deviation of the run length of the EWMA chart with time-varying control limits. Numerical results are given for monitoring the mean of a normal distribution. Not surprisingly, the results show that EWMA charts with time-varying control limits has shorter average run lengths (ARLs) than EWMA charts with asymptotic control limits for start up quality problems. The effect for out-of-control mean values is more pronounced than for the in-control case, especially for large process shifts. As a result, EWMA control charts with time-varying control limits are appropriate in all situations where the initial quality level is suspect. This is useful because processes are fairly likely different from the target value when a control scheme is initiated due to start-up problems or because of ineffective control action after the previous out-of-control signal. In addition, often after a process change or adjustment we wish to quickly confirm that the change had the desired effect.

Using time-varying control limits has an effect similar to the fast initial response (FIR) feature recommended by Lucas and Crosier (1982) for CUSUM charts, since it helps detect problems with the start up quality. For CUSUMs, the FIR feature substantially decreases the ARL for an out-of-control process while decreasing the ARL of an in-control process only marginally. A one-sided tabular CUSUM control chart designed to monitor for upward process mean shifts consists of plotting: $Y_i = \max\left(0, Y_{i-1} + \bar{x}_i - (\mu_0 + k)\right)$, where $Y_0 = 0$, μ_0 is the in-control mean, and *k* is the reference value (Montgomery, 1991). The process is assumed to be in-control as long as $Y_i < h$, and is deemed to shifted if at any

i $Y_i \geq h$. The FIR feature sets the initial CUSUM value Y_0 at some non-zero value, typically $Y_0 = h/2$. A two-sided version involves monitoring two one-sided CUSUM charts one for positive shifts and the other for negative shifts.

For EWMA charts, Lucas and Saccucci (1990) suggested the simultaneous use of two one-sided EWMA charts with initial states different than zero as an implementation of the FIR feature. One EWMA chart monitors for increases in the process parameter, while the other chart monitors for decreases. Rhoads, Montgomery and Mastrangelo (1996) adapt the Lucas and Saccucci approach by allowing the one-sided EWMA to have time-varying control limits as given by (2) and discussed in Section 2. Rhoads et al. (1996) compare the run length properties determined through simulation. Both these implementations of FIR-EWMA charts require monitoring two EWMA charts to monitor a process for two-sided shifts.

This article shows that the use of time-varying control limits makes a EWMA chart more sensitive to start up quality problems than the traditional asymptotic limits. If additional protection to start up quality problems is desired the further narrowing of the control limits according to an exponential weighting scheme mimics the FIR feature. The derivation of time-varying control limits for an EWMA is presented in Section 2, and the effect of time-varying control limits is illustrated for a simple example. Section 3 uses numerical results to contrast and compare EWMA's with time-varying control limits and EWMA's with asymptotic limits. Section 4 introduces a FIR feature for two-sided EWMA charts and shows that this approach is superior to methods suggested previously by Lucas and Saccucci (1990) and Rhoads et al. (1996). In the Appendix, it is shown that the run length properties of an EWMA chart with time-varying control limits can be approximated using a non-homogenous Markov chain.

2. EWMA Control Charts with Time-Varying Control Limits

From (1) the mean value and variance of z_t are easily derived (Montgomery, 1991). Assuming the $\bar{x}_i$`s are independent random variables with mean μ_x and variance σ_x^2/n, where n is the sample size used at each time interval to calculate $\bar{x}_i$, we get

$$\mu_{z_t} = \mu_x, \qquad \text{and}$$
$$\sigma_{z_t}^2 = \frac{\sigma_x^2}{n}\left(\frac{\lambda}{2-\lambda}\right)\left[1-(1-\lambda)^{2t}\right]. \tag{2}$$

Notice that the variance of the EWMA test statistic z_t is a function of time. This should be expected since the number of observations used to derive the EWMA test statistic varies with time and the influence of the initial fixed value z_0 slowly decreases.

Control limits for an EWMA control chart are typically derived based on $\pm L$ sigma limits, where L is usually equal to three as in the design of Shewhart control chart limits. Thus, the time-varying upper and lower EWMA control limits, $UCL(t)$ and $LCL(t)$ respectively, are given by

$$UCL(t) \;=\; \mu_x + L\sigma_x\sqrt{\frac{\lambda\left[1-(1-\lambda)^{2t}\right]}{(2-\lambda)n}} \quad \text{and}$$
$$LCL(t) \;=\; \mu_x - L\sigma_x\sqrt{\frac{\lambda\left[1-(1-\lambda)^{2t}\right]}{(2-\lambda)n}}\;, \tag{3}$$

where, in applications, μ_x and σ_x are typically estimated from preliminary data as the sample mean and sample standard deviation. As t increases the control limits $UCL(t)$ and $LCL(t)$ converge to the asymptotic control limits, denoted as UCL and LCL, given by $\mu_x \pm L\sigma_x\sqrt{\lambda/(2-\lambda)n}$. The rate of convergence to this asymptotic values depends critically on λ with the convergence being much slower for small λ.

To illustrate the effect of time-varying limits consider the following example used by Lucas and Crosier (1982) to show the effect of the FIR feature on a CUSUM chart. The raw data points in time sequence are (.8, 1.9, 1.4, 2, 1.1, .7, 2.6, .5, 1.2) and represent an initial out-of-control situation. Figure 1 shows the resulting EWMA charts for different values of λ with control limits calculated based on initial estimates of $\mu_x = 0$ and $\sigma_x = 1$ and L set to three. The time-varying upper control limit $UCL(t)$ is shown as a solid line, whereas the asymptotic control limit UCL is shown as a dashed line. Figure 1 shows only the upper control limits to aid display; normally both upper and lower control limits are shown. The number of observations needed to generate an out-of-control signal depends on both the value of λ and whether the time-varying control limits are used. When λ equals .05, .1 or .25 an EWMA chart with time-varying control limits signals after only four observations whereas using the asymptotic limits a signal will not be generated until observation seven for $\lambda = .1$ and $\lambda = .25$, or observation nine for $\lambda = .05$. When $\lambda = .5$, the time-varying control limit quickly converges to the asymptotic value and thus has little effect. When $\lambda = .5$ a signal occurs after seven observations using either $UCL(t)$ or UCL as the control limit.

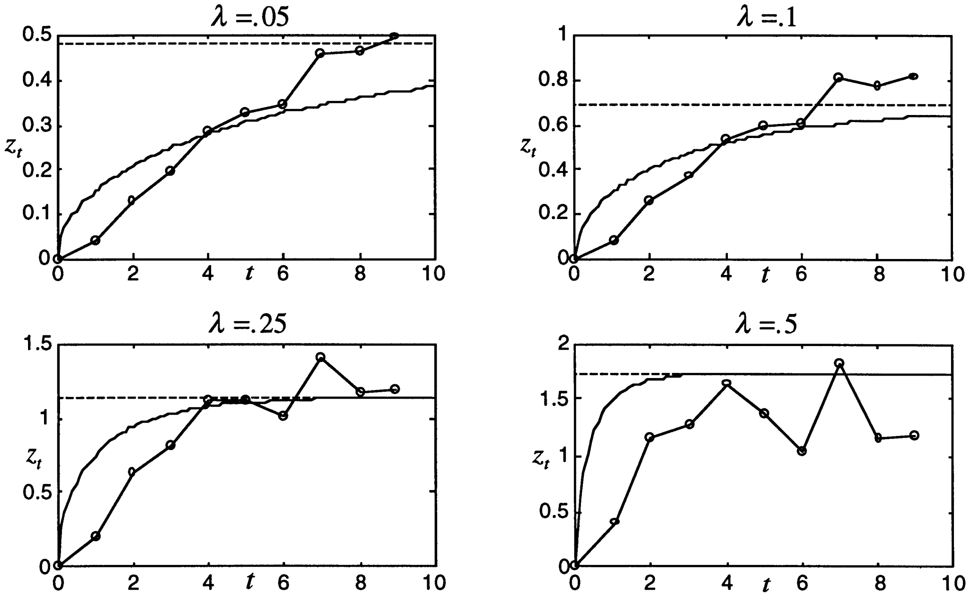

Figure 1: Plot of EWMA Control Charts with Time-varying Control Limits
dashed lines show the asymptotic control limits
solid lines show the time-varying control limits generated by (3)
circles represent the EWMA values

As can be seen in Figure 1, using asymptotic control limits rather than the time varying limits makes the EWMA chart much less sensitive to process shifts in the first few observations. This could be a significant problem if a large shift occurs early, or if after an out-of-control condition the process is not properly reset.

3. Run Length Properties of EWMA Charts with Time-varying Control Limits

In this section, the run length properties of EWMA charts with time-varying control limits, such as ARL, are compared with the run length properties of EWMA charts with asymptotic control limits. As will be shown, while the process is in-control, the ARLs of EWMA control charts with time-varying control limits are nearly identical to the ARLs of traditional EWMA charts with asymptotic control limits. However, when the initial process level is out-of-control the ARL of the two charts may differ substantially depending on the value of the EWMA weight λ.

It is important to quantify the effect of using time-varying control limits since EWMA control charts are usually designed to have given average run lengths (ARLs) under certain operating conditions. For an EWMA the design parameters include λ and L. However, since the time-varying control limits converge to the constant asymptotic values as time increases, for process shifts that occur later in time the two charts will have similar run length properties. As a result, EWMA control charts with time-varying control limits can be designed in the same manner as EWMA with asymptotic limits. See Crowder (1987) for guidelines.

The run length properties of EWMA control charts with asymptotic control limits were determined by Crowder (1987) using an integral equation approach. Unfortunately, this integral equation solution approach is not applicable for EWMA charts with time-varying control limits. However, the run length properties of the EWMA chart with time-varying control limits can be approximated using a non-homogeneous discrete Markov chain. Using a Markov chain the feasible state space is approximated through discretization and the probability of moving from any one state to any other state for each time period is determined. By using a greater number of distinct states the approximation of the run length properties can be made more precise. A detailed explanation of the solution procedure is given in the Appendix.

The effect of time-varying control limits on the ARL is illustrated in Figure 2. The results were derived using $L = 3.0$ as the control limit constant, and without loss of generality assuming an in-control mean and standard deviation of zero and unity respectively. In Figure 2, the horizontal axis gives the initial true process mean in $\sigma_{\overline{X}}$ units, the standard deviation of the sample mean. The results are given only for positive shifts, but since the problem is symmetric the same pattern is observed for negative shifts. These results are also tabulated in the Table A1 of the Appendix. ARL values for the asymptotic case are taken from Crowder (1987), while ARL results for EWMA with time-varying control limits are determined using the methodology presented in the Appendix. Figure 2 shows that the effect of using time-varying control limits on the ARL of the EWMA is substantial when the process is not initially in-control, especially when λ is small. The figure uses log(ARL) to improve the visual comparison.

As an example, from Table A1 assuming the initial process mean value is 2.0 $\sigma_{\overline{X}}$ units greater than the in-control value used to set up the EWMA chart, then for $\lambda = .05$ the ARL of the EWMA with time-varying control limits is 2.8 which is much shorter than the ARL of 6.0 required for an EWMA using asymptotic control limits. The effect of the time-varying control limits, however, has very little influence on the in-control run length as shown in Figure 2 and by the $\sigma_{\overline{X}} = 0.0$ row in Table A1. As such time-varying control limits are recommended for all EWMA charts, since their performance will be substantially better than asymptotic limit EWMAs when the process is fairly likely to start out-of-control.

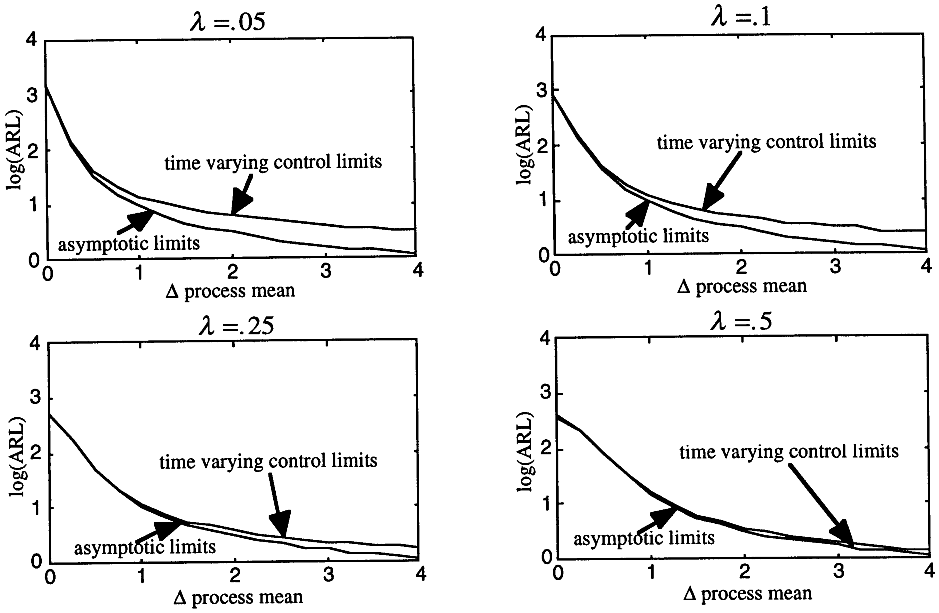

Figure 2: Plot of the ARLs for EWMA charts with time-varying and asymptotic control limits

Standard deviation values for the asymptotic EWMA control charts are also given in Crowder (1987). Table A2 in the Appendix reproduces the Crowder results and gives the standard deviation values for the time-varying case also calculated using the time non-homogenous Markov chain methodology presented in the Appendix. Table A2 shows that the standard deviation of the run lengths are nearly identical for the asymptotic EWMA and the EWMA with time-varying control limits.

It is also of interest to examine how the distribution of the run length of an EWMA chart changes when time-varying control limits are adopted. The run length distribution can be determined using equations (A3) given in the Appendix. Figures 3 and 4 show the run length distributions for EWMAs with time-varying control limits and asymptotic control limits when the initial process is in-control and shifted one $\sigma_{\bar{X}}$ unit respectively.

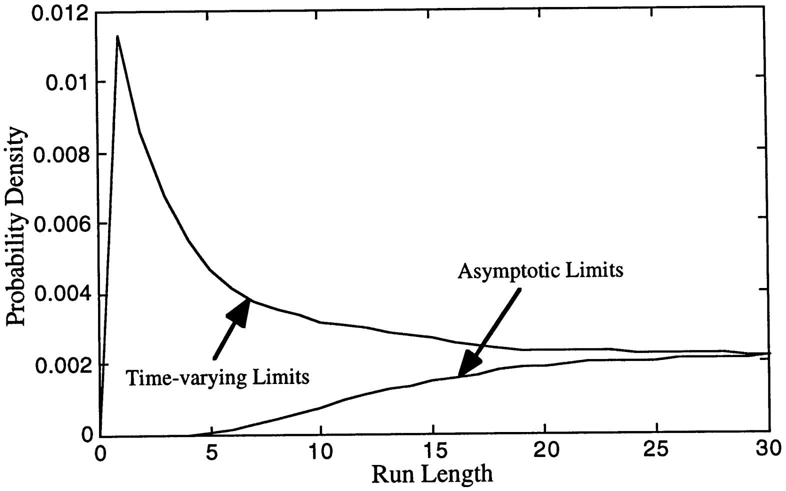

Figure 3: In-control Run Length Distribution of EWMA with Time-varying and Asymptotic Control Limits, $\lambda = .05, L = 2.587$

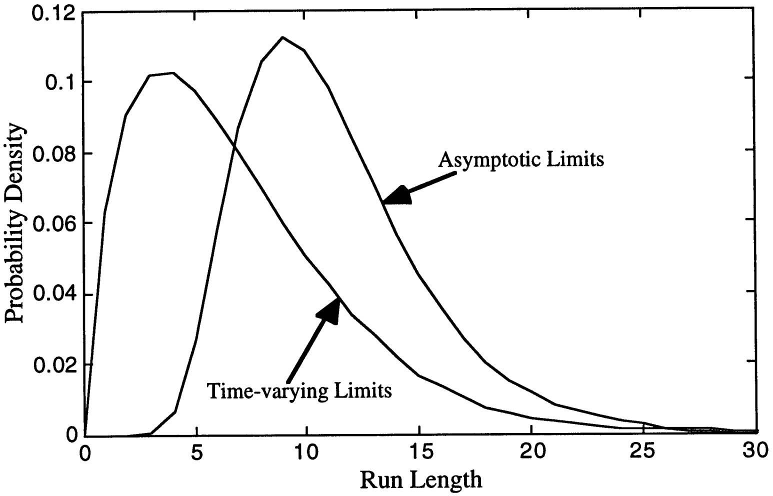

Figure 4: Out-of-control Run Length Distribution of EWMA with Time-varying and Asymptotic Control Limits initial mean shift equals one standard deviation unit

Figure 3 shows an initial spike in the run length probability density for the EWMA with time-varying control limits, with the two probability densities nearly converging for long run lengths. This greater probability of a short run length is undesirable since the initial process state is in-control and we would like the run length to be very long. However, since the probabilities involved are still very small, this spike has a corresponding small influence on the ARL. In Figure 4, by contrast, the bulk of the probability density for the two cases is quite different, and the ARL under the time-varying control limits will be substantially shorter. Of course, given an initial out-of-control state a short ARL is desirable.

Comparing the run length distribution plots shown in Figure 3 and 4 with similar plots for CUSUM and FIR CUSUM in Lucas and Crosier (1982) and for FIR-EWMA charts in Lucas and Saccucci (1990) suggests that the effect of the time-varying limits is similar to that achieved with the FIR feature. The effect of the time-varying limits appears less pronounced than the FIR-CUSUM which suggests that an additional narrowing of the time-varying control limits for small values of t may be appropriate to make the EWMA chart even more sensitive to start-up quality problems.

4. EWMA Control Charts with Fast Initial Response (FIR)

EWMA charts with time-varying control limits were shown in the previous section to have properties similar to the FIR feature when compared with asymptotic EWMA. However, using time-varying control limits is not the same as the FIR feature for CUSUMs since the adjustment of the control limits only corrects the control limits to take into account the time dependent nature of the EWMA statistic given by (1). In effect, the time-varying control limits do not make detecting start up quality problems any more likely than detecting later quality problems.

A few authors have suggested adaptations to the EWMA scheme to build in a true FIR feature. As discussed in the introduction, to create a two-sided EWMA chart that reacts quickly Lucas and Saccucci (1990) suggested the use of two one-sided EWMA charts with initial states different than zero. Rhoads, Montgomery and Mastrangelo (1996) adapt the Lucas and Saccucci approach by allowing each one-sided chart to have time-varying control limits. Both these methods have the desired effect of making the chart more sensitive to start up quality problems, but are rather awkward since they require the simultaneous use of two EWMA charts to accomplish the task previously achieved with just one chart.

Here a different approach is suggested that retains the simplicity of a single control chart. To give EWMA charts with time-varying control limits a FIR feature, the control limits are narrowed further for the first few sample points. This approach is easily implemented since the control limits are already time-varying. Since the time-varying control limits exponentially approach the asymptotic limits it is reasonable to use an exponentially decreasing adjustment to further narrow the limits. Let $FIR_{adj} = 1-.5^{1+a(t-1)}$. With this setup the FIR adjustment makes the control limits for the first sample point ($t = 1$) half the original distance from the starting value. Half the distance was chosen to mimic the 50% head start typically suggested for FIR CUSUM charts. The effect of the FIR adjustment decreases with time to ensure that the long term run length properties of the EWMA will be virtually unchanged. A reasonable setup would be to set the adjustment parameter a so that the FIR adjustment has very little effect after observation 20. This should be sufficient to allow the detection of quality problems in the startup. Using $a = 0.3$ means that the adjust for the control limits at observation 20 is .99, and the difference between the adjusted limit and the time-varying limit is negligible. Using this adjustment factor and (3), the FIR-EWMA control limits are:

$$\mu_x \pm L\sigma_x\left(1-(1/2)^{0.7+0.3t}\right)\sqrt{\frac{\lambda\left[1-(1-\lambda)^{2t}\right]}{(2-\lambda)n}} \qquad (4)$$

The control limits given by (4) are time-varying, thus the run lengths properties of the proposed FIR-EWMA can also be determined using the non-homogeneous Markov chain methodology presented in the Appendix.

Figure 5 shows the effect of using limits (4) on the simple example initially discussed in Section 2, and previously illustrated in Figure 1. In Figure 5, the advantage of the additional narrowing of the control limits in detecting start up quality problems is clearly demonstrated. For all the different values of λ the FIR-EWMA signals in just two observations. This is a substantial improvement over the run lengths obtained with only the time-varying control limits, especially for large values of λ.

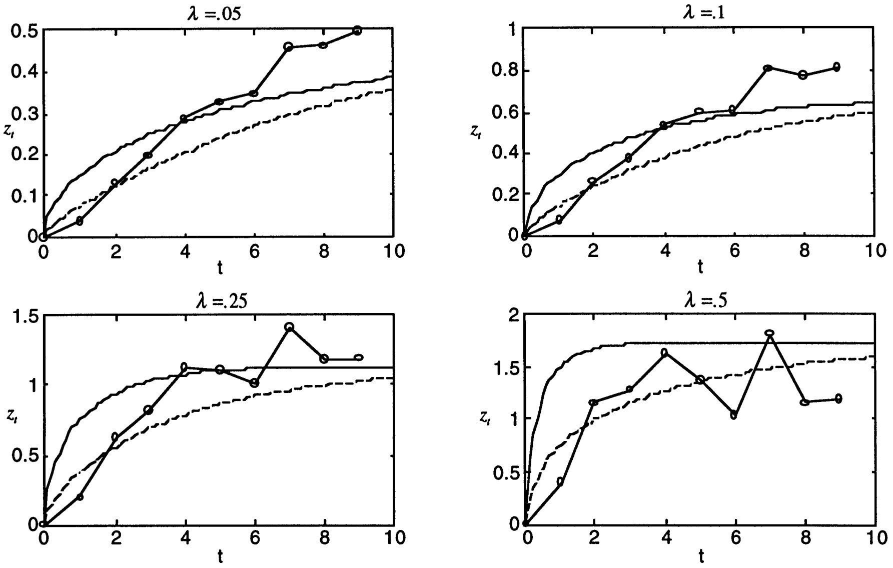

Figure 5: EWMA Control Charts with Time-varying Control Limits
dashed lines show the FIR time-varying control limits from (4)
solid lines show the time-varying control limits generated by (3)
circles represent the EWMA values

Table 1 compares the ARL of the Lucas and Saccucci (1990) FIR-EWMA, denoted L-FIR, the Rhoads et al. (1996) FIR-EWMA, denoted R-FIR, and a FIR-EWMA with adjusted time varying control limits given by (4). The results for the L-FIR and the R-FIR are taken from simulation results published in Rhoads et al. (1996), and the run length results for the proposed FIR-EWMA were approximated using the methodology described in the Appendix. For all the FIR-EWMAs the control limit multiple L has been adjusted so that, in-control, all methods have approximately the same ARL.

Table 1: Average Run Length Comparison of EWMAs with FIR

	λ=0.25			λ=0.1		
$\mu_x/\sigma_{\bar{x}}$	L-FIR	R-FIR	FIR	L-FIR	R-FIR	FIR
	L=2.81	L= 3.0	L=3.07	L=2.81	L= 3.0	L=2.91
0.0	483	452	468	463	466	459
0.5	42.1	39.3	33.5	24.2	22.2	19.6
1.0	8.5	7.6	5.2	6.9	5.4	4.5
1.5	3.9	3.2	2.3	3.7	2.4	2.1
2.0	2.5	1.9	1.5	2.7	1.6	1.4
3.0	1.5	1.1	1.1	1.8	1.1	1.1
4.0	1.1	1.0	1.0	1.3	1.0	1.0

	λ=0.05			λ=0.03		
$\mu_x/\sigma_{\bar{x}}$	L-FIR	R-FIR	FIR	L-FIR	R-FIR	FIR
	L=2.62	L=2.72	L=2.69	L=2.44	L=2.54	L=2.55
0.0	421	417	419	383	384	391
0.5	19.7	17.0	16.5	18.6	14.9	13.8
1.0	7.0	4.4	4.2	7.4	3.9	3.6
1.5	4.1	2.2	2.0	4.6	2.0	1.8
2.0	3.1	1.5	1.4	3.4	1.4	1.3
3.0	2.1	1.1	1.1	2.4	1.1	1.0
4.0	1.7	1.0	1.0	1.9	1.0	1.0

The results in Table 1 suggest that proposed FIR-EWMA is superior to the previous approaches. For example, with λ=.1 and a mean shift of one standard deviation unit, the proposed FIR-EWMA requires on average only 4.5 observations to signal, while the Lucas and Saccucci FIR-EWMA, and the Rhoads et al. FIR-EWMA require 6.9 and 5.4 observations respectively. The reduction in out-of-control ARLs appears to be greatest when λ is not small. In addition to the benefit of better run length properties, the EWMAs with time-varying control limits also provide two-sided protection from start up quality problem through only a single control chart. This is a major advantage from an implementation perspective.

Summary

This article derives the run length properties for EWMA control charts with time-varying control limits. Since the variance of the EWMA test statistic is a function of time, time-varying control limits result in improved process shift detection capabilities if the process is initially out-of-control, or if it goes out-of-control quickly. The magnitude of the benefit of using time-varying control limits over traditional asymptotic limits depends on the EWMA constant λ, and size of the initial process shift. Results are presented that quantify the difference for an EWMA designed to monitor the process mean. In general, time varying control limits are useful if λ is small, say less than .3.

In situations where at the start of process monitoring there is good chance the process is out-of-control further narrowing of the time-varying control limits is shown to provide an additional fast initial response benefit. Adjusting the control limits to start at half the regular value and then exponentially approach the regular time-varying limits for 20 observations is shown to perform better than previously suggested approaches to create a FIR-EWMA. The proposed approach has the additional benefit of retaining the benefit of the EWMA chart that allows the two-sided detection of problems with a single chart. In this respect the proposed FIR-EWMA chart is also preferable to a FIR CUSUM.

Appendix

In this appendix, approximations for the distribution, expected value and variance of the run length of EWMA charts with time-varying control limits are derived. The solution procedure utilizes a non-homogenous Markov chain with g distinct states. In the solution the state space between the control limits is divided into g-1 distinct discrete states, and the out-of-control condition corresponds to the gth state. The different states are defined as $\mathbf{s} = \left(s_1, s_2, \ldots, s_{g-1}\right) = (LCL + w,\ LCL + 2w,\ \ldots,\ UCL - 2w,\ UCL - w)$, where $w = (UCL - LCL)/g$ and UCL and LCL are the asymptotic control limits as given by setting $t = \infty$ in (3). As g increases the approximation improves.

Assume that the transition probability matrix for time period t is given by

$$P_t = \begin{bmatrix} {}_tp_{11}, {}_tp_{12}, \ldots, {}_tp_{1g} \\ {}_tp_{21}, \quad \cdots \quad , {}_tp_{2g} \\ \vdots \qquad\qquad\quad \vdots \\ {}_tp_{g1}, \quad \cdots \quad , {}_tp_{gg} \end{bmatrix} = \begin{bmatrix} R_t & ,(I - R_t)\mathbf{1} \\ 0, \ldots, 0, & 1 \end{bmatrix} \tag{A1}$$

where I is the g by g identity matrix, **1** is a g by 1 column vector of ones, and ${}_tp_{ij}$ equals the transition probability from state s_i to state s_j for time period t. The last row and column correspond to the absorbing state that represents an out-of-control signal. The R_t matrix equals the transition probability matrix with the row and column that correspond to the absorbing (out-of-control) state deleted. R_t will be used to derive the run length properties of the EWMA control chart with time-varying control limits.

Since the time-varying control limits (3) asymptotically approach constant values, the state transition probabilities ${}_tp_{ij}$ converge to probabilities ${}_\infty p_{ij}$ and the matrix R_t converges to the infinite time transition matrix R_∞ as $t \to \infty$. The values for ${}_\infty p_{ij}$ can be determined by making some process assumptions. Assuming a normal model with X_i ~ $N(\mu_x, \sigma_x^2)$ and given the current EWMA value, the distribution of the future EWMA value z_{t+1} is $N(\lambda\mu_x + (1-\lambda)z_t, \lambda^2\sigma_x^2)$. Thus, the infinite time transition probabilities are:

$$\begin{aligned} {}_\infty p_{ij} &= \Pr\left(s_j - \frac{w}{2} < z < s_j + \frac{w}{2}\right), \text{ for } \qquad j = 1,2,\ldots,g-1 \text{ and} \\ {}_\infty p_{ig} &= \Pr\left(z > s_{g-1} + \frac{w}{2}\right) + \Pr\left(z < s_1 - \frac{w}{2}\right), \end{aligned} \qquad \text{(A2)}$$

where $z \sim N(\lambda\mu + (1-\lambda)s_i, \lambda^2\sigma^2)$. These values can be easily calculated to determine P_∞ and R_∞.

The time dependent transition matrices R_t can be determined from R_∞ by changing the transitions probabilities that lead to an earlier signal. Transitions probabilities in R_∞ from starting values (rows) that are outside the time-varying control limits and to ending values (columns) that result in out-of-control signals are set to zero. For each value of t, the appropriate rows and columns are identified by comparing the time-varying control limits with the states in the state space. In other words, to determine R_t the first $f_1(t)$ and last $f_2(t)$ rows and columns of R_∞ are set to zero vectors, where $f_1(t)$ equals the largest integer for which $s_{f_1} - w/2 \le LCL(t)$ and $f_2(t)$ is the smallest integer for which

$s_{f_2} + w/2 \geq UCL(t)$. In an attempt to consistently yield run length values less than the true value any state whose transition probability is at all effected by the changing control limit is set to zero. A state s_i is effected if the time-varying control limit is either closer to zero than s_i or within $w/2$ of s_i. Using this procedure estimates for R_1, R_2, etc. are obtained.

Determining the expected run length and the variance of the run length can now proceed using the matrices R_t. Letting *RL* equal the run length of the EWMA we have

$$\Pr(RL \leq t) \quad = \quad \left(I - \prod_{i=1}^{t} R_i\right)\mathbf{1}, \text{ and}$$

$$\Pr(RL = t) \quad = \quad \left(\prod_{i=1}^{t-1} R_i - \prod_{i=1}^{t} R_i\right)\mathbf{1} \quad \text{for } t \geq 1. \qquad \text{(A3)}$$

Thus,

$$E(RL) \quad = \quad \sum_{t=1}^{\infty} t\Pr(RL = t) \quad = \quad \sum_{t=1}^{\infty}\left(\prod_{s=1}^{t} R_s \mathbf{1}\right). \qquad \text{(A4)}$$

Similarly, the variance of the run length is

$$Var(RL) \quad = \quad I + \sum_{t=1}^{\infty}\left[(2t+1)\left(\prod_{s=1}^{t} R_s \mathbf{1}\right)\right]. \qquad \text{(A5)}$$

These expressions are gx1 vectors that give the average run length and variance from any starting value or state s_i. The values that correspond to the starting with $z_0 = \overline{\overline{X}}$ are easily found. Assuming that the control limits are symmetric about $\overline{\overline{X}}$ the corresponding state is $s_{g/2}$.

(A4) and (A5) give the moments of the run length in terms of infinite sums that converge for large *t*. These expression can be simplified in this case, since the control limits converge asymptotically, and thus the transition probability matrices R_t also converge to R_∞ as *t* increases. Replacing all R_t matrices for large *t* values with R_∞, the infinite sums (A4) and (A5) can be written as:

$$E(RL) \quad = \quad \sum_{t=1}^{t_{\max}-1}\left(\prod_{s=1}^{t} R_s \mathbf{1}\right) + \left(\prod_{s=1}^{t} R_s\right)(I - R_\infty)^{-1}\mathbf{1}, \quad \text{and} \qquad \text{(A6)}$$

$$Var(RL) \quad = \quad \mathbf{1} + \sum_{t=1}^{\mathrm{t_{max}}-1}\left[(2t+1)\left(\prod_{s=1}^{t} R_s\,\mathbf{1}\right)\right] + (2\mathrm{t_{max}}+1)\left(\prod_{s=1}^{\mathrm{t_{max}}} R_s\right)(1-R_\infty)^{-1}\,\mathbf{1}$$
$$+2\left(\prod_{s=1}^{\mathrm{t_{max}}} R_s\right)R_\infty(1-R_\infty)^{-2}\,\mathbf{1}, \tag{A7}$$

where $t_{\max}$ equals the number of time period for which different transition probability matrices are used. For the computations, $t_{\max}$ was chosen based on λ and g so that the matrix $R_{t_{\max}}$ is indistinguishable from R_∞. In this way increasing $t_{\max}$ further will have no influence on the solution accuracy. The minimum value for $t_{\max}$ is derived by realizing that if the time-varying control limits at time $t_{\max}$ differ from the asymptotic limits by less than $w/2$ then the matrix $R_{t_{\max}}$ is the same as R_∞. Solving $UCL - UCL(t) \le w/2$ and $LCL - LCL(t) \le w/2$ for the minimum t value yields $t_{\max}$ as the smallest integer larger than

$$\log\left(\frac{12nw(2-\lambda)\sigma\sqrt{\lambda/n(2-\lambda)} - w}{36\lambda\sigma^2}\right)\Bigg/ 2\log(1-\lambda).$$

For computational efficiency and accuracy, $E(RL)$ and $Var(RL)$ are determined using Gaussian elimination rather than by finding the matrix inverse directly as suggested by (A6) and (A7).

In general, as g increases the $E(RL)$ and $Var(RL)$ values obtained through (A6) and (A7) increase and more closely approximate the true values. The values increase because the procedure always underestimates the true run length. The run lengths are underestimated for two reasons; first, the absorbing boundaries for R_∞ are narrower than the control limits since they are set at $LCL + w/2$ and $UCL - w/2$, and second for R_t the absorbing probabilities are conservatively calculated since all states even marginally effected by the control limit are assumed to lead to absorption.

The advantage of consistently underestimating the run lengths of the EWMA are that we can use the rate of increase to estimate the true values. The values shown in the Tables A1, A2, and A3 were derived by estimating the true value $E(RL)_{g=\infty}$ based on fitting the model $E(RL) = E(RL)_{g=\infty} + B/g + C/g^2$ derived using the results generated with g = 50, 100, and 150. Verification of this approach using simulation suggests that our results differ from the true value by less than 1% except for very large process shifts when the average run length is near unity. For very large shifts, the values in the transition probability matrix R_t become smaller and calculations required to derive $E(RL)$ become more prone to rounding error. As a result, for large shifts the $E(RL)$ estimate may not increase as g increases. If this occurs, we use the largest obtained $E(RL)$ as an estimate of the true $E(RL)_{g=\infty}$, and the estimate may be off by as much as 10%. A similar problem is also reported in Lucas and Crosier (1982). However, in our case, for comparison purposes, the results are adequate.

Tables A1 and A2 give the detailed results required to generate Figures 1 and 2 in the text. The initial shift in the process mean is given $\sigma_{\bar{X}}$ units. Note that the value for $\mu_x/\sigma_{\bar{x}}$ = 0.0 and λ = .05 is incorrectly given as 1623.50 in Crowder (1987), the correct value is given in Table A2.

Table A1: Average Run Length for Two-sided EWMA Charts
Zero State Results, L = 3.0

	Asymptotic Control Limits				Time-varying Control Limits			
$\mu_x/\sigma_{\bar{x}}$	λ=.50	λ=.25	λ=.10	λ=.05	λ=.50	λ=.25	λ=.10	λ=.05
.00	398	503	842	1379	382	500	828	1353
.25	209	171	145	135	207	170	140	127
.50	75.4	48.5	37.4	37.4	74.5	47.6	34.5	32.5
.75	31.5	20.2	17.9	20.0	30.8	19.5	15.3	15.6
1.00	15.7	11.2	11.4	13.5	15.2	10.2	9.1	9.0
1.50	6.1	5.5	6.6	8.3	5.7	4.7	4.5	4.5
2.00	3.5	3.6	4.7	6.0	3.2	2.9	2.8	2.8
2.50	2.4	2.8	3.7	4.8	2.2	2.1	2.0	2.0
3.00	1.9	2.3	3.1	4.0	1.6	1.6	1.6	1.6
3.50	1.5	2.0	2.6	3.4	1.3	1.3	1.3	1.3
4.00	1.3	1.7	2.3	3.0	1.2	1.2	1.2	1.1

Table A2: Standard Deviation of the Run Length for Two-sided EWMA Charts
Zero State Results, $L = 3.0$

	Asymptotic Control Limits				Time-varying Control Limits			
$\mu_x/\sigma_{\bar{x}}$	λ=.50	λ=.25	λ=.10	λ=.05	λ=.50	λ=.25	λ=.10	λ=.05
.00	396	499	833	1363	396	499	834	1364
.25	207	167	133	113	207	167	133	113
.50	73.2	43.8	27.6	22.0	73.2	43.8	28.0	23.0
.75	29.3	15.9	10.2	8.8	29.2	16.0	10.6	9.7
1.00	13.6	7.5	5.3	4.9	13.6	7.4	5.7	5.5
1.50	4.3	2.7	2.3	2.3	4.2	2.8	2.5	2.6
2.00	1.9	1.4	1.3	1.4	1.9	1.5	1.5	1.5
2.50	1.1	0.9	0.9	1.0	1.1	1.0	1.0	1.0
3.00	0.8	0.6	0.7	0.8	0.8	0.7	0.7	0.7
3.50	0.6	0.5	0.6	0.6	0.5	0.5	0.5	0.5
4.00	0.5	0.5	0.5	0.5	0.4	0.4	0.4	0.4

To provide more details, Table A3 gives the ARL values for EWMAs with time-varying control limits for some different values of L.

Table A3: Average Run Length for Time-varying Control Limits EWMA Charts
Zero State Results

	$L = 2.75$				$L = 2.5$			
$\mu_x/\sigma_{\bar{x}}$	λ=.50	λ=.25	λ=.10	λ=.05	λ=.50	λ=.25	λ=.10	λ=.05
.00	184	240	410	664	90.5	122	213	343
.25	107	96.9	92.3	88.7	57.8	58.2	62.3	63.7
.50	43.6	32.0	27.1	25.6	26.8	22.4	21.1	20.5
.75	20.1	14.7	13.2	12.6	13.7	11.3	10.8	10.4
1.00	10.9	8.5	8.1	7.7	8.0	6.9	6.8	6.4
1.50	4.7	4.3	4.2	3.8	3.9	3.6	3.6	3.2
2.00	2.8	2.7	2.6	2.3	2.4	2.4	2.3	2.0
3.00	1.5	1.5	1.5	1.3	1.4	1.4	1.4	1.2
4.00	1.1	1.1	1.1	1.0	1.1	1.1	1.1	1.0

Results are derived for the two-sided case but the methodology can be easily adapted for one-sided case EWMA charts defined as $z_t = \max\left(\lambda \bar{x}_t + (1-\lambda) z_{t-1}, z_0\right)$. In addition, the examples provided assume the distribution of the observed process parameter is normal. However, similar results are easily derived for other underlying distributions.

Acknowledgments

This research was supported, in part, by the Natural Sciences and Engineering Research Council of Canada and the Manufacturing Research Council of Ontario.

References

Box, G.E.P., Jenkins, G.M., and MacGregor, J.F. (1974), “Some Recent Advances in Forecasting and Control,” *Applied Statistics*, 23, 158-179.

Brook, D., Evans, D.A. (1972), “An approach to the probability distribution of cusum run length,” *Biometrika*, 59, 539-549.

Crowder, S.V. (1987), “Run-Length Distributions of EWMA Charts,” *Technometrics*, 29, 401-407.

Lucas, J.M. and Crosier, R.B. (1982), "Fast Initial Response for CUSUM Quality Control Schemes," *Technometrics*, 24, 199-205.

Lucas, J.M. and Saccucci, M.S. (1990), “Exponentially Weighted Moving Average Control Schemes: Properties and Enhancements,” with discussion, *Technometrics*, 32, 1-29.

Montgomery, D.C. (1991), *Introduction to Statistical Quality Control*, Second Edition, John Wiley and Sons, New York.

Page, E. S. (1954), "Continuous Inspection Schemes," *Biometrika,* Vol. 41, 100-114.

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

Rhoads, T.R., Montgomery, D.C. and Mastrangelo, C.M. (1996-97) “Fast Initial Response Scheme for the Exponentially Weighted Moving Average Control Chart,” *Quality Engineering*, 9, 317-327.