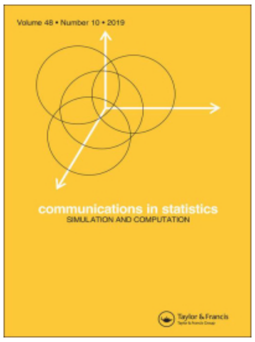
Volume 48 • Number 10 • 2019
communications in statistics
SIMULATION AND COMPUTATION
Taylor & Francis

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/lssp20

Gradient analysis of Markov-type control schemes and its applications

Emmanuel Yashchin

COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION®
2022, VOL. 51, NO. 4, 1814–1836
https://doi.org/10.1080/03610918.2019.1687718

Taylor & Francis
Taylor & Francis Group

Gradient analysis of Markov-type control schemes and its applications

Emmanuel Yashchin

IBM, Thomas J. Watson Research Center, Yorktown Heights, New York, USA

ABSTRACT
This paper presents formulas for gradients of the Average Run Length (ARL) and other Run Length characteristics by control scheme parameters, in the context of Markov Chain approach to analysis of Markov-type control schemes. We illustrate use of these formulas in several problems related to control charting, including (a) design of control schemes, (b) inference on Run Length characteristics based on Phase-1 data and (c) performance analysis for highly complex input data distributions.

ARTICLE HISTORY
Received 31 December 2018
Accepted 24 October 2019

KEYWORDS
Average run length; Detection; Monitoring; Statistical process control

1. Introduction

Markov-type control schemes have become increasingly popular in advanced quality control applications, mostly due to their statistical power and relative ease of use. Performance of this type of schemes has been proven to be statistically superior to their classical counterparts—Shewhart schemes ($\bar{x}$ - charts, p - charts, etc.) in the sense that with the same degree of protection against false alarms, they have a much better sensitivity with respect to out-of-control situations. These schemes are easily "analyzable"; in other words, it is possible to examine, by analytic means, the Run Length behavior of a scheme for any given stochastic pattern of incoming independent and identically distributed (iid) observations. They are relatively easily "designable": procedures for selecting control scheme parameters are available for many schemes of this type. One representative is the class of Cusum–Shewhart schemes, for which the literature covers an extensive range of methods and applications (Lucas 1982; Woodall 1983, 1986; Yashchin 1985a, 1985b; Hawkins and Olwell 1998; Capizzi 2015; Saleh et al. 2015; Knoth 2018). Approximate results for some non-iid cases are also available (Johnson and Bagshaw 1974; Bagshaw and Johnson 1975; Yashchin 1993). Another class is the Exponentially Weighted Moving Average (EWMA) schemes, available in several versions (Lucas and Saccucci 1990; Amin et al. 1999). The class of schemes introduced by Girshick and Rubin (1952), also known as the Roberts–Shiryaev class, has also received considerable attention in the literature, e.g., see Polunchenko and Raghavan (2018).

In recent years, computational issues related to application of control schemes necessitate research effort directed towards design and analysis, as monitoring systems are

CONTACT Emmanuel Yashchin yashchi@us.ibm.com IBM, Thomas J. Watson Research Center, Box 218, Yorktown Heights, Yorktown, NY 10598, USA.

becoming more complex and are expected to handle large and intense volumes of data (Dietrich, Plachy, and Norton 2014; Negandhi et al. 2015). In some applications, control schemes are used as part of *search engines* (Yashchin 2018), leading to necessity to maintain a large number of control schemes “on the fly”. Gradient analysis is most valuable for this type of applications, enabling the system administrators to maintain the desired balance between false alarm intensity, sensitivity requirements, problem prioritization and computing resources. In IBM, early large-scale monitoring applications became feasible due to efficient scheme design capability driven by gradient analysis, see Yashchin (1986). Such applications covered substantial parts of the IBM business, including microelectronics, storage systems, computer manufacturing, supply chain and finance, e.g., see Philips, Yashchin, and Stein (2003).

In this paper, we discuss an approach to gradient analysis that, we believe, has some potential, especially in light of its connection to the method of control scheme analysis that involves discretization of the scheme values. Essentially, the analysis is handled within the framework of Markov Chains—and this enables one to obtain some results that apply to all Markov-type schemes. A number of results are available in the literature (Lele 1996; Fu and Hu 1999; Fu, Lele, and Vossen 2009; Shu, Huang, and Jiang 2014; Huang, Shu, and Jiang 2016, 2018). These methods are most useful in conjunction with control scheme analysis based on solving integral equations, and they can give very accurate results—however, dealing with integral equations typically involves a non-trivial computational burden. Our procedures provide useful estimates of gradients at a much lower computational cost. Furthermore, as the procedures are formulated in terms of the matrix theory, the computer codes for analysis and gradient computations of various Markov-type schemes exhibit a high degree of similarity, facilitating their implementation.

In Sec. 2 we provide basic information about Markov-type control schemes and their analysis. In Sec. 3 we discuss the gradient by the signal level (threshold) of schemes. In Sec. 4 we consider gradient analysis by other scheme parameters. In Sec. 5 we explore performance issues. Finally, in Sec. 6 we discuss applications.

2. Preliminaries

In this section, we provide some information needed for the derivation of our main results and introduce the appropriate notation. We shall assume that the observations $X_1, X_2, \ldots$ form a sequence of iid random variables with distribution function $F(x)$.

Definition 2.1. The *upper* Markov-type scheme $(h, \boldsymbol{\eta}, s_0)$ is an operator transforming the sequence $X_1, X_2, \ldots$ into a set of random variables $S_0, S_1, \ldots$ defined by

$$S_0 = s_0, \;\; S_n = \psi(S_{n-1}, X_n, \boldsymbol{\eta}), \qquad n = 1, 2, \ldots, \tag{1}$$

where ψ is a suitably-chosen function and $\boldsymbol{\eta}$ is the vector of control scheme parameters. The value $h \geq 0$ serves as the decision threshold of the scheme; by definition, the Run Length (RL) N of the scheme is the first index n for which $S_n > h$. If $N < \infty$, we say that the scheme signals at time N.

If an additional signal criteria is introduced that calls for triggering an out-of-control signal at time i if a single observation X_i satisfies $X_i > c,$ the above control procedure is called a Markov-type scheme supplemented by Shewhart's control limit, c.

In a similar way, one can define the *lower* Markov-type scheme as an *upper* scheme $(h^-, \boldsymbol{\eta}^-, s_0^-)$, possibly supplemented by the Shewhart's limit c^-, applied to the reflected sequence of observations, $\{-X_i\}$. To detect changes in both directions, one can use a *two-sided* Markov-type scheme, in which the upper scheme $(h^+, \boldsymbol{\eta}^+, s_0^+)$ and the lower scheme $(h^-, \boldsymbol{\eta}^-, s_0^-)$, possibly supplemented by the respective Shewhart limits (c^+, c^-), are run in parallel (Knoth 2018). In two-sided schemes, the function ψ for the constituent upper and lower schemes is typically the same, but this assumption can be relaxed when deemed beneficial.

Below are three examples of Markov-type schemes.

(a) *Page's (Cusum–Shewhart) scheme.* Introduced by Page (1954), it is defined in terms of four parameters, $(h \geq 0, k, 0 \leq s_0 \leq h, c)$ based on the the process:

$$S_0 = s_0, \;\; S_n = max\; [0, S_{n-1} + (X_n - k)], \qquad n = 1, 2, ... \tag{2}$$

Details about the role of the scheme parameters can be found, for example, in Hawkins and Olwell (1998). Though schemes based on only two parameters, (h, k), are sometimes satisfactory for practical purposes, supplementing the scheme by a Shewhart's limit c is generally advisable as it improves the sensitivity of the scheme with respect to substantial increases in the process level. In other words, it removes some of the "inertia" of a Cusum scheme when facing a sharp change of the process (e.g., see Lucas 1982; Woodall and Mahmoud 2005). This type of scheme has a number of advantages related to design, analysis and implementation. One of them is related to the fact that analysis of two-sided schemes can be performed based on analysis of its one-sided components, see Knoth (2018). As this paper demonstrates, gradient analysis of such schemes is also considerably simpler than analysis of other Markov-type schemes, greatly enhancing its practical appeal. A number of optimality results are available for the Page's scheme, e.g., see Moustakides (1986).

(b) *Geometric Cusum.* This scheme can be viewed as a generalization of the Reflected EWMA, see Yashchin (1987). The basic decision process (1) in this case is:

$$S_0 = s_0, \;\; S_n = max\; [0, \gamma S_{n-1} + (X_n - k)], \qquad n = 1, 2, ..., \tag{3}$$

where $0 \leq \gamma \leq 1$ is the evidence-discounting factor. The additional parameter provides flexibility with respect to detecting other types of abrupt changes in the process level, such as drifts.

(c) *Girshik–Rubin scheme.* This scheme, also known as a Roberts–Shiryaev scheme, was introduced by Girshick and Rubin (1952) in the context of Bayesian analysis of control procedures. Several optimality properties of this procedure are discussed in the literature, e.g., see Pollak and Tartakovsky (2009). The decisions process can be formulated as a direct generalization of the Cusum scheme:

$$S_0 = s_0, \;\; S_n = \alpha \times ln(1 + e^{S_{n-1}/\alpha}) + X_n - k, \qquad n = 1, 2, ..., \tag{4}$$

where $\alpha > 0$ is the fourth parameter of the scheme. With $\alpha = \sigma^2/(\mu_1 - \mu_0), k = (\mu_1 + \mu_0)/2,$ where σ is the standard deviation of the observations, the scheme (4) has a certain optimality property in detecting a change in mean from μ_0 to μ_1, see Roberts

(1966). Note, however, that the Page's scheme is a special case corresponding to $\alpha \to 0$, and it has optimality properties too. Furthermore, in most practical applications, Markov-type schemes are supplemented by Shewhart's limits. Therefore, in practice it is convenient to think of α as just an additional parameter that can be used for manipulating the operating characteristics of the scheme.

A number of methods for analysis of Markov-type control schemes are discussed in the literature, e.g., see Hawkins and Olwell (1998). In this paper, we focus on the method proposed by Brook and Evans (1972), which is based on discretizing the values of $\{S_n\}$, and then treating it as a Markov Chain. It is clear that $S_0, S_1, \ldots$ form a Markov Chain which is discrete in time, but may be continuous in space. The levels $(0, h)$ are reflecting and absorbing barriers of the chain, respectively. This method can be extended to cover general Markov-type schemes, and our approach to analysis of schemes, including gradient analysis, is based on this approach.

For computational purposes we discretize the values of $S_n, n = 1, 2, \ldots$ using intervals of length

$$\delta = h/(d - 0.5), \tag{5}$$

where the positive integer d is the *level of discretization* that determines the number of states of the Markov Chain in the interval $[0, h]$. The states of the discretized scheme are thus $i * \delta, \quad i = 0, 1, \ldots, (d-1)$ and the boundaries of the corresponding subintervals are at points $(i + 0.5) * \delta$.

In other words, the values of $S_0, S_1, \ldots,$ are rounded to the center of a corresponding sub-interval. Such method of discretization usually leads to approximations of good quality relative to the computational effort invested. Studies show that levels of discretization of the order $d \approx 30$ are sufficient for most practical purposes (Yashchin 1985a, Table 1). One reason for that is related to the fact that we discretize *the states* of the Markov-type schemes rather than the observations themselves. There are also other reasons, see Sec. 5. Thus, relatively high accuracy of this method is explained by compensation of roundoff errors when computing values of the scheme.

The transition matrix P of the corresponding Markov chain can be expressed in terms of $F(x)$:

$$P_{d+1} = \begin{pmatrix} R & (I_d - R)\mathbf{1} \\ \mathbf{0}^T & 1 \end{pmatrix}, \tag{6}$$

where the elements $r_{ij}(i,j = 0, 1, \ldots d - 1)$ of $R_{d\times d}$ are computed based on the type of the scheme and properties of $F(x)$, $\mathbf{1}_{d\times 1}$ is a vector of *ones*, and I_d is a $(d \times d)$ *identity* matrix. In what follows, bold letters will represent column-vectors. In particular, we denote columns and rows of R by $\{\boldsymbol{c}_i\}$ and $\{\boldsymbol{r}_i^T\}$, where $i = 0, 1, \ldots, (d-1)$.

Analysis of the run length distribution can be performed as follows. The vector $\boldsymbol{\mu}$ containing ARL's corresponding to headstarts $0, \delta, \ldots, (d-1)\delta$ is

$$\boldsymbol{\mu} = (I_d - R)^{-1}\mathbf{1} \tag{7}$$

see Brook and Evans (1972). The survival function of the run length (for all values of the headstart) is

$$P\{RL > n\} = R^n\mathbf{1}. \tag{8}$$

Under the assumption that all the eigenvalues of R have the same algebraic and geometric multiplicities, there exist a spectral representation of the form

$$R = U \times \left[\text{diag } (\lambda_0, \lambda_1, \dots \lambda_{d-1})\right] \times U^{-1}, \tag{9}$$

where $\lambda_0 > |\lambda_1| \geq \dots \geq |\lambda_{d-1}|$ are the eigenvalues of R (note that R is *primitive* and λ_0 is its Perron-Frobenius eigenvalue); the columns of U are the corresponding right eigenvectors and the rows of U^{-1} are the corresponding left eigenvectors, see Yashchin (1985b).

Denoting by $\boldsymbol{u}_0, \boldsymbol{u}_1, \dots \boldsymbol{u}_{d-1}$ the columns of U, we obtain that

$$\begin{aligned} P\{RL > n\} &= \{w_0\boldsymbol{u}_0, w_1\boldsymbol{u}_1, \dots, w_{d-1}\boldsymbol{u}_{d-1}\} \\ &\quad \times (\lambda_0^n, \lambda_1^n, \dots \lambda_{d-1}^n)^T, \quad n = 0, 1, \dots \end{aligned} \tag{10}$$

where the weights $w_0, w_1, \dots, w_{d-1}$ are chosen so that $\sum_{i=0}^{d-1} w_i\boldsymbol{u}_i = \mathbf{1}$. In what follows, we shall always assume that the right eigenvectors are scaled so that their sum is $\mathbf{1}$, which is equivalent to requirement that $U^{-1}\mathbf{1} = \mathbf{1}$. The first term of the expansion (10) leads to the asymptotic formula discussed in Brook and Evans (1972).

The assumption we made about the same geometric and algebraic multiplicity of the eigenvectors of R holds in almost all practical situations; otherwise, use of a canonical Jordan matrix representation instead of (9) leads to results analogous to those obtained in the present work. We do not consider this possibility in the context of gradient analysis, since in the rare cases where it is relevant, it is always possible to use alternative methods to re-compute the quantities associated with a control scheme.

3. Gradient by the signal level

Let us suppose that, after performing analysis of a Markov-type scheme, we would like to examine the effect of the increase of h by δ. The matrix $\tilde{R}$ corresponding to these conditions is:

$$\tilde{R} = \begin{pmatrix} R & \boldsymbol{c}_d \\ \boldsymbol{r}_d^T & r_{dd} \end{pmatrix}, \tag{11}$$

where elements of $\boldsymbol{r}_d$ and $\boldsymbol{c}_d$ are computed in accordance with the type of scheme used.

In this section, we derive the basic quantities associated with the new scheme in terms of those corresponding to the original one. First, we denote by $\boldsymbol{p}^* = (I_d - R)^{-1}\boldsymbol{c}_d$. The j-th component of this vector has the following probabilistic meaning:

$$p_j^* = P\{\text{At the moment of signal the state of the chain is } d | s_0 = j\delta\}, \tag{12}$$

which follows from treating d as a separate absorbing state in the original scheme and analyzing the resulting transition matrix. Our first result yields $\tilde{\boldsymbol{\mu}}$, the new set of ARL's.

Theorem 3.1. *The set of ARL's, $\tilde{\boldsymbol{\mu}}$ of the modified scheme is:*

$$\tilde{\boldsymbol{\mu}} = \begin{pmatrix} \boldsymbol{\mu} + \boldsymbol{p}^* \times \ell \\ \ell \end{pmatrix}, \tag{13}$$

where

$$\ell = (1 + \boldsymbol{r}_d^T\boldsymbol{\mu})/(1 - r_{dd} - \boldsymbol{r}_d^T\boldsymbol{p}^*) \tag{14}$$

Proof. Let the modified scheme start from the state $s_0 = \delta d$. Then ℓ is its average run length, and (13) follows from the argument that the run lengths of the original and modified schemes differ only in the case where the state of the original scheme at the time of signal is d. This occurs with the probabilities $\boldsymbol{p}^*$ computed using the formula shown above; the additional run length of the modified scheme under this condition is ℓ.

The expression (14) then follows from the relationship

$$\ell = 1 + r_{dd} \times \ell + \boldsymbol{r}_d^T(\boldsymbol{\mu} + \boldsymbol{p}^* \times \ell), \tag{15}$$

that is based on (13) and the probabilities in the bottom row of the modified matrix $\tilde{R}$, see (11). This completes the proof.

As one can see, $\tilde{\boldsymbol{\mu}}$ can be computed with a minimal additional effort, as $(I_d - R)^{-1}$ is available from the analysis of original scheme. The gradient vector by h (for the set of headstarts) is thus $(\ell/\delta) \times \boldsymbol{p}^*$.

In practical applications, it is important to consider not only the ARL's, but also the Standard Deviation of the Run Length (SDRL), as well as the overall Run Length distribution, e.g., see Yashchin (1985a). Accordingly, some design procedures for h focus on quantiles of the Run Length. One approach to such design is based on the spectral representation (9). To compute gradients of Run Length quantiles, we need methods for efficient computation of (9) for the augmented matrix $\tilde{R}$. To obtain the new set of eigenvalues $\tilde{\lambda}_0, ..., \tilde{\lambda}_d$ and the right eigenvectors $\tilde{\boldsymbol{u}}_0, ..., \tilde{\boldsymbol{u}}_d$, we can use the following approach. First of all, let us find the weights $v_0, v_1, ..., v_{d-1}$ satisfying

$$\sum_{i=0}^{d-1} v_i \boldsymbol{u}_i = \boldsymbol{p}^*, \tag{16}$$

and denote

$$\begin{aligned}
C &= \{v_0\boldsymbol{u}_0, v_1\boldsymbol{u}_1, ..., v_{d-1}\boldsymbol{u}_{d-1}\} \\
D(\lambda) &= \ \text{diag}\ \left\{(\lambda - \lambda_0)^{-1}, (\lambda - \lambda_1)^{-1}, ..., (\lambda - \lambda_{d-1})^{-1}\right\} \\
\boldsymbol{a}(\lambda) &= \left\{\frac{1-\lambda_0}{\lambda - \lambda_0}, \frac{1-\lambda_1}{\lambda - \lambda_1}, ..., \frac{1-\lambda_{d-1}}{\lambda - \lambda_{d-1}}\right\} \\
\boldsymbol{\ell}_1(\lambda) &= D(\lambda) \times \mathbf{1}; \qquad \boldsymbol{\ell}_2(\lambda) = D(\lambda) \times \boldsymbol{a}(\lambda).
\end{aligned} \tag{17}$$

Then one can prove the following

Theorem 3.2. *Suppose that none of the eigenvalues $\lambda_0, ..., \lambda_{d-1}$ is also an eigenvalue of $\tilde{R}$. Then the set $\tilde{\lambda}_0, \tilde{\lambda}_1, ..., \tilde{\lambda}_d$ of eigenvalues of $\tilde{R}$ are solutions of the equation*

$$\lambda - r_{dd} - \boldsymbol{r}_d^T C \boldsymbol{a}(\lambda) = 0. \tag{18}$$

The right eigenvector $\tilde{\boldsymbol{u}}_i$ of $\tilde{R}$ corresponding to the single eigenvalue $\tilde{\lambda}_i$ is given by

$$\tilde{\boldsymbol{u}}_i = \begin{pmatrix} C \times \boldsymbol{a}(\tilde{\lambda}_i) \\ 1 \end{pmatrix}. \tag{19}$$

If all the eigenvalues of $\tilde{R}$ are distinct and weights $\tilde{w}_i$ are chosen as

$$\tilde{w}_i = \left[1 + \boldsymbol{r}_d^T U \times \boldsymbol{\ell}_1(\tilde{\lambda}_i)\right] \Big/ \left[1 + \boldsymbol{r}_d^T C \times \boldsymbol{\ell}_2(\tilde{\lambda}_i)\right], \tag{20}$$

then

$$\sum_{i=0}^{d} \tilde{w}_i \boldsymbol{u}_i = \mathbf{1}. \tag{21}$$

Proof. By the formula for a determinant of a partitioned matrix and in light of our assumptions, the characteristic equation for $\tilde{R}$ is

$$\| \tilde{R} - \lambda I_{d+1} \| = \| R - \lambda I_d \| \times \left[r_{dd} - \lambda - \boldsymbol{r}_d^T \times (R - \lambda I_d)^{-1} \boldsymbol{c}_d \right] = 0. \tag{22}$$

Further, by (9),

$$-(R - \lambda I_d)^{-1} \boldsymbol{c}_d \equiv U \times D(\lambda) \times U^{-1} \boldsymbol{c}_d \equiv C\boldsymbol{a}(\lambda), \tag{23}$$

since for $\lambda = 1$ the LHS of (23) is $\boldsymbol{p}^*$. Therefore, the equation for eigenvalues of $\tilde{R}$ reduces to (22).

To find the right eigenvector corresponding to a detected eigenvalue $\tilde{\lambda}_i$ of $\tilde{R}$, we can set its last component to 1 and then identify the remaining components as $-(R - \tilde{\lambda}_i I_d)^{-1} \boldsymbol{c}_d = C\boldsymbol{a}(\tilde{\lambda}_i)$.

Now let us find the weights $\tilde{w}_i$ so that (21) holds. First of all, the left eigenvector corresponding to a detected eigenvalue $\tilde{\lambda}_i$ can be identified as $(-\boldsymbol{r}_d^T (R - \tilde{\lambda}_i I_d)^{-1}, 1)$. Multiplying (21) by this eigenvector and using the bi-orthogonality property of systems of left and right eigenvectors results in $\tilde{w}_i = [1 + \boldsymbol{r}_d^T U D(\tilde{\lambda}_i) U^{-1} \mathbf{1}] / [1 + \boldsymbol{r}_d^T U D(\tilde{\lambda}_i) U^{-1} C\boldsymbol{a}(\tilde{\lambda}_i)]$. Finally, (20) follows from $U^{-1}\mathbf{1} = \mathbf{1}$ and $C = U \times \text{ diag } \{v_0, v_1, \ldots v_{d-1}\}$, completing the proof.

The assumption that none of the eigenvalues of R are also eigenvalues of $\tilde{R}$ in the above theorem is by no means crucial. One can see that such situation occurs if and only if $\boldsymbol{r}_d^T$ is orthogonal to appropriate right eigenvector of R. The "new" eigenvectors associated with such eigenvalues are obtained by attaching a trailing zero component to the respective "old" eigenvectors. Furthermore, (18) and (19) still enables one to find the remaining eigenvalues and eigenvectors of $\tilde{R}$.

In many practical cases we are interested not only in the set of new ARL's or new spectral representation of the transition matrix, but in the matrix $(I_{d+1} - \tilde{R})^{-1}$. This matrix can be used not only for purposes of sensitivity analysis of higher order moments, but also for solving the following problem: with all the other parameters fixed, find the maximal h for which ARL $\leq m$, where m is some prescribed number (we assume that s_0 is a multiple of δ and h increases in steps of size δ). As we shall see, there exist a simple relation between $(I_d - R)^{-1}$ and $(I_{d+1} - \tilde{R})^{-1}$. By using this relation and repeating the procedure (13), one can efficiently solve the above problem.

Let us denote

$$a = \ell / (1 + \boldsymbol{r}_d^T \boldsymbol{\mu}); \qquad \boldsymbol{b}^T = a \times \boldsymbol{r}_d^T (I_d - R)^{-1}. \tag{24}$$

Then the mentioned relation is

$$(I_{d+1} - \tilde{R})^{-1} = \begin{pmatrix} (I_d - R)^{-1} + \boldsymbol{p}^* \boldsymbol{b}^T & \boldsymbol{p}^* a \\ \boldsymbol{b}^T & a \end{pmatrix}. \tag{25}$$

The relation (25) can be proved by applying an inversion formula for partitioned matrices (ex. see Anderson 1984, p.18) to

$$(I_{d+1} - \tilde{R}) = \begin{pmatrix} I_d - R & -\boldsymbol{c}_d \\ \boldsymbol{r}_d^T & 1 - r_{dd} \end{pmatrix}. \tag{26}$$

Note that (25) can be used to prove (13) and (15); however, the proof given in the Theorem 3.1 is more interesting as it is based on a probabilistic argument only. It is also clear how one can make a "step down" i.e., find $(I_d - R)^{-1}$ once $(I_{d+1} - \tilde{R})^{-1}$ is known: determine $\boldsymbol{p}^*$ and $\boldsymbol{b}^T$ from the last column and row of $(I_{d+1} - \tilde{R})^{-1}$, respectively, and then subtract $\boldsymbol{p}^*\boldsymbol{b}^T$ from its $(d \times d)$ principal minor. Therefore, (13) enables one to compute $\boldsymbol{\mu}$ efficiently based on $\tilde{\boldsymbol{\mu}}$ and $(I_{d+1} - \tilde{R})^{-1}$. For the case of Cusum-Shewhart schemes, several examples can be found in Yashchin (1986).

4. Gradient by other scheme parameters

In this section, we consider the situation in which the basic quantities associated with the scheme have been computed and one would like to examine the effect of changing other components of the parameter vector $\boldsymbol{\eta}$ in (1). In the wake of perturbing $\boldsymbol{\eta}$, the main part of the transition matrix of the modified scheme can be represented as $R + E$, where R corresponds to the basic scheme and E is typically "small" and it depends on the parameter(s) being varied. To avoid trivialities, we assume that the upper-left elements of both R and $R + E$ are less than 1; this implies that the ARL's of both schemes are finite. We start by outlining the approach for finding the set of ARL's corresponding to the modified scheme. Denote $K = (I_d - R)^{-1}$ and $\tilde{K} = (I_d - R - E)^{-1}$. The new set of ARL's is given by

$$\begin{aligned} \bar{\boldsymbol{\mu}} &= \tilde{K} \times \mathbf{1} = (I_d - KE)^{-1} \times K\mathbf{1} \\ &\equiv \left[I_d + KE + \ldots + (KE)^n + (I_d - KE)^{-1}(KE)^{n+1}\right] \times \boldsymbol{\mu}, \end{aligned} \tag{27}$$

for every integer n; note that under our assumptions the matrices $I_d - R, I_d - R - E$ and, consequently, $I_d - KE$ are invertible. Since the norm of E is typically "small", especially for higher levels of discretization, (27) enables one to evaluate $\bar{\boldsymbol{\mu}}$ iteratively, starting from $\boldsymbol{\mu}$. Furthermore, the gradient can be approximated by the formula *Gradient* $\approx (1/\delta)(\mathrm{KE}) \times \boldsymbol{\mu}$. Denote $\boldsymbol{t}_n = (KE)^n\boldsymbol{\mu}, \quad n = 0, 1, \ldots,$ and $\boldsymbol{\mu}_n = \sum_{i=0}^n \boldsymbol{t}_i$. Then, at any stage of the iterative procedure, one is able to assess the relative approximation error by using the following

Statement 4.1. Suppose that for some n and $\epsilon > 0, |E\boldsymbol{t}_n| \le \epsilon \times \mathbf{1},$ componentwise. Then

$$|\bar{\boldsymbol{\mu}} - \boldsymbol{\mu}_n| \le \epsilon \times \boldsymbol{\mu}. \tag{28}$$

Proof. Since $\tilde{K}$ is a fundamental matrix of an absorbing Markov chain (Seneta 1981, p.122) all its elements are non-negative; therefore, the function $\psi(\boldsymbol{x}) = \tilde{K} \times \boldsymbol{x}$ is non-decreasing in every component of $\boldsymbol{x}$. Thus, by (27),

$$|\bar{\boldsymbol{\mu}} - \boldsymbol{\mu}_n| = |\tilde{K}E\boldsymbol{t}_n| \le \epsilon \times |\tilde{K} \times \mathbf{1}| \le \epsilon \times \boldsymbol{\mu}, \tag{29}$$

proving the statement.

Clearly, (28) implies that the absolute error of approximation of $\bar{\boldsymbol{\mu}}$ by $\boldsymbol{\mu}_n$ cannot exceed $|\boldsymbol{\mu}_n \times \epsilon/(1-\epsilon^2)|$.

One should take into consideration that the iterative procedure does not need, in general, to converge. There are several sets of sufficient conditions for convergence; for example, it takes place when the norm of (KE) is less than 1; when the absolute value of the dominant eigenvalue of (KE) is less than 1, etc. Define the norm of a matrix $A = (a_{ij})$ by $\parallel A \parallel = \max_i \sum_j |a_{ij}|$. Then a simple criterion for convergence of the iterative procedure can be based on the inequalities $\parallel KE \parallel \le \parallel K \parallel \times \parallel E \parallel$ and

$$\parallel K \parallel = \ \text{ARL}(0) \ \le \left(1 - \sum_j r_{0j}\right)^{-1}, \tag{30}$$

where ARL(0) is the average run length of the basic scheme with headstart 0. To prove (30) we first note that since all the elements of K are non-negative, its norm is equal to the maximal component of $K \times \mathbf{1}$, i.e., to ARL(0). Further, consider the following modification of the Markov chain corresponding to the basic scheme; if the process starts from the headstart 0, it also stays there until absorption occurs. It is clear that the average time to absorption of this chain, $(1-\sum_j r_{0j})^{-1}$, is greater than or equal to ARL(0). It is also not difficult to show that strict inequality holds in (30), provided at least one of the elements $r_{01}, r_{02}, \ldots, r_{0,d-1}$ is positive.

Next we consider two special cases pertaining to the Cusum–Shewhart schemes, see (2). In this case, we need to evaluate the effect of increasing k or the supplemental Shewhart's limit c by δ. The elements of R are given by

$$r_{ij} = \begin{cases} F^*(k + (-i + 0.5)\delta), & j = 0 \\ F^*(k + (j - i + 0.5)\delta) - F^*(k + (j - i - 0.5)\delta), & j > 0, \end{cases} \tag{31}$$

where the (improper) cumulative distribution function (cdf) F^* is defined as

$$F^*(x) = \begin{cases} F(x), & x < c \\ F(c), & x \ge c. \end{cases} \tag{32}$$

4.1. Cusum scheme gradient by the Shewhart's limit

For a Cusum–Shewhart scheme, only values $c \le h + k$ have an impact on the Run Length characteristics. Furthermore, when $c \le k$, we get a pure Shewhart scheme, for which a gradient analysis is straightforward. Therefore, we limit our discussion to the case $c \in (k, h + k)$.

In light of (31) and (32), for the i-th row of R, the trailing elements are zero starting with the first column index j for which $k + (j - i - 0.5)\delta \ge c$ (provided that this index is less than or equal to $d-1$). Consequently, for the first row ($i = 0$), the trailing zeros will be present only when the condition $0.5 + (c-k)/\delta \le d - 1$ is satisfied, and the number of trailing zeros is then

$$m = [d - (0.5 + (c-k)/\delta)] = [(d - 0.5)(1 - ((c-k)/h))]. \tag{33}$$

When $h + k \ge c > h + k - \delta$, there are no trailing zeros in the first row of R. Computing a gradient in this region is more involved, since increasing c by δ sets it to a level where it becomes irrelevant (note that the theoretical gradient has a discontinuity at $c = h + k$). One could still use the method described below—however, the increment in the argument c would need to be smaller than δ. In the rest of this section, we will assume, for the sake of simplicity, that the increment in the gradient analysis is δ and values of c are restricted to the grid $h + k - i\delta$, where i is a positive integer.

So, in all cases under consideration $m > 0$, i.e., the upper-right triangular part of the matrix R consists of zeros and E is of the following form: its right upper $(m \times m)$ submatrix is $e \times I_m$, where e is the probability of one step passage from the state 0 to $(d - 1) - m$; the remaining elements of E are zeros. First, we prove that in this case $\boldsymbol{\mu}_n \to \bar{\boldsymbol{\mu}}$ as $n \to \infty$. Indeed, if at least one of the elements $r_{01}, r_{02}, ..., r_{0,m-1}$ is positive or if $e < 1 - r_{00}$, then

$$\| KE \| < \left(1 - \sum_j r_{0j}\right)^{-1} \times e \leq 1; \tag{34}$$

therefore, $\boldsymbol{\mu}_n \to \bar{\boldsymbol{\mu}}$. Otherwise (i.e., when R is a lower triangular matrix and $e = 1 - r_{00}$), it might happen that $\| KE \| = 1$. However, as we will see from the argument below, $(KE)^n \to 0$ as $n \to \infty$, i.e., convergence still takes place.

Next we discuss a simple direct procedure for finding $\tilde{K}$. Denote by K_1 and K_2 the upper left $(d - m) \times m$ and lower left $(m \times m)$ minors of K, respectively. Then direct verification shows that

$$(KE)^n = e^n \begin{pmatrix} \mathbf{0} & K_1 K_2^{n-1} \\ \mathbf{0} & K_2^n \end{pmatrix}; \tag{35}$$

therefore,

$$\tilde{K} = \begin{pmatrix} I_{(d-m)} & e \times K_1 (I_m - e \times K_2)^{-1} \\ \mathbf{0} & (I_m - e \times K_2)^{-1} \end{pmatrix} \times K. \tag{36}$$

One can see that $(e \times K_2)^n \to 0$ as $n \to \infty$. This follows from the probabilistic interpretation of K as a fundamental matrix of a Markov Chain: it is non-negative, and its diagonal elements are all greater or equal to 1. Furthermore, (a) sums of rows of K are non-increasing and equal to $ARL(i)$, $i = 0, 1, ..., (d - 1)$, and (b) $e \times ARL(0) \leq 1$ because of (30) and (34). Therefore, the matrix $(e \times K_2)$ is *sub-stochastic*, and if $m \leq d/2$ then sums of elements in every row of this matrix are strictly less than 1. In other words, $(e \times K_2)$ can be viewed as a transition matrix of the m - state Markov Chain with an additional absorbing state, and a direct transition to an absorbing state is possible from any of the states. Such a chain will terminate in an absorbing state with probability 1, therefore, $(e \times K_2)^n \to 0$.

When $m > d/2$, some of the diagonal elements of $e \times K$ will be captured in a subdiagonal of $e \times K_2$. The number of positive elements in this sub-diagonal is $d - m$, thus it is possible that the first $d - m$ states of the chain corresponding to the transition matrix $e \times K_2$ do not permit direct passage to the absorbing state. However, the sum of the rows of $e \times K_2$ for the remaining $2m - d$ (bottom) states is strictly less than 1, and so direct transition to the absorbing state is possible for them. In order to prove that under these conditions the chain will terminate in the absorbing state with probability 1, we need to show that, for any of the top $d - m$ states, the transition to one of the bottom $2m-d$ states is possible. To show that, we can use the mentioned positive sub-diagonal inherited from K: based on its elements, we can conclude that passage from state 1 to the state corresponding to the sub-diagonal element is possible. If the latter state is among the bottom $2m-d$, we found the path to the absorbing state; otherwise, passage from this state to the state corresponding to its own sub-diagonal element is possible. In this way, we proceed down the states corresponding to sub-diagonal elements, until we eventually reach one of the bottom $2m-d$ states. Based on this reasoning, we conclude that the relationship $(e \times K_2)^n \to 0$ holds in this case as well.

Based on this result and (35), we conclude that $(KE)^n \to 0$ as $n \to \infty$. Furthermore, it proves that the matrix $I_m - e \times K_2$ in (36) is invertible. The formula (36) enables one to find $\boldsymbol{\mu}$ directly; it is especially useful in cases where $m \ll d$. This relationship holds in many practical cases, except when h is high and the change in mean that we wish to detect is small, e.g., $h = 8, k = 0.25, c = 4$, see Lucas (1982). However, even in such cases, savings related to using this formula are sizeable. One can also see that (36) provides an easy way to compute K using $\tilde{K}$, and, consequently, an easy way to evaluate the effect of decreasing c by δ.

4.2. Cusum scheme gradient by the reference value

In the case when the reference value is increased by δ, the matrix E becomes

$$E = (\boldsymbol{c}_1, \boldsymbol{c}_2 - \boldsymbol{c}_1, \ldots, \boldsymbol{c}_d - \boldsymbol{c}_{d-1}), \tag{37}$$

where $\boldsymbol{c}_i$ is the i-th column of R $(i = 1, \ldots, d-1)$ and $\boldsymbol{c}_d$ is defined in (11). For any vector $\boldsymbol{x} = (x_0, x_1, \ldots, x_{d-1})^T$, by using summation by parts we obtain

$$E\boldsymbol{x} = R \times (0, x_0 - x_1, \ldots, x_{d-2} - x_{d-1})^T + \boldsymbol{c}_d \times x_{d-1} \tag{38}$$

and consequently,

$$KE\boldsymbol{x} = (K - I) \times (0, x_0 - x_1, \ldots, x_{d-2} - x_{d-1})^T + \boldsymbol{p}^* \times x_{d-1}. \tag{39}$$

As gradient analysis by h is usually performed before analysis by k (i.e., $\boldsymbol{p}^*$ is available from the previous computations), the computational effort needed to perform a single step of iterative procedure is essentially equivalent to that needed to multiply a $(d \times d)$ matrix by a vector. Clearly, the iterative procedure of computing the ARL's corresponding to $\tilde{R}$ is initiated by assigning the value of μ to $\boldsymbol{x}$.

5. Performance

In this section, we discuss several performance aspects of the proposed methodology. In the presented examples we use Cusum–Shewhart schemes—however, the methodology is applicable to more general Markov-type schemes. First, let us focus on a case that is quite typical in applications: the observations are distributed symmetrically with the mean and standard deviation (μ, σ), but the tail of $F(x)$ is somewhat longer than that of the Normal distribution. Specifically, assume that X_i are iid with the cdf

$$F(x|\mu, \sigma) = F_\nu[(x - \mu) \times (\sigma_\nu / \sigma)], \tag{40}$$

where F_ν is the cdf of the T - distribution with ν degrees of freedom, and $\sigma_\nu^2 = \nu/(\nu - 2)$ is its variance. We will consider the case $\nu = 10, \mu = 0, \sigma = 1$ and apply the upper Cusum–Shewhart scheme ($h=5, k=1, s_0=0, c=4.5$) to the observations. We will then examine the properties of the gradient. Similar properties hold for other parameter values, but from the practical perspective, fewer alternative tools for exploring the RL properties are available when the process is on-target, and this is why we focus on $\mu=0$.

The levels of discretization used in this study are $d=2^i, \quad i=4,5,\ldots,11$, with the corresponding discretization intervals δ given by (5). We denote the ARL corresponding to the discretized scheme with $s_0=0$ by $ARL[1/d|h,k,c]$, or simply $ARL[1/d]$ (here and in what follows, the square brackets indicate that we view the ARL as a function of the discretization level). First, note that the Brook and Evans (1972) approach is in effect solving the integral equation that yields the values $ARL(s_0)$, see Hawkins and Olwell (1998, p. 154), via the method of composite mid-point quadrature rule (Faires and Burden 2003, p. 123). According to the properties of this method,

$$ARL[1/d]-ARL[0]=\xi_0\times h\times(1/d)^n+o(1/d)^n, \quad d\to\infty, \tag{41}$$

where $n=2$ and ξ_0 is a constant. Hawkins (1992) used this property to enhance the Brook and Evans approach with the help of the *Richardson Extrapolation* method, see Faires and Burden (2003, p. 209). In the context of (41), the Richardson Extrapolation constructs an improved evaluation of $ARL[0]$ in terms of the current discretization $1/d$ and the previous discretization $1/(\kappa d)$:

$$\frac{\kappa^n ARL[1/(\kappa d)]-ARL[1/d]}{\kappa^n-1}=ARL[0]+O(1/d)^{n+1}; \tag{42}$$

Instead of $\kappa=2$ used in our study, a practitioner could adopt a value more suitable for the situation at hand. The relationship (41) is the key reason why the Markov Chain approximation works well with relatively low d, as mentioned earlier, even without the enhancement (42). In Table 1 we illustrate the effect of discretization on computing the $ARL[0]$ of the continuous Cusum–Shewhart scheme. The absolute and relative deviation (in percentages) indicate relevance of the quadratic convergence property in (41), and the acceleration property of the Richardson extrapolation.

The quadratic convergence property of $ARL[1/d]$ does not hold, unfortunately, for its gradient components, $ARL'_h[0], ARL'_k[0]$ and $ARL'_c[0]$, as illustrated in the Tables 2–4. Only the linear convergence with $n=1$, in line with the Taylor's expansion, can be counted on. For derivatives by k and c this behavior follows from the expansion (27).

Table 1. Effect of discretization on evaluation of the $ARL[0]$. (AE, RE) are the absolute and relative deviation from the "true" value, assumed to be the last element in the column ARL_R corresponding to the Richardson extrapolation. (AE_R, RE_R) are the absolute and relative errors of the values obtained via the Richardson extrapolation.

d	δ	ARL	AE	RE	ARL_R	AE_R	RE_R
16	0.323	3478.314	−12.77	−0.3658	N/A	N/A	N/A
32	0.159	3487.943	−3.143	−0.0900	3491.152	0.066	0.0019
64	0.079	3490.517	−0.569	−0.0163	3491.375	0.289	0.0083
128	0.039	3490.910	−0.176	−0.00505	3491.041	−0.0454	−0.0013
256	0.020	3491.040	−0.0463	−0.00133	3491.083	−0.0030	−8.6E-05
512	0.010	3491.074	−0.0117	−0.00034	3491.086	−0.0002	−5.7E-06
1024	0.005	3491.084	−0.0020	−5.8E-05	3491.087	0.0012	3.5E-05
2048	0.002	3491.086	−0.0005	−1.4E-05	3491.086	0	0

For the derivative by h, the linear rate of convergence follows from (12) and (13), as the components p_j^* of $\boldsymbol{p}^*$ are $O(\delta)$ and, under the typical regularity conditions, follow the Taylor expansion formula; also, note that ℓ is $O(1)$ as $\delta \to 0$.

The Richardson extrapolation enables one to achieve a quadratic rate of convergence, in line with the basic formula (42), see Tables 2–4. For example, Table 2 shows that values of $ARL_h'[0]$ can be evaluated with 1.7% accuracy based on the level $d = 32$. Of course, to achieve this level of accuracy we need to perform an additional analysis with $d = 16$—however, the computational cost of this analysis is relatively low.

For $ARL_k'[0]$, we analyze both the evaluation obtained by the direct computation of $ARL(0)$ for the reference value $k + \delta$ and the evaluation based on the first term of the expansion (27). Table 3 shows that the level of discretization $d = 32$, in conjunction with the linear term of the expansion (27) and Richardson extrapolation, provides accuracy of 2.6%, which is sufficient for most practical applications. In this particular case, the result is even better than the one based on the direct computation of $ARL(0)$ based on $k + \delta$ in conjunction with the Richardson extrapolation, which gives a 15.3% accuracy (without the Richardson extrapolation, we get accuracy of only 35.5%.

Similar analysis is done for $ARL_c'[0]$, see Table 4, where we analyze the evaluations obtained by the direct computation of $ARL(0)$ for the supplemental Shewhart limit $c + \delta$ and the evaluation based on the first term of the expansion (27). With $d = 32$, we get 1.7% accuracy by employing just the linear term of (27), in conjunction with the Richardson extrapolation.

To summarize this section, the gradient of $ARL(0)$ can be obtained based on relatively low levels of discretization, and use of the Richardson extrapolation is highly advisable. A word of caution is in order, however, especially when computing $ARL_k'[0]$. For the distribution of the observations with mean μ, the behavior of the control scheme trajectory changes qualitatively as the reference value changes from the region $k < \mu$ to the region $k > \mu$, see Hawkins and Olwell (1998). When analyzing the off-target behavior for an upper scheme (i.e., $k < \mu$) with very low d, one can encounter the situation where incrementing k by δ switches the type of the underlying Markov Chain and, consequently, the behavior of the trajectories. In such cases one can run into situations where the Richardson approximation behaves erratically for small d. In practice, one can typically establish the regime of the trajectories in advance and ensure the desired degree of accuracy by taking simple precautionary measures.

In practice, especially in conjunction with the applications considered in the next section, it is rarely worthwhile to invest computational effort to secure many accurate

Table 2. Effect of discretization on evaluation of the $ARL_h'[0]$, denoted by ARL_h. (AE, RE) are the absolute and relative deviation from the "true" value, assumed to be the last element in the column ARL_h_R corresponding to the Richardson extrapolation. (AE_h_R, RE_h_R) are the absolute and relative errors of the values obtained via the Richardson extrapolation.

d	δ	ARL_h	AE	RE	ARL_h_R	AE_h_R	RE_h_R
16	0.323	517.359	−111.1	−17.68	N/A	N/A	N/A
32	0.159	567.540	−60.94	−9.697	617.721	−10.76	−1.712
64	0.079	596.435	−32.05	−5.099	625.329	−3.154	−0.502
128	0.039	612.207	−16.28	−2.590	627.980	−0.504	−0.0802
256	0.020	620.269	−8.215	−1.307	628.330	−0.154	−0.0245
512	0.010	624.357	−4.127	−0.657	628.445	−0.0389	−0.0062
1024	0.005	626.415	−2.069	−0.329	628.474	−0.0102	−0.0016
2048	0.002	627.450	−1.034	−0.165	628.484	0	0

Table 3. Effect of discretization on evaluation of the $ARL'_k[0]$, denoted by ARL_k. (AE, RE) are the absolute and relative deviation from the "true" value, assumed to be the last element in the column ARL_k_R corresponding to the Richardson extrapolation. (AE_k_R, RE_k_R) are the absolute and relative errors of the values obtained via the Richardson extrapolation. ARL_k1 is the gradient estimated based on the disturbance matrix E given by (37), in conjunction with $n = 1$ in the expansion formula (27), and ARL_k1_R is the corresponding value of the Richardson extrapolation. (AE_k1_R, RE_k1_R) are its absolute and relative errors.

d	δ	ARL_k	AE	RE	ARL_k_R	AE_k_R	RE_k_R	ARL_k1	ARL_k1_R	AE_k1_R	RE_k1_R
16	0.323	1146	−1439	−55.67	N/A	N/A	N/A	2023	N/A	N/A	N/A
32	0.159	1669	−916.9	−35.46	2191	−394.5	−15.26	2271	2519	−66.88	−2.587
64	0.079	2066	−519.4	−20.09	2464	−121.9	−4.715	2419	2566	−19.23	−0.7439
128	0.039	2310	−275.6	−10.66	2554	−31.68	−1.225	2500	2581	−4.266	−0.165
256	0.020	2444	−141.8	−5.484	2578	−8.024	−0.3103	2542	2584	−1.068	−0.04131
512	0.010	2514	−71.85	−2.779	2584	−1.919	−0.07424	2564	2585	−0.1802	−0.00697
1024	0.005	2549	−36.12	−1.397	2585	−0.3894	−0.01506	2575	2586	0.04454	0.001723
2048	0.002	2567	−18.06	−0.6985	2586	0	0	2580	2586	0.1083	0.004188

Table 4. Effect of discretization on evaluation of the $ARL'_c[0]$, denoted by ARL_c. (AE, RE) are the absolute and relative deviation from the "true" value, assumed to be the last element in the column ARL_c_R corresponding to the Richardson extrapolation. (AE_c_R, RE_c_R) are the absolute and relative errors of the values obtained via the Richardson extrapolation. ARL_c1 is the gradient estimated based on the disturbance matrix E, in conjunction with $n = 1$ in the expansion formula (27), and ARL_c1_R is the corresponding value of the Richardson extrapolation. (AE_c1_R, RE_c1_R) are its absolute and relative errors.

d	δ	ARL_c	AE	RE	ARL_c_R	AE_c_R	RE_c_R	ARL_c1	ARL_c1_R	AE_c1_R	RE_c1_R
16	0.323	5603	692.6	14.1	N/A	N/A	N/A	3688	N/A	N/A	N/A
32	0.159	5280	369.5	7.526	4957	46.48	0.9466	4258	4827	−83.04	−1.691
64	0.079	5099	188.7	3.843	4918	7.845	0.1598	4573	4889	−21.21	−0.4321
128	0.039	5005	94.69	1.928	4911	0.6898	0.01405	4739	4904	−6.124	−0.1247
256	0.020	4957	47.44	0.9662	4910	0.1937	0.003946	4824	4909	−1.489	−0.03033
512	0.010	4934	23.74	0.4836	4910	0.04544	0.000925	4867	4910	−0.3717	−0.00757
1024	0.005	4922	11.88	0.2419	4910	0.01428	0.000291	4888	4910	−0.08959	−0.00183
2048	0.002	4916	5.939	0.121	4910	0	0	4899	4910	−0.02588	−0.00053

digits, as the role of the gradient is typically to assist in design and analysis of control procedures, and low-accuracy approximations can often be tolerated. For example, if our objective is to find the threshold h using some gradient method, and the gradient is only "reasonably" accurate, the search procedure can still be made reliable and accurate to any desired level. Furthermore, even the said search procedures can tolerate some inaccuracy: if the desired h should provide, say, an $ARL(0) = 1000$ but is actually providing $ARL(0)$ within 5% of that, it is not likely to cause long-term issues, given the high uncertainty about the assumed nature of the underlying models and other system parameters. When allocating computational resources in a large-scale monitoring system, one will typically take into account many factors, and accuracy is only one of them. For such systems, ability to run on demand within a pre-specified time segment, run at a higher clip (or in real time) and energy consumption/carbon footprint are likely to emerge as factors that justify some sacrifice of accuracy.

6. Applications

In this section, we discuss several applications, with focus on high-complexity cases encountered in practice.

6.1. Design and analysis of schemes, known F(x)

This problem is encountered in systems for monitoring massive data streams when on-target data distribution is either assumed known, or is set to some "worst" member of the family of distributions that define acceptable process performance. Essentially, we are seeking scheme parameters that will, for example, ensure a low overall rate of false alarms for any possible on-target scenario. Methods described above will typically enable efficient design based on the desired tradeoff between false alarms and sensitivity requirements.

In this setting, we do not rely on Phase-1 data, since there is no guarantee that the future on-target behavior will be similar to that represented by the historic data: rather, we are focusing on the set of scenarios represented by various hypothesized data models. An example of a system based on this approach can be found in Yashchin (2018).

In the process of scheme design, it is typically important to obtain a good initial set of scheme parameters, and then to use gradient analysis to fine-tune them. In this context, diffusion process approximations tend to be very helpful (Johnson and Bagshaw 1974; Bagshaw and Johnson 1975). Of special importance are the threshold overshoot corrections that greatly reduce the approximation error, see Siegmund (1985). For Cusum schemes, under the assumption that the mean and standard deviation of the iid observations is (μ_0, σ), the $ARL(0)$ approximation is

$$ARL(h|k, s_0 = 0, \mu_0, \sigma) \approx 2\tilde{h}^2 * \frac{1}{(2a)^2}[\exp(-2a) + 2a - 1] \tag{43}$$

where

$$\tilde{h} = h/\sigma + \zeta, \quad a = -\tilde{h} * (k - \mu_0)/\sigma, \tag{44}$$

and ζ is the distribution-dependent threshold overshoot correction. When $a = 0$, (43) yields the limit value, $(\tilde{h})^2$. For Normally distributed observations, $\zeta = 1.166$. It is not difficult to obtain ζ for other distributions using simulation or formulas provided in Siegmund (1985). In many cases, using $\zeta = 1.166$ leads to acceptable results even for non-Normal families, so it is advisable to try it first before investing effort in obtaining distribution-specific corrections. Several gradient analysis formulas for the diffusion-approximated RL characteristics can be found in Yashchin (1986).

We illustrate use of our methodology for rapid design of Cusum–Shewhart schemes used in conjunction with a *search engine* application of type described in Yashchin (2018). In such applications, analyses of numerous data sequences are performed simultaneously, at times that are either scheduled in advance or determined based on other criteria, such as the current process status. The control scheme parameters (especially signal levels) often need to be determined at the time of analysis, based on the intensity of data streams or on other concomitant information. For example, choice of the on-target $ARL(0)$ can depend on the amount of data that happen to fall into the current analysis window: larger volumes will often require setting $ARL(0)$ at a higher level in order to establish a tighter control of per-analysis false alarm probabilities. Ability to design numerous control schemes "on the fly" is thus highly valued for this type of applications and it is often a limiting factor in ability of the search engine to operate at an aggressive schedule.

Let us suppose that for a particular control sequence $\{X_i\}$, the elements are iid from the cdf (40) with the parameters $\nu = 10, \mu = 0, \sigma = 1$. We need to design an upper Cusum–Shewhart scheme with $(k = 1, s_0 = 0)$; the parameters $(h,\ c)$ should be determined based on (a) $ARL(0) = 3500$ and (b) the requirement that introduction of the Shewhart's supplemental level c reduce the $ARL_{pure}(0)$ of the pure Cusum scheme by 10%. In other words, h is determined based on the requirement that $ARL_{pure}(0) = 3500/(1-0.1) = 3889$.

Such a kind of requirement is often motivated by practical considerations. Generally, pure Cusum schemes are rarely used in practice because the magnitude of unfavorable changes is typically unpredictable and pure Cusum schemes are too slow in detecting large changes. Therefore, a supplemental Shewhart limit is necessary, and we are willing to pay a price in terms of $ARL(0)$ when the process is on target (in our case, $\mu = 0$). However, we wish to limit its impact to 10%—in other words, we want the vast majority of false alarms to involve accumulation of evidence.

So, our first step is to design a pure Cusum chart with $ARL_{pure}(0) = 3889$, and we start with the Brownian Motion approximation (43) with $\zeta = 1.166$. Solving it yields the initial value, $h_0 = 3.315$. Subsequent analysis using the Richardson extrapolation based on the levels of discretization $d = (16, 32)$ of type shown in Table 1 gives the corresponding $ARL_0(0) = 1197$. This value is too far from 3889, so we compute the gradient by h, which is also based on the Richardson extrapolation with $d = (16, 32)$, as in Table 2: $ARL_h R = 1717$. To get the next approximation, h_1, we use the fact that for $\mu_0 < k$, the dominant term in (43) is $\exp(-2a)$—therefore, $ln[ARL(0)]$ shows less curvature as a function of h than $ARL(0)$. The initial value of $ln[ARL(0)]$ is $ln(1197) = 7.088$ and its gradient by h is $1717/1197 = 1.434$. The target value for $ln[ARL(0)]$ is $ln(3889) = 8.266$. Thus, the next approximation to the signal level is $h_1 = 3.315 + (8.226 - 7.088)/1.434 = 3.315 + 0.822 = 4.137$.

Performing the Markov Chain analysis of the pure Cusum $(h = 4.137, k = 1)$ using the Richardson extrapolation with $d = (16, 32)$ gives $ARL(0) = 3849$, which is roughly within 1% of the target value of 3889, so our search for h is complete: only one step in the direction of the gradient was needed to achieve that. Next we will find the supplemental Shewhart limit c that reduces the $ARL(0)$ to 3500. To obtain the *upper bound* for the initial value c_0 we solve the rate equation

$$\frac{1}{3849} + P\{X_i > c | \nu = 10, \mu = 0, \sigma = 1\} = \frac{1}{3500}, \tag{45}$$

which gives $c = 6.02$. Since this value falls outside the interval $(k, h + k) = (1, 5.137)$, we will set c_0 to the 90%—value inside this interval: $c_0 = 1 + 0.9 \times 4.137 = 4.885$. Since our intent is to use the level of discretization $d = 32$, let us adjust this value to the grid $h + k - i \times \delta:$ as by (5), $\delta = 4.137/(32 - 0.5) = 0.1313$, the adjusted value is $c_0 = 5.137 - 2 \times 0.1313 = 4.874$. Analysis of the scheme $h = 4.137, k = 1, c = 4.874$ with $d = 32$ gives $ARL(0) = 3402$. The gradient by c based on $d = 32$ and the linear term in the expansion (27) is 1691 (with the Richardson extrapolation we get a more accurate value, 1938—but this requires an additional analysis with $d = 16$, so let us skip it). Once again, we switch to the log-space: the current value of $ln[ARL(0)]$ is $ln(3402) = 8.132$ and its gradient by c is $1691/3402 = 0.497$. The target value is $ln(3500) = 8.161$, therefore the next iteration is $c_1 = 4.874 + (8.161 - 8.132)/0.497 = 4.932$. The final verification of the scheme $h = 4.137, k = 1, c = 4.932$ with $d = 32$, using Richardson extrapolation gives $ARL(0) = 3517$; all the digits are correct, as confirmed by a more refined analysis. Since this result is within 1% of the desired value, the design of the Cusum–Shewhart scheme is complete. As in the case of h, we only needed one iteration to achieve the desired accuracy. With the help of the mentioned Richardson extrapolation in the gradient computation, we obtain $ARL(0) = 3504$. Though the improvement in accuracy is sizeable, it is of no practical importance here.

6.2. Design and analysis based on phase-1 data

In this class of applications, we have a sample $(x_1, x_2, \ldots, x_n)$ of the data series, which is assumed to be iid. The objective is to analyze the performance of a given Markov-type control scheme, when applied to the population from which the data is sampled. With the help of such samples, one can then design a control scheme to ensure low false alarm rate or other performance characteristics.

In this context, we can estimate the unknown $F(x)$ by a suitably chosen $\hat{F}(x)$, and then estimate the ARL, SDRL and other RL-related quantities by substituting $\hat{F}$ for F in the Markov-chain analysis. This yields the estimates like $A\hat{R}L$, and one can subsequently obtain confidence bounds for ARL using standard techniques like re-sampling analysis or delta-method. In the analysis based on $\hat{F}$, it is advisable to use re-parametrization of the composite parameters like ARL and SDRL. Specifically, good results are typically obtained by using new parameters $(\boldsymbol{\theta}, \boldsymbol{\rho})$ defined via:

$$\boldsymbol{\theta} = \frac{1}{ARL}; \qquad \boldsymbol{\rho} = 1 - \left(\frac{SDRL}{ARL}\right)^2, \tag{46}$$

see Yashchin (1992). For example, for an *upper* scheme in the context of the Markov Chain analysis, the above formula pertains to the corresponding sets $\{ARL(i), SDRL(i)\}$, and the new parameters are, in light of (7),

$$\boldsymbol{\theta}^+ = \frac{1}{\boldsymbol{\mu}}; \qquad \boldsymbol{\rho}^+ = 2 - \frac{2K\boldsymbol{\mu} - \boldsymbol{\mu}}{\boldsymbol{\mu}^2}. \tag{47}$$

Analogously, one can compute $(\boldsymbol{\theta}^-, \boldsymbol{\rho}^-)$ for lower schemes. This transformation enables one not only to evaluate SDRL, but also to judge how far away is the RL distribution from the geometric form, which typically approximates it well under on-target conditions. Furthermore, this parametrization simplifies the analysis of two-sided procedures: for example, for the Cusum–Shewhart case one can take advantage of the relations $\boldsymbol{\theta} = \boldsymbol{\theta}^+ + \boldsymbol{\theta}^-$ and $\boldsymbol{\rho} = \boldsymbol{\rho}^+ + \boldsymbol{\rho}^-$ between the characteristics of one-sided and two-sided schemes. Finally, working with $(\boldsymbol{\theta}, \boldsymbol{\rho})$ ensures that the corresponding estimates have moments, which helps to secure their good theoretical properties.

Re-sampling analysis: To obtain confidence bounds for $(\boldsymbol{\theta}, \boldsymbol{\rho})$ based on $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\rho}})$ one could apply Bootstrap, Jackknife or Infinitesimal Jackknife. For every re-sampled data set, a new empirical Markov transition matrix is constructed, and the re-sampled estimates like $(\hat{\boldsymbol{\theta}}_b, \hat{\boldsymbol{\rho}}_b)$ for $b = 1, 2, \ldots, B$ are obtained. Use of gradient analysis is very helpful in this process, since the re-sampled estimate of the transition matrix can be represented as a perturbation of the basic one, i.e., $\hat{R}_b = \hat{R} + E_b$, and methods of Sec. 3 can be applied to approximate the re-sampled estimates $(\hat{\boldsymbol{\theta}}_b, \hat{\boldsymbol{\rho}}_b)$.

In this context, Bootstrap estimation tends be expensive relative to Jackknife in the case of Cusum–Shewhart schemes. One of the reasons for that is that the Jackknife analysis takes better advantage of the special form of R in this case. Specifically, removal of the j-th data point leads to the transition matrix

$$\hat{R}_{(j)} = \hat{R} + E_{(j)}, \tag{48}$$

where $E_{(j)} = (\hat{R} - B_j)$ and B_j is a very sparse 0/1 matrix. This leads to an efficient approximation of the gradient via

$$\boldsymbol{\mu}_{(j)} = [I_d + (KE_{(j)}) + O(KE_{(j)})^2] \times \boldsymbol{\mu} \tag{49}$$

that is used in the Infinitesimal Jackknife, and to simple estimates $\hat{\sigma}(\hat{\boldsymbol{\theta}})$.

Parametric analysis: In this case F is known up to the vector of parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_k)$. Based on our data sample, we obtain an estimate $F(x|\hat{\boldsymbol{\beta}})$. Our primary objective is inference for $[\boldsymbol{\theta}(\boldsymbol{\beta}), \boldsymbol{\rho}(\boldsymbol{\beta})]$; this can be achieved with the help of the *delta-method*. The gradient needed for this method is obtained by using techniques of Sec. 3. For example, to compute the gradient $\nabla\boldsymbol{\theta}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$, we analyze the effect of change in $\hat{R}$ resulting from the perturbation of the i-th component, $\hat{\beta}_i \Rightarrow \hat{\beta}_i + \epsilon$. This change can be represented as $\hat{R} \Rightarrow \hat{R} + E_i$, leading to efficient approximation.

Note that re-sampling analysis can also be used in this setting (e.g., parametric bootstrap) and gradient methods can be used in this context as well.

6.3. Simulation-based analysis of control schemes

In many practical situations, observations $\{X_n\}$ are complex functions of other data. Obtaining values of $F(x)$ needed for constructing R can be very costly. At the same time, one can generate random variables from this distribution inexpensively, based on the assumed model for $\{X_n\}$. We can use this (a) to produce a large sample of observations from $F(x)$, (b) to construct an estimate $\hat{F}(x)$, (c) to conduct inference for $(\boldsymbol{\theta}, \boldsymbol{\rho})$ and other RL characteristics based on this estimate, as illustrated in Sec. 5.

This approach has proven to be of high practicality in many monitoring problems. For example, in semiconductor manufacturing it is very important to control not only the overall process variability, but also the variance components. For some complex variance components it is difficult to represent $F(x)$ and construct the transition matrix—yet, simulation schemes related to these components are readily available. Using gradient analysis in conjunction with simulation leads to relatively easy analysis and design of control schemes in this setting, e.g., see Yashchin (1995).

6.4. Design and analysis of schemes with serially correlated observations

In practice, data streams often deviate from the iid model. For example, observations $\{X_n\}$ can be serially correlated. In such situations, the described gradient analysis can still lead to useful approximations. For example, Yashchin (1993) showed that performance characteristics of schemes involving serially correlated observations can often be approximated by those based on suitably transformed iid observations. The transformation from the iid $F(x)$ to the approximating $F^*(x)$ that accounts for onset of serial correlation can be represented in terms of a suitable perturbation E, enabling approximation of ARL's and other RL characteristics of schemes for serially correlated observations using methodology described in Secs. 3 and 4.

In this section, we consider the situation where the monitored observations $\{X_i\}$ are generally iid Gaussian with standard deviation σ_m and a Cusum scheme (2) is used to monitor its mean, μ. However, under some process conditions, onset of serial correlation causes substantial increase in the rate of false alarms, and it is necessary to robustify the scheme so as to prevent them. Such conditions are sometimes predictable, e.g., when they follow equipment maintenance or activation of a control loop; in other cases, they need to be detected separately, using procedures available in the literature, see Knoth and Schmid (2004). Under the correlation-inducing change, $\{X_i\}$ follow the $AR(1)$ process:

$$X_i = \mu + \phi(X_{i-1} - \mu) + E_i, \quad i = 1, 2, \ldots, \tag{50}$$

where the innovations E_i are iid Gaussian with mean 0 and standard deviation σ_0. In this context, σ_m is the standard deviation of the marginal distribution, and $\sigma_m^2 = \sigma_0^2/(1-\phi^2)$. In what follows, we will assume, without loss of generality, that (a) σ_m is unaffected by the change and (b) the on-target mean is $\mu_0 = 0$. The adjustments to the arguments below needed to accommodate situations where (a) or (b) do not hold are straightforward.

The signal level h of our scheme (2) is selected so as to provide some nominal false alarm rate characterized by $ARL0$—unless stated otherwise, we will assume that it corresponds to μ_0 and $s_0 = 0$. Upon discovery of the onset of serial correlation, one can take one of two actions. The first one is to continue the scheme (2) as is, but adjust h to some h_x, so as to preserve the nominal $ARL0$. The second course of action is to switch to the transformed sequence,

$$Y_i = (X_i - \phi X_{i-1})/(1-\phi), \quad i = 1, 2, \ldots, \tag{51}$$

and apply to it (2) with the same reference value k and a signal level h_y chosen so as to preserve $ARL0$. In this paper, we focus on the second approach (Yashchin 1993 discusses relative merits of both approaches). One advantage of this approach is that $\{Y_i\}$ form an iid sequence that preserve the mean of $\{X_i\}$ and enable a more straightforward implementation of a supplemental Shewhart limit. In what follows, we will show how h_y can be approximated with the help of the gradient analysis. Note that

$$\sigma_Y = \sigma_m f_\phi, \quad \text{where} \quad f_\phi = \sqrt{(1+\phi)/(1-\phi)}. \tag{52}$$

When designing (2) for the iid $\{X_i\}$, it is useful to save the gradients by the key scheme parameters. For our application, we use the *gradient exchange ratios*: for example, the ratio g_{kh} is defined as

$$g_{kh} = -ARL_k'[0]/ARL_h'[0], \tag{53}$$

where the derivatives are computed at the nominal parameters of the scheme (2) for the iid $\{X_i\}$ with on-target mean $\mu_0 = 0$. The interpretation of g_{kh} is as follows: if k in (2) is decreased by some Δ_k, then one will need to increase h by $\Delta_h = \Delta_k \times g_{kh}$ to preserve the nominal $ARL0$.

To obtain h_y, let us divide the scheme (2) for $\{Y_i\}$ and its signal level h_y by f_ϕ and denote the scaled down values using the "tilde" symbol; the corresponding scheme is:

$$\tilde{S}_0 = \tilde{s}_0, \; \tilde{S}_n = max \left[0, \tilde{S}_{n-1} + (\tilde{Y}_n - k/f_\phi)\right], \qquad n = 1, 2, ..., \tag{54}$$

signal when $\tilde{S}_n > \tilde{h} = h_y/f_\phi$. Clearly, when a supplemental Shewhart limit is used, it needs to be scaled down by f_ϕ as well. Since the distribution of $\tilde{Y}_n$ is the same as that of the iid X_n for which the $ARL0$ was originally computed, i.e., it is $N(0, \sigma_m)$, we can use the gradient analysis to obtain $\tilde{h}$ and then h_y. The formula is:

$$h_y = f_\phi \times \left[h + k\left(\frac{1}{f_\phi} - 1\right) \times g_{kh}\right]. \tag{55}$$

To illustrate this method, consider the scheme $(h = 3.93, k = 0.5)$ applied to the standard Gaussian distribution, i.e., $\mu_0 = 0, \sigma_m = 1$. For this scheme, $ARL0 = 312$; we do not use the supplemental Shewhart limit here since $h + k = 4.43$ is small. The gradient analysis ($d = 32$, using linear term in (27) with Richardson extrapolation) gives $ARL'_k[0] = 2027$ and $ARL'_h[0] = 322.7$, so the gradient exchange ratio is $g_{kh} = -2027/322.7 = -6.28$. Onset of serial correlation turns $\{X_i\}$ into an AR(1) series with $(\phi = 0.2, \sigma_0 = 0.9798)$, so that $\sigma_m = 1$. Since $f_\phi = 1.2247$, the formula (55) yields $h_y = 1.2247 \times [3.93 + 0.5 \times (1/1.2247 - 1) \times (-6.28)] = 5.52$. This value gives a slightly lower ARL0 of 290 instead of the nominal 312 (which corresponds to $h_y = 5.59$)—however, it is reasonably close to be of practical value. Note that if the onset of serial correlation affects σ_m (so that the marginal standard deviation of $\{X_i\}$ becomes, say, σ'_m), the argument remains the same, except that a modified value $f_\phi = (\sigma'_m/\sigma_m) \times \sqrt{(1 + \phi)/(1 - \phi)}$ is used in the formulas. Methods in this section can also be easily extended to the case of general stationary $ARMA(p, q)$ processes.

6.5. Design of more complex schemes

Methods presented in this paper can be used with any Markov-type scheme—however, the procedures are especially efficient for the Cusum–Shewhart schemes, due to the special structure of the matrices involved. One can use this fact to enhance the performance of more complex schemes, as many of them contain the Cusum–Shewhart scheme as a special case. For example, consider the Geometric Cusum (3). Introducing a moderate value of the evidence-discounting factor γ enables one to enhance the scheme performance with respect to changes other than abrupt shift of process parameters from one level to another—however, finding h that provides the necessary $ARL(0)$ involves a somewhat more tedious process. In this process, getting an initial level h_0 helps one to greatly speed up the search process. This can be done using the following procedure:

Step 1. After designing the Cusum–Shewhart scheme, compute the gradients of ARL(0) by h and γ. The latter is computed by creating a perturbation matrix E_γ based on a small decrement of γ from 1, and using the expansion (27); in most cases, the linear term will provide sufficient accuracy. Create the gradient exchange ratio $g_{\gamma h} = -ARL'_\gamma[0]/ARL'_h[0]$, as in (53).

Step 2. For the target level of γ_0 (which depends on the application, and is typically around 0.8) compute the necessary adjustment of h to the value h_0 that preserves ARL(0), using $g_{\gamma h}$ and a formula similar to (55). As γ decreases, the corresponding h will decrease too.

Step 3. Validate the solution, adjust as needed.

In a similar way, one can design a Girshik–Rubin scheme: as noted in Sec. 2, Cusum–Shewhart is a special case corresponding to $\alpha = 0$. Typically, one will have a target value α_0 for α based on the particular application and the desired level of tradeoff between two forms of optimality involved. For the initial Cusum–Shewhart scheme, gradient analysis is performed. For α, this analysis is based on the perturbation matrix E_α in conjunction with (27). Subsequently, gradient exchange ratio $g_{\alpha h}$ is computed as above and used to adjust the signal level as α increases. The details are omitted.

7. Conclusions

Based on our research and practical experience, gradient and perturbation analysis play a key role in several problems related to development and deployment of Markov-type control schemes. As illustrated in Secs. 3, 4 and 5, gradient analysis typically requires a small fraction of resources related to design of control schemes.

Discretization of control schemes is a useful technique that enables both efficient performance analysis and gradient analysis. The resulting formulation of the design and analysis in terms of Markov Chains has a universal appeal, and it can be implemented efficiently, taking into account that many analytic and software tools available to practitioners. It is especially useful in conjunction with the Richardson extrapolation that helps to achieve high accuracy using low discretization levels.

Phase 1 data can be used efficiently, in conjunction with gradient analysis, to design control schemes and to evaluate their properties. In the absence of such data, gradient analysis is useful in conjunction with pre-defined acceptable/unacceptable regions in the parameter space, and with a set of scenarios.

Re-sampling techniques are useful for inference and for working with analytically complex distributions that are relatively easy to simulate. They can be used effectively in several contexts, in conjunction with gradient analysis.

Gradient analysis is useful in design of complex schemes, including schemes with serially correlated data, based on simple schemes that can be typically designed with minimal computational investment. In many cases, just one step in the gradient direction yields results that are sufficiently accurate for practical use.

There are still many open problems in this area that justify research investment. Many of the control schemes used in practice are non-Markovian, and availability of efficient design methodologies would be highly valuable. For example, schemes based on Generalized Likelihood Ratio (GLR) are of this type. Furthermore, monitoring for multivariate serially correlated data or spatio-temporal data could benefit greatly from gradient-type methods. As analysis of more complex monitoring problems often include simulation, development of gradient-based techniques in conjunction with simulation-assisted design of control systems would also be highly valued by practitioners.

Acknowledgments

I am grateful to the Editors and the Referees, whose work helped to improve substantially the quality of this article.

References

Amin, R. W., H. Wolff, W. Besenfelder, and R. Baxley. 1999. EWMA control charts for the smallest and largest observations. *Journal of Quality Technology* 31 (2):189–206. doi:10.1080/00224065.1999.11979914.

Anderson, T. W. 1984. *An introduction to multivariate statistical analysis*. New York: Wiley.

Bagshaw, M., and R. A. Johnson. 1975. The effect of serial correlation on the performance of Cusum tests II. *Technometrics* 17 (1):73–80. doi:10.2307/1268003.

Brook, D., and D. A. Evans. 1972. An approach to the probability distribution of Cusum run length. *Biometrika* 59 (3):539–49. doi:10.2307/2334805.

Capizzi, G. 2015. Recent advances in process monitoring: Nonparametric and variable-selection methods for phase I and phase II (with discussion). *Quality Engineering* 27 (1):44–80. doi:10.1080/08982112.2015.968046.

Dietrich, B. L., E. C. Plachy, and M. F. Norton. 2014. *Analytics across the enterprise*. New York: IBM Press, Pearson.

Faires, J. D., and R. Burden. 2003. *Numerical methods*. 3rd ed. Pacific Grove: Thomson Learning.

Fu, M. C., and J.-Q. Hu. 1999. Efficient design and sensitivity analysis of control charts using Monte Carlo simulation. *Management Science* 45 (3):395–413. doi:10.1287/mnsc.45.3.395.

Fu, M. C., S. Lele, and T. W. M. Vossen. 2009. Conditional Monte Carlo gradient estimation in economic design of control limits. *Production and Operations Management* 18 (1):60–77. doi:10.1111/j.1937-5956.2009.01005.x.

Girshick, M., and H. Rubin. 1952. A Bayes approach to a quality control model. *The Annals of Mathematical Statistics* 23 (1):114–25. doi:10.1214/aoms/1177729489.

Hawkins, D. M. 1992. Evaluation of average run lengths of cumulative sum charts for an arbitrary data distribution. *Communications in statistics—Simulation and computation* 21 (4):1001–20. doi:10.1080/03610919208813063.

Hawkins, D. M., and D. H. Olwell. 1998. *Cumulative sum charts and charting for quality improvement*. New York: Springer.

Huang, W., L. Shu, and W. Jiang. 2016. A gradient approach to the optimal design of Cusum charts under unknown mean—shift sizes. *Journal of Quality Technology* 48 (1):68–83. doi:10.1080/00224065.2016.11918152.

Huang, W., L. Shu, and W. Jiang. 2018. A gradient approach to efficient design and analysis of multivariate EWMA control charts. *Journal of Statistical Computation and Simulation* 88 (14):2707–25. doi:10.1080/00949655.2018.1483367.

Johnson, R. A., and M. Bagshaw. 1974. The effect of serial correlation on the performance of Cusum tests. *Technometrics* 16 (1):103–22. doi:10.2307/1267498.

Knoth, S. 2018. New results for two-sided Cusum–Shewhart control charts. In *Frontiers in statistical quality control*, ed. S. Knoth and W. Schmid, vol. 12, 45–63. Berlin: Springer.

Knoth, S., and W. Schmid. 2004. Control charts for time series: A review. In *Frontiers in Statistical Quality Control*, ed. H.J. Lenz and P.T. Wilrich, vol. 7 210–36. Berlin: Springer.

Lele, S. 1996. Steady state analysis of three process monitoring procedures in quality control. PhD thesis, University of Michigan.

Lucas, J. M. 1982. Combined Shewhart-Cusum quality control schemes. *Journal of Quality Technology* 14 (2):51–9. doi:10.1080/00224065.1982.11978790.

Lucas, J. M., and M. S. Saccucci. 1990. Exponentially weighted moving average control schemes: Properties and enhancements. *Technometrics* 32 (1):1–12. doi:10.1080/00401706.1990.10484583.

Moustakides, G. V. 1986. Optimal stopping times for detecting changes in distributions. *The Annals of Statistics* 14 (4):1379–87. doi:10.1214/aos/1176350164.

Negandhi, V., L. Sreenivasan, R. Giffen, M. Sewak, and A. Rajasekharan. 2015. *IBM predictive maintenance and quality 2.0 technical overview*. Poughkeepsie: IBM Redbooks.

Page, E. 1954. Continuous inspection schemes. *Biometrika* 41 (1–2):100–15.

Philips, T., E. Yashchin, and D. Stein. 2003. Using statistical process control to monitor active managers. The Journal of Portfolio Management 30 (1):86–94. doi:10.3905/jpm.2003.319922.

Pollak, M., and A. G. Tartakovsky. 2009. Optimality properties of the Shiryaev–Roberts procedure. *Statistica Sinica* 19:1729–39.

Polunchenko, A. S., and V. Raghavan. 2018. Comparative performance analysis of the cumulative sum chart and the Shiryaev–Roberts procedure for detecting changes in autocorrelated data. *Applied Stochastic Models in Business and Industry* 34 (6):922–48. doi:10.1002/asmb.2372.

Roberts, S. W. 1966. A comparison of some control chart procedures. *Technometrics* 8 (3): 411–30. doi:10.1080/00401706.1966.10490374.

Saleh, N. A., M. A. Mahmoud, L. A. Jones-Farmer, I. Zwetsloot, and W. H. Woodall. 2015. Another look at the EWMA control chart with estimated parameters. *Journal of Quality Technology* 47 (4):363–82. doi:10.1080/00224065.2015.11918140.

Seneta, E. 1981. *Non-negative matrices and Markov chains*. New York: Springer Verlag.

Shu, L., W. Huang, and W. Jiang. 2014. A novel gradient approach to optimal design and sensitivity analysis of EWMA control charts. *Naval Research Logistics* (*NRL*) 61 (3):223–37. doi:10.1002/nav.21579.

Siegmund, D. 1985. *Sequential analysis*. New York: Springer.

Woodall, W. H. 1983. The distribution of the run length of one-sided Cusum procedures for continuous random variables. *Technometrics* 25 (3):295–300. doi:10.2307/1268615.

Woodall, W. H. 1986. The design of Cusum quality control charts. *Journal of Quality Technology* 18 (2):99–102. doi:10.1080/00224065.1986.11978994.

Woodall, W. H., and M. A. Mahmoud. 2005. The inertial properties of quality control charts. *Technometrics* 47 (4):425–36. doi:10.1198/004017005000000256.

Yashchin, E. 1985a. On the analysis and design of Cusum–Shewhart control schemes. *IBM Journal of Research and Development* 29 (4):377–91. doi:10.1147/rd.294.0377.

Yashchin, E. 1985b. On a unified approach to the analysis of two-sided cumulative sum schemes with headstarts. *Advances in Applied Probability* 17 (3):562–93. doi:10.2307/1427120.

Yashchin, E. 1986. Sensitivity analysis of Cusum–Shewhart control schemes. IBM Research Report RC 12078, Yorktown Heights, New York. Available online at http://researcher.watson.ibm.com/researcher/files/us-yashchi/Sensitivity_analysis_of_cusum_shewhart_schemes_rc_12078_OCR_1986.pdf.

Yashchin, E. 1987. Some aspects of the theory of statistical control schemes. *IBM Journal of Research and Development* 31:199–205.

Yashchin, E. 1992. Analysis of Cusum and other Markov—type control schemes by using empirical distributions. *Technometrics* 34 (1):54–63. doi:10.2307/1269552.

Yashchin, E. 1993. Performance of Cusum control schemes for serially correlated observations. *Technometrics* 35 (1):37–52. doi:10.2307/1269288.

Yashchin, E. 1995. Likelihood ratio methods for monitoring parameters of a nested random effect model. *Journal of American Statistical Association* 90 (430):729–38. doi:10.2307/2291085.

Yashchin, E. 2018. Statistical monitoring of multi-stage processes. In *Frontiers in statistical quality control*, ed. S. Knoth and W. Schmid, vol. 12, 185–209. Berlin: Springer.