Optimal Design of Second-Order Linear Filters for Control Charting

Chang-Ho CHIN

School of Mechanical and Industrial Systems Engineering
Kyung Hee University
Yongin-si, Gyeonggi-do 446-701
Republic of Korea
(*chin@khu.ac.kr*)

Daniel W. APLEY

Department of Industrial Engineering
and Management Sciences
Northwestern University
Evanston, IL 60208-3119
(*apley@northwestern.edu*)

In many common control charting situations, the statistic to be charted can be viewed as the output of a linear filter applied to the sequence of process measurement data. In recent work that has generalized this concept, the charted statistic is the output of a general linear filter in impulse response form, and the filter is designed by selecting its impulse response coefficients to optimize its average run length performance. In this work, we restrict attention to the class of all second-order linear filters applied to the residuals of a time series model of the process data. We present an algorithm for optimizing the design of the second-order filter that is more computationally efficient and robust than the algorithm for optimizing the general linear filter. We demonstrate that the optimal second-order filter performs almost as well as the optimal general linear filter in many situations. Both methods share a number of interesting characteristics and are tuned to detect any distinct features of the process mean shift as it manifests itself in the residuals.

KEY WORDS: Autocorrelation; Control chart; Linear filtering; Markov chain method; Statistical process control; Time series.

1. INTRODUCTION

Many common control charting methods are based on linear filtering in the following sense. The sequence of statistics $\{y_t : t = 1, 2, 3, \ldots\}$ to be charted is calculated as the output of a linear filter applied to the sequence of process observations $\{x_t : t = 1, 2, 3, \ldots\}$. An alarm is sounded at observation number t if y_t falls outside a set of control limits. Note that a general (stationary) linear filter applied to a sequence $\{x_t\}$ is defined to be of the form $y_t = h_0 x_t + h_1 x_{t-1} + h_2 x_{t-2} + \cdots$, where $\{h_j : j = 0, 1, 2, \ldots\}$ are weighting coefficients (Box et al. 1994, chap. 1). In other words, the output of a linear filter is simply a weighted sum of past observations. If B denotes the time-series backshift operator, then the linear filter can be written as $y_t = H(B)x_t$, where $H(B) = h_0 + h_1 B + h_2 B^2 + \cdots$. A linear filter $H(B)$ expressed in this manner is said to be in impulse response form, and $\{h_j : j = 0, 1, 2, \ldots\}$ are called the impulse response coefficients. Linear filters are often expressed in transfer function form, such as (1). All transfer function forms of stationary linear filters have an equivalent impulse response form. (See Box et al. 1994, chap. 1, for an introductory discussion of linear filters and how to convert between impulse response and transfer function forms.)

A classic example of control charts based on linear filters is the exponentially weighted moving average (EWMA) control chart of Roberts (1959), in which y_t is an EWMA of x_t. The Shewhart individual chart is a trivial case with y_t equal to x_t. When x_t is an autocorrelated process, EWMA charts and Shewhart individual charts on the residuals of an autoregressive moving average (ARMA) model of the process (see, e.g., Montgomery and Mastrangelo 1991; Lu and Reynolds 1999a) constitute two more examples. This is because the residuals themselves can be viewed as the output of a linear filter applied to x_t.

More recent examples, in which the linear filter has a more complex structure than an EWMA, include the ARMA chart of Jiang, Tsui, and Woodall (2000) and Jiang (2001) and the Proportional Integral Derivative (PID) chart of Jiang, Wu, Tsung, Nair, and Tsui (2002). A more complex filter structure, with more filter design parameters, creates the potential for better control chart performance, especially when the process data are autocorrelated. It may be difficult to take advantage of this potential, however, because of difficulty in properly selecting the design parameters. The only available guidelines are heuristic and rather anecdotal. An ARMA or PID chart that is not optimized may perform worse than a well-designed EWMA.

Recently, Apley and Chin (2006) proposed a compete generalization of the concept of a control chart based on linear filtering. They considered a control chart statistic of the form $y_t = H(B)x_t$, where $H(B)$ is a general linear filter (GLF) in impulse response form. They treated this as an optimal filter design problem and developed a method for finding the filter impulse response coefficients, $\{h_j : j = 0, 1, 2, \ldots\}$, that minimize the out-of-control average run length (ARL) for a specified mean shift of interest, under the constraint that the in-control ARL equals some desired value. They demonstrated that for step mean shifts in independently identically distributed (iid) data, the optimal GLF (OGLF) coincides with a simple EWMA. For many autocorrelated processes, however, the OGLF has an intricate structure and can achieve much better ARL performance than an optimized EWMA. We note that Apley and Chin (2006) directly optimized the design of a GLF of the form $y_t = H(B)e_t$, where e_t denotes the residuals of an ARMA process model (see Sec. 2). There is no loss of generality in optimizing a GLF applied to e_t versus one applied to x_t, and vice versa, if the ARMA model is assumed to be stable and invertible.

TECHNOMETRICS, AUGUST 2006, VOL. 48, NO. 3
DOI 10.1198/004017006000000020

One disadvantage of the method of Apley and Chin (2006) is that calculating the ARL for a GLF is so complex that certain approximations and Monte Carlo simulations are required in the GLF optimization algorithm. Moreover, the GLF can be somewhat cumbersome to implement, because it requires storage of the entire set of impulse response coefficients (up to a suitably large truncation time, after which the coefficients are essentially 0). To avoid these drawbacks, we propose as a control chart statistic a second-order linear filter (SLF) of the form

$$y_t = \gamma\left[\frac{1-\beta B}{1-\alpha_1 B-\alpha_2 B^2}\right]e_t, \tag{1}$$

where $\alpha_1, \alpha_2, \beta$, and γ are the SLF design parameters to be determined. We include the scaling constant γ, because we use normalized control limits ± 1. The filter in (1) is a special case of the GLF considered by Apley and Chin (2006), with the restrictions that its transfer function form is second order, and it is applied to e_t instead of x_t.

Following Apley and Chin (2006), we focus on optimizing the design of the filter. Specifically, we develop an approach for selecting the SLF parameters $\alpha_1, \alpha_2, \beta$, and γ to minimize the out-of-control ARL under the constraint that the in-control ARL equals some desired value. In Section 2 and the Appendix we describe our approach for calculating the ARL of the SLF and its gradient with respect to the filter design parameters, which is needed in the optimization algorithm.

Our focus on filter design optimization is one aspect that distinguishes this work from the work on ARMA and PID charts. The heuristic design procedures suggested by Jiang et al. (2000, 2002) for the ARMA and PID charts are somewhat ambiguous and may result in control charts that perform far from optimally. We demonstrate this in Section 3.4 using the same vibration data example considered by Jiang et al. (2000, 2002).

Another difference between this work and the ARMA chart of Jiang et al. (2000) is that our SLF is applied to the residuals, whereas the ARMA chart is applied to the original data x_t. Applying the SLF to the residuals has two advantages. First, for reasons that become apparent in the next section, it allows a more computationally feasible approach for calculating the ARL. This is important when optimizing the performance of the SLF. The approach is applicable for any ARMA process, regardless of the model order. Second, it appears that applying the SLF to the residuals results in better ARL performance than applying the SLF to the original data, evidence of which we present in Section 3. Indeed, for many of the examples that we consider in Section 3, the performance of our optimized SLF is almost equal to that of the most general linear filter optimized by Apley and Chin (2006).

2. AVERAGE RUN LENGTH CALCULATION AND FILTER OPTIMIZATION STRATEGY

Throughout this article, we assume that x_t follows an ARMA process model of the form $x_t = \Phi^{-1}(B)\Theta(B)a_t + \mu_t$, where μ_t represents the deterministic process mean, t is a time index, a_t is an iid Gaussian process with mean 0 and variance σ^2, and $\Phi(B) = (1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)$ and $\Theta(B) = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)$ are the AR and MA polynomials of order p and q. The model residuals (i.e., the one-step-ahead prediction errors) are generated by the linear filtering operation (Apley and Shi 1999),

$$e_t = \frac{\Phi(B)}{\Theta(B)}x_t = \frac{\Phi(B)}{\Theta(B)}\left[\frac{\Theta(B)}{\Phi(B)}a_t + \mu_t\right] = a_t + \tilde{\mu}_t,$$

where $\tilde{\mu}_t = \Theta^{-1}(B)\Phi(B)\mu_t$ is a filtered version of the deterministic mean shift μ_t. The residuals are an independent sequence of Gaussian random variables with variance σ^2 and time-varying mean $\tilde{\mu}_t$. Apley and Shi (1999) referred to $\tilde{\mu}_t$ as the fault signature.

The objective is to find the SLF parameters that minimize the out-of-control ARL for a specified mean shift μ_t (e.g., a step shift of size μ, represented by $\mu_t = 0$ for $t < \tau$ and $\mu_t = \mu$ for $t \geq \tau$), while simultaneously constraining the in-control ARL to some desired value. To accomplish this, we express the ARL as a function of the filter parameters using the following variation of the Markov chain approach of Brook and Evans (1972). Define the vector $\mathbf{V}_t = (y_t, z_t)^\top$, where $z_t = \alpha_2 y_{t-1} - \gamma\beta e_t$, and note that $\mathbf{V}_t$ can be written as

$$\begin{aligned}\mathbf{V}_t &= \begin{bmatrix} y_t \\ z_t \end{bmatrix} = \begin{bmatrix} \alpha_1 & 1 \\ \alpha_2 & 0 \end{bmatrix}\begin{bmatrix} y_{t-1} \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} \gamma \\ -\gamma\beta \end{bmatrix} e_t \\ &= \mathbf{D}\mathbf{V}_{t-1} + \mathbf{W}e_t, \end{aligned} \tag{2}$$

where $\mathbf{D} = \left[\begin{smallmatrix} \alpha_1 & 1 \\ \alpha_2 & 0 \end{smallmatrix}\right]$ and $\mathbf{W} = \left[\begin{smallmatrix} \gamma \\ -\gamma\beta \end{smallmatrix}\right]$.

Because $\mathbf{V}_t$ is a two-dimensional vector Markov process, two-dimensional Markov chain methods similar to those used by Runger and Prabhu (1996), VanBrackle and Reynolds (1997), and Jiang (2001) can be used to calculate the ARL of the control chart on y_t. Readers familiar with automatic control theory might recognize (2) as the observable canonical form (Åström and Wittenmark 1990) of the filter. Although Jiang (2001) suggested using a three-dimensional vector Markov chain representation for such a process, invoking the canonical form of the filter yields the two-dimensional representation of (2). The reduction in dimensionality substantially reduces the computational expense involved in calculating the ARL. It also eliminates the need for the Monte Carlo simulation used by Apley and Chin (2006) for optimizing a GLF. The result is that the algorithm for optimizing the SLF is much more computationally efficient than the algorithm for optimizing the GLF.

The two-dimensional Markov chain approach used by Runger and Prabhu (1996), VanBrackle and Reynolds (1997), and Jiang (2001) can be applied to the present situation to calculate the ARL, as described in the Appendix. The strategy for optimizing the vector of SLF parameters $\boldsymbol{\zeta} = [\alpha_1 \ \alpha_2 \ \beta \ \gamma]^\top$ is summarized in the remainder of this section. As inputs to the optimization routine, the user specifies the ARMA process model, a mean shift that is of particular interest (the type as well as the magnitude), and a desired in-control ARL. The optimization algorithm then finds the filter parameters that minimize the out-of-control ARL for the specified mean shift, while providing the desired in-control ARL. The efficiency of the optimization routine is substantially improved by incorporating the gradient $\partial ARL/\partial\boldsymbol{\zeta}$, an expression for which we derive in the Appendix.

Using (A.2) and (A.3) of the Appendix for the ARL and its gradient, we have coded in MATLAB a straightforward gradient-based algorithm for optimizing the SLF parameters,

which is available on request. The Appendix provides some additional details on the algorithm, and further details have been given by Chin (2004). Note that numerical evaluation of $\partial ARL/\partial \boldsymbol{\zeta}$ in (A.3) involves roughly the same computational expense as evaluation of the ARL in (A.2).

3. DISCUSSION AND EXAMPLES

3.1 Comparison With the Optimal EWMA and the OGLF

In this section we compare the optimal SLF (OSLF) with the OGLF of Apley and Chin (2006) and with an optimized residual-based EWMA. The parameters of all three charts are optimized to minimize the out-of-control ARL for a shift occurring at initial time $\tau = 1$, while constraining the in-control ARL to equal 500. Although we must specify a shift time of occurrence, τ, to implement the optimization algorithm, we demonstrate in Section 3.5 that the ARL of an OSLF optimized for $\tau = 1$ is nearly independent of when the shift actually occurs. In other words, if we optimize the OSLF for zero-state ARL performance (i.e., $\tau = 1$), then we can be assured that it is nearly optimal for steady-state ARL performance as well.

The residual-based EWMA is defined as

$$y_t = (1 - \lambda)y_{t-1} + ge_t,$$

where $0 < \lambda \leq 1$ is the EWMA parameter and g is a scaling constant. The chart signals when the EWMA statistic y_t falls outside the control limits ± 1. Although the last term in the EWMA equation is usually written as λe_t, we use the additional constant g to account for the fact that our control limits are fixed at ± 1. Note that the EWMA can be written as a first-order linear filter $y_t = H(B)e_t$, where $H(B) = (1 - (1 - \lambda)B)^{-1}g$, which has impulse response coefficients $h_j = g(1 - \lambda)^j$. Consequently, the optimal EWMA can be viewed as a more restrictive counterpart of the OSLF, whereas OGLF can be viewed as a more general counterpart.

The performance of all three charts depends heavily on the form and magnitude of the residual mean and on the ARMA model describing the process. Because of this, we compare performance for the same broad combination of scenarios that Apley and Chin (2006) considered, which are represented by the 28 examples listed in Table 1. The process models are all ARMA(1, 1) models of the form $x_t - \phi x_{t-1} = a_t - \theta a_{t-1}$, which includes their special cases of first-order AR and iid. Without loss of generality, we assume that $\sigma = 1$ for the remainder of the article. We also consider three different types of mean shifts—step, spike, and sinusoidal—and a range of mean shift sizes depending on the specific example. The step mean shift was defined in the previous section, and the spike mean shift is defined as $\mu_1 = \mu$ and $\mu_t = 0$ for $t \neq 1$. The sinusoidal shifts are denoted by S_1–S_4 in Table 1. S_1, S_2, and S_3 are sinusoidal functions with amplitude .75 and periods of two, four, and eight observations. S_4 has amplitude 1.5 and a period of eight observations. The phase of each sinusoid is such that it achieves its maximum value on the initial observation.

Table 1 lists the out-of-control ARL values for all three charts for the 28 examples. All ARL values listed are zero-state values (corresponding steady-state ARL values are discussed in Sec. 3.5), and the in-control ARL was 500 in all cases. Although the Markov chain method was used to optimize the EWMA and the SLF, all ARL values shown in Table 1 were from Monte Carlo simulation with 250,000 replications. The standard errors of the ARL estimates are shown in parentheses. The optimized parameters for the EWMA and SLF are also shown. In the subsequent discussion, when of interest, we give the impulse response coefficients for the OGLFs. The OGLF impulse response coefficients for all 28 examples have been given by Apley and Chin (2006).

As shown by Apley and Chin (2006), the OGLF reduces to the simple-structured EWMA for the case of step mean shifts in iid processes (Examples 1–4). Consequently, because the SLF is contained within the class of GLFs, the OSLF also reduces to an EWMA. This is evident from $\alpha_2 = \beta = 0$ in Table 1 for Examples 1–4. Note that the optimal value of the EWMA parameter ($\lambda = 1 - \alpha_1$) becomes larger as the size of the mean shift increases, which is well known (Lucas and Saccucci 1990). It should not be surprising that for step shifts in iid data the OGLF and OSLF reduce to an EWMA. It is well known that a properly designed EWMA can approximately match the performance of any (two-sided) CUSUM, and Moustakides (1986) has shown that (one-sided) CUSUMs are optimal in a sense similar to ours over the class of all possible control charting procedures, linear or nonlinear.

In many of the examples listed in Table 1, substantial performance improvement can be achieved by increasing the complexity of the filter from the first-order EWMA to the SLF and the GLF. In most of the examples where the EWMA and OGLF performance differs significantly, the OSLF also performs much better than the optimal EWMA and almost as good as the OGLF. The exception to this is Example 28, for which the OSLF performs only slightly better than the optimal EWMA and substantially worse than the OGLF. We provide an explanation for this in Section 3.3.

Consider Example 8, which is a step mean shift of magnitude 4σ in an AR(1) process with $\phi = .9$. Note that the variance of x_t is $\sigma_x^2 = (1 - \phi^2)^{-1}\sigma^2$, and a mean shift of 4σ translates to only $1.74\sigma_x$. As illustrated in Figure 1(a), the residual mean in this case experiences a pronounced initial spike before dropping down to a much smaller steady-state value. In situations like this, Lin and Adams (1996) and Lu and Reynolds (1999b) recommended using a combined Shewhart–EWMA scheme. (See Lucas and Saccucci 1990 and Reynolds and Stoumbos 2001 for more discussion of combined Shewhart–EWMA charts.) The shapes of the OSLF and OGLF impulse response functions for Example 8 shown in Figure 1(b) indicate that they closely correspond to a combined Shewhart–EWMA scheme. Note that the plots in Figure 1(b) were obtained by expressing both the OGLF and the OSLF in their impulse response forms, $y_t = H(B)e_t$, with $H(B) = h_0 + h_1B + h_2B^2 + \cdots$. Also note that the OSLF and OGLF are almost identical, in the sense that their impulse response coefficients nearly coincide.

The connection to a combined Shewhart–EWMA scheme becomes more apparent if we write the OSLF in the following form, using the Example 8 parameters from Table 1:

$$y_t = .298\left[\frac{1 - .847B}{1 - .863B - .105B^2}\right]e_t$$

Table 1. ARL Comparison for the OSLF, the Optimal EWMA, and the OGLF

	Time series model		Shift		OGLF	Optimal EWMA			OSLF				
No.	ϕ	θ	Type	Size μ	ARL	λ	g	ARL	α_1	α_2	β	γ	ARL
1	0	0	Step	.5	28.82 (.03)	.047	.1167	28.82 (.03)	.953	.000	.000	.1167	28.82 (.03)
2				1.5	5.45 (.01)	.242	.2179	5.45 (.01)	.758	.000	.000	.2179	5.45 (.01)
3				3	1.86 (.00)	.676	.3067	1.86 (.00)	.324	.000	.000	.3067	1.86 (.00)
4				4	1.21 (.00)	.887	.3216	1.21 (.00)	.113	.000	.000	.3216	1.21 (.00)
5	.9	0	Step	.5	355.31 (.57)	.002	.0527	355.31 (.57)	.998	.000	.000	.0527	355.31 (.57)
6				1.5	130.64 (.18)	.007	.0654	130.64 (.18)	.993	.000	.000	.0654	130.64 (.18)
7				3	46.91 (.10)	.021	.0887	49.43 (.07)	.863	.105	.784	.2754	47.26 (.10)
8				4	13.72 (.06)	.038	.1080	29.78 (.05)	.863	.105	.847	.2983	13.72 (.06)
9	.9	0	Spike	.5	495.39 (.98)	1.000	.3236	497.12 (1.00)	−.070	.046	.869	.2368	496.83 (1.00)
10				1.5	422.01 (.98)	1.000	.3236	454.46 (.99)	−.071	.037	.872	.2364	427.08 (.98)
11				3	82.72 (.54)	1.000	.3236	177.83 (.76)	−.103	.001	.844	.2360	85.12 (.55)
12				4	6.72 (.14)	1.000	.3236	28.70 (.32)	−.069	.035	.872	.2367	7.12 (.15)
13	0	0	Sinusoid	S_1	15.79 (.02)	1.000	.3236	103.6 (.51)	−.558	.322	.326	.1506	15.79 (.02)
14				S_2	30.69 (.04)	1.000	.3236	170.8 (.81)	−.026	−.903	−.243	.1494	30.69 (.04)
15				S_3	32.90 (.04)	.608	.2986	137.6 (.57)	1.160	−.716	−1.208	.0849	43.30 (.08)
16				S_4	10.61 (.01)	.616	.2997	26.31 (.05)	1.024	−.636	−1.070	.1068	11.46 (.01)
17	.9	−.9	Step	.5	447.66 (.75)	.002	.0527	447.66 (.75)	.998	.000	.000	.0527	447.66 (.75)
18				1.5	139.26 (.54)	.003	.0557	255.72 (.39)	−.924	.007	−.039	.1399	163.10 (.71)
19				2	41.54 (.36)	.004	.0584	194.09 (.28)	−.924	.007	−.039	.1399	43.31 (.37)
20				3	3.12 (.03)	1.000	.3236	76.23 (.49)	−.861	−.045	−.084	.2051	3.21 (.04)
21	.9	.5	Step	.5	205.04 (.30)	.004	.0584	205.58 (.30)	.996	.000	.000	.0584	205.58 (.30)
22				1.5	50.28 (.07)	.021	.0887	50.28 (.07)	.979	.000	.000	.0887	50.28 (.07)
23				3	10.77 (.03)	.120	.1662	10.80 (.03)	.879	.000	−.020	.1639	10.77 (.03)
24				4	2.74 (.01)	.304	.2374	2.88 (.01)	.696	.000	.000	.2374	2.88 (.01)
25	.9	.5	Spike	.5	497.47 (.99)	1.000	.3236	497.61 (.99)	−.238	−.001	−.038	.3172	497.47 (1.00)
26				1.5	461.86 (.99)	1.000	.3236	469.74 (.99)	−.222	−.005	−.163	.3231	469.23 (.99)
27				3	208.77 (.80)	1.000	.3236	259.67 (.87)	−.220	−.006	−.185	.3234	259.77 (.88)
28				4	50.75 (.41)	1.000	.3236	86.10 (.56)	−.230	−.004	−.156	.3227	83.72 (.55)

NOTE: The simulation standard errors are shown in parentheses.

$$= \left[\frac{.264}{1+.108B}\right]e_t + \left[\frac{.034}{1-.971B}\right]e_t$$

$$\cong .264e_t + \frac{.034}{1-.971B}e_t.$$

The first term by itself represents a scaled version of a Shewhart individual chart on the residuals. The second term by itself represents a scaled version of an EWMA with a small value $\lambda = 1 - .971 = .029$ for the EWMA parameter. Consequently, the OSLF and OGLF for Example 8 are essentially a weighted combination of a Shewhart individual chart and an EWMA chart, as can be seen in Figure 1(b). Whereas the typical combined Shewhart–EWMA scheme charts the two statistics separately but simultaneously, the OSLF combines them

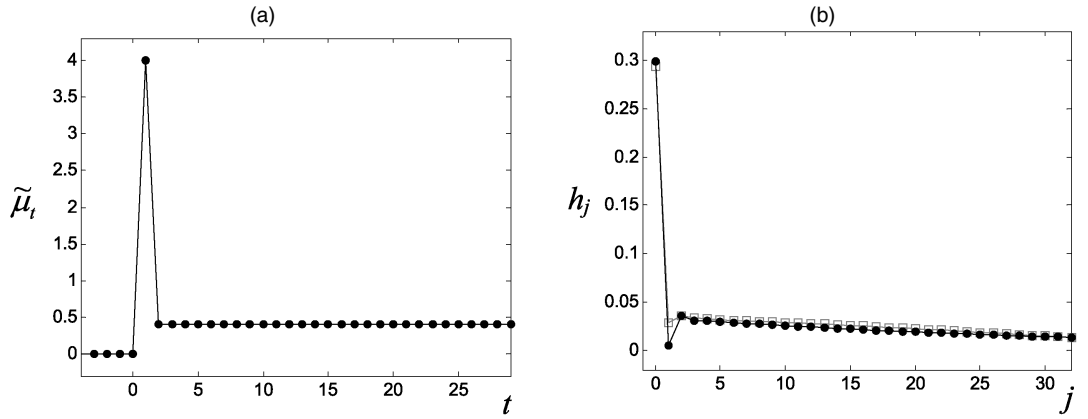

Figure 1. Residual Mean (a) and Impulse Responses of the OSLF (—●—) and the OGLF (—□—) (b) for Example 8.

together into a single statistic y_t. Despite this difference, we would expect the two charts to behave similarly; the Shewhart component is effective in detecting the initial spike in the residuals, and the EWMA component with small λ is effective in detecting the small but sustained steady-state shift in the residual mean. One attractive feature of the OSLF is that the relative weighting of the two components is selected optimally, to minimize the ARL.

The OSLFs in Examples 5–7 can be viewed similar to combined Shewhart–EWMA charts, where the relative weighting of the Shewhart component decreases as the size of the mean shift decreases. When the mean shift size decreases, so does the prominence of the initial spike in the residuals, and one must rely more heavily on the EWMA component to detect the small but sustained shift in the residual mean.

The OSLF and the OGLF are also quite similar to each other for the AR(1) processes with $\phi = .9$ and spike mean shift (Examples 9–12 in Table 1), and both outperform the optimal EWMA for large mean shifts. Figure 2(a) shows the residual mean for Example 12, and Figure 2(b) shows the corresponding impulse response coefficients for the OSLF and the OGLF. The reason why the OSLF and OGLF outperform the optimal EWMA in this case is apparent from Figure 2. The residual mean oscillates above and below 0 on the first two observations after the shift. Both the OSLF and the OGLF are tuned to detect this oscillation, in the sense that their impulse response coefficients also oscillate. A more precise explanation of this phenomenon is provided in Section 3.2.

Figures 3 and 4 show the residual mean and the OGLF and OSLF impulse responses for Examples 13 and 15 (sinusoidal mean shifts in iid processes). Note that the sinusoidal mean in Example 13, with a period of two observations, is characteristic of mixture distributions in which successive observations alternate between two different distributions. In both examples, the OSLF impulse responses are damped (i.e., exponentially decaying) sinusoids with period equal to the period of the sinusoidal shift. We can view both the OSLF and the OGLF as being tuned to detect the sinusoidal dynamics of the residual mean. The first-order EWMA obviously does not have a structure that is sufficiently flexible to be thus tuned, and consequently, its ARL is much larger.

3.2 Applying an SLF to e_t versus x_t

Let $H_e(B)e_t$ denote the OGLF applied to e_t, and suppose that we also optimize a GLF applied to x_t, denoted by $H_x(B)x_t$. Because $e_t = \Theta^{-1}(B)\Phi(B)x_t$, if the ARMA process model is stable and invertible, then it must be the case that the two optimized GLFs satisfy the relationship $H_x(B) = H_e(B)\Theta^{-1}(B) \times \Phi(B)$ or, equivalently, $H_e(B) = H_x(B)\Phi^{-1}(B)\Theta(B)$. The performance of the two charts is identical, and there is no loss of

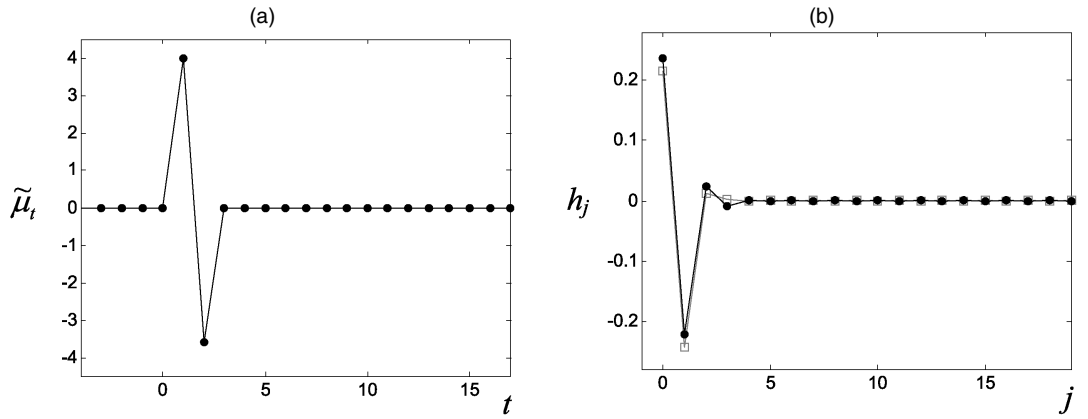

Figure 2. Residual Mean (a) and Impulse Responses of the OSLF (—●—) and the OGLF (—□—) (b) for Example 12.

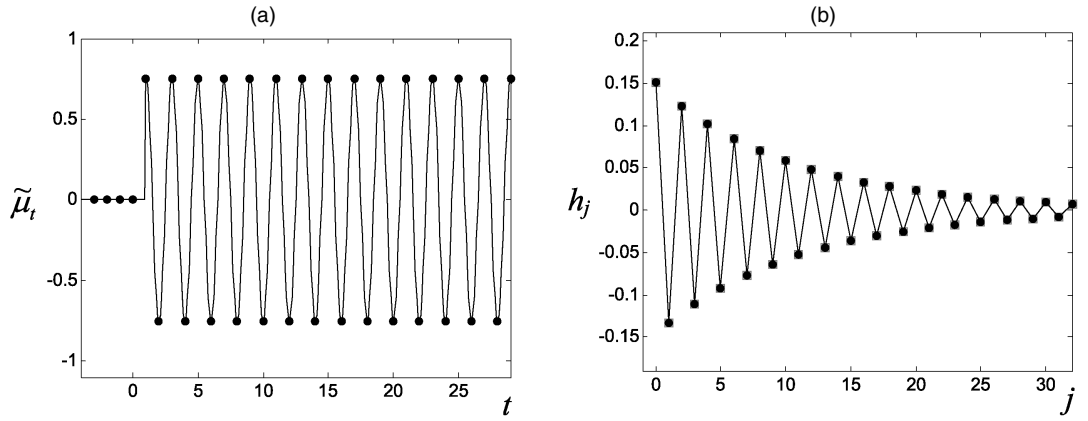

Figure 3. Residual Mean (a) and Impulse Responses of the OSLF (—●—) and the OGLF (—□—) (b) for Example 13.

generality in optimizing a GLF applied to e_t, versus one applied to x_t.

The same cannot be said about the SLFs. If we optimize an SLF applied to e_t, then its performance may be very different from that of an optimized SLF applied to x_t. The results in the previous section provide some evidence that better performance can be achieved by applying the SLF to the residuals. Specifically, we see that the OSLFs are almost identical to the OGLFs for many examples [e.g., those shown in Figs. 1(b), 2(b), and 3(b)], and the OGLF is the optimal linear filter of any order, applied to either e_t or x_t.

Another advantage of applying the SLF to the residuals is that the ARL computation is more tractable. Because the residuals are a sequence of independent random variables, we can represent an SLF applied to the residuals through the two-dimensional vector Markov process described in Section 2. In contrast, it can be shown that an SLF applied to an ARMA(p, q) process x_t requires a vector Markov chain representation of dimension $\max\{p+2, q+2\}$.

3.3 Second-Order versus Higher-Order Filters

The similarity of the OGLF and OSLF in many of the examples also implies that it is often unnecessary to consider filters of order higher than 2. One explanation for this relates to the mechanisms by which the OGLF or OSLF may achieve substantially better performance than the optimal EWMA. We can write each of the control chart statistics in terms of their filter impulse response coefficients through

$$y_t = \sum_{j=0}^{t-1} h_j e_{t-j} = \sum_{j=0}^{t-1} h_j \tilde{\mu}_{t-j} + \sum_{j=0}^{t-1} h_j a_{t-j}.$$

Thus, at any time t, the control chart statistic y_t is normally distributed with mean $\sum_{j=0}^{t-1} h_j \tilde{\mu}_{t-j}$ and variance $\sigma^2 \sum_{j=0}^{t-1} h_j^2$. Comparing the residual mean functions and the OGLF impulse response coefficients shown in Figures 1–4 shows that the OGLF is tuned so that its impulse response function is highly "correlated" with the residual mean, in the sense that the mean $\sum_{j=0}^{t-1} h_j \tilde{\mu}_{t-j}$ of y_t is large in magnitude after a mean shift.

Although the structure of the simple first-order EWMA filter is not sufficiently flexible to allow it to be tuned to correlate closely with a complex residual mean function, the increased flexibility of the second-order filter evidently does allow this. Comparing the OGLFs and OSLFs in Figures 1–4 clearly shows that the OSLF is capable of capturing the dynamics of a variety of impulse response functions.

One mild exception to this occurs in Example 15 (Fig. 4), in which the OSLF impulse response is a damped sinusoid of period eight. Although this approximates the general characteristics of the OGLF, the differences are not negligible, and the

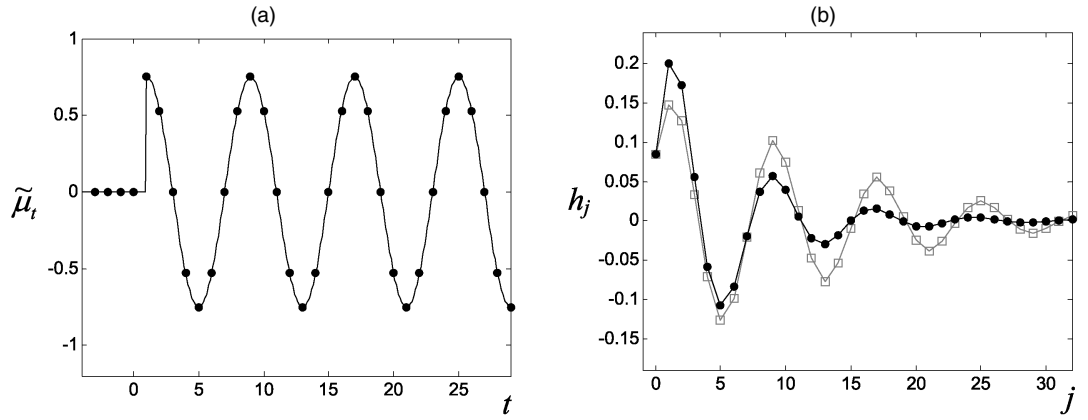

Figure 4. Residual Mean (a) and Impulse Responses of the OSLF (—●—) and the OGLF (—□—) (b) for Example 15.

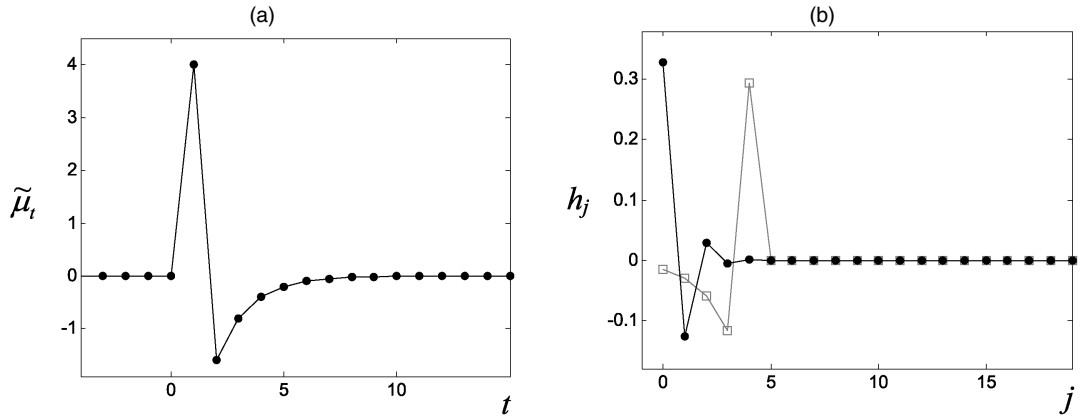

Figure 5. Residual Mean (a) and Impulse Responses of the OSLF (—●—) and the OGLF (—□—) (b) for Example 28.

OGLF performs substantially better than the OSLF. The ARLs for this example are 32.9 for the OGLF and 43.3 for the OSLF. Example 28, the residual mean and filter impulse response coefficients for which are shown in Figure 5, is a more extreme case. The OSLF cannot adequately approximate the OGLF, which closely mimics the oddly shaped residual mean. The result is that the ARL for the OSLF is almost 70% larger than the ARL for the OGLF (83.7 vs. 50.8).

3.4 Comparison With the PID Chart

The PID chart statistic of Jiang et al. (2002) can be written as $y_t = [1 - (1 - k_P - k_D - k_I)B - (k_P + 2k_D)B^2 + k_D B^3]^{-1} \times [1 - B]x_t$, where k_P, k_D, and k_I are design parameters. We can view this as a third-order filter on the original data x_t, with the restriction that the numerator polynomial of the filter is the differencing operator $[1 - B]$. The PID chart signals when y_t falls outside a set of symmetric control limits $\pm L$. One difficulty with the PID chart is that it is not straightforward how to select the filter design parameters. For computational reasons, it would be quite difficult to optimize the chart using a Markov chain approach. If x_t is an ARMA(p, q) process, then this would require a vector Markov chain representation of dimension $\max\{p+3, q+2\}$. For comparison purposes, however, we consider the same example of a mechanical vibration system studied by Jiang et al. (2002) using the values for k_P, k_D, and k_I that they recommended for this example.

The mechanical vibration data, originally considered by Pandit and Wu (1983), were fitted by the ARMA(2, 1) model,

$$x_t - 1.439x_{t-1} + .600x_{t-2} = a_t + .519a_{t-1}, \qquad (3)$$

with $\sigma = 2.21$ ($\sigma_x = 9.13$). For notational simplicity, we henceforth assume that the data are scaled so that $\sigma = 1$ ($\sigma_x = 4.13$). We also ignore any modeling errors.

Table 2 compares the zero-state out-of-control ARLs for the OSLF and the PID charts with design parameters taken directly from table 1 of Jiang et al. (2002), who selected these chart parameters according to their heuristic design guidelines. More specifically, they chose each of the three sets of PID parameters to optimize a pair of capability indices for one of the mean shifts in Table 2, which, they argued, should translate to reasonably good ARL performance for that size mean shift. For purposes of comparison with the OSLF, the column labeled "PID*" lists the ARL for the most effective of the three PID charts for each mean shift. All ARLs were evaluated based on Monte Carlo simulation with 250,000 replicates. The out-of-control condition was simulated by generating data that follow the model (3) with a step mean shift of size μ added at the initial observation. The in-control ARL was 370 for all charts. Table 2 also includes the OGLF and a Shewhart individual chart on the residuals with control limits ± 3.

As in many of the previous examples, the OSLF and OGLF perform comparably. The OSLF and OGLF both perform substantially better than the best of the three PID charts for each

Table 2. ARLs of the OSLF, the OGLF, the Residual-Based Shewhart Chart, and the PID Charts for the Mechanical Vibration Example

Shift	OSLF						Residual-based Shewhart chart		PID $\{k_P, k_I, k_D\} = \{-.8, 0, 0\}$	PID $\{k_P, k_I, k_D\} = \{-.3, 1.8, 0\}$	PID $\{k_P, k_I, k_D\} = \{-.8, 0, .5\}$
$(\Delta = \mu/\sigma_x)$	α_1	α_2	β	γ	ARL	OGLF		PID*	$(L = 2.596)$	$(L = 2.978)$	$(L = 2.531)$
0					370	370	370	370	370	370	370
					(.68)	(.68)	(.74)	(.73)	(.73)	(.73)	(.72)
.5	.986	.000	.000	.0843	76.88	61.26	200.02	118.14	141.43	351.12	118.14
					(.10)	(.15)	(.56)	(.22)	(.27)	(.72)	(.22)
1	−.529	.000	.000	.2855	1.59	1.40	3.56	37.25	44.89	118.24	37.25
					(.03)	(.01)	(.06)	(.06)	(.08)	(.53)	(.06)
2	.000	.000	.000	.3334	1.00	1.00	1.00	1.00	11.57	1.00	10.94
					(.00)	(.00)	(.00)	(.00)	(.02)	(.00)	(.01)
3	.000	.000	.000	.3334	1.00	1.00	1.00	1.00	5.44	1.00	5.60
					(.00)	(.00)	(.00)	(.00)	(.01)	(.00)	(.00)

NOTE: The simulation standard errors are in parentheses. The PID* column lists the best of the three PID charts for each mean shift size.

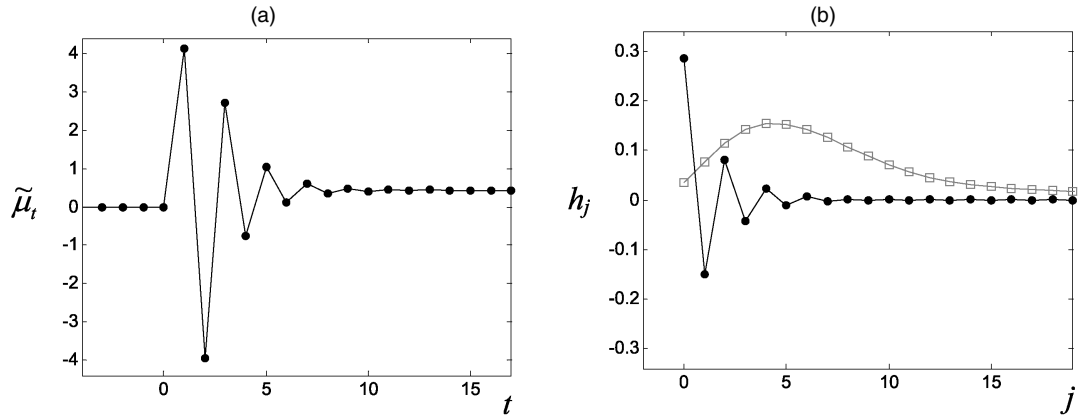

Figure 6. Residual Mean (a) and Impulse Responses of the OSLF (—●—) and the PID Chart (—□—) (b) for the Mechanical Vibration Example.

mean shift (except for the large shifts, which each chart detects immediately). The reason for this is clear from Figure 6. Figure 6(a) shows the residual mean, and Figure 6(b) shows the impulse response coefficients for the OSLF and the best of the three PID charts ($\{k_P, k_I, k_D\} = \{-.8, 0, .5\}$) for the $\Delta = 1$ case. To have a common basis for comparison, we plot the impulse response coefficients for the equivalent form of the PID chart filter when we view it as being applied to the residuals. In other words, we show the impulse response coefficients for the filter $y_t = H(B)e_t$ with $H(B) = [1 - (1 - k_P - k_D - k_I)B - (k_P + 2k_D)B^2 + k_D B^3]^{-1}[1 - B]\Phi^{-1}(B)\Theta(B)$. Clearly, the OSLF is tuned to detect the oscillatory nature of the residual mean, whereas the PID chart is not.

3.5 Robustness Properties

The design and analysis of the OSLF was based on a number of assumptions, including the following: (a) The shift time of occurrence τ coincides with the initial observation, as opposed to some later point in time (zero-state ARLs, as opposed to steady-state ARLs); (b) we are able to specify a single shift size of particular interest, for which the OSLF is optimized; and (c) for situations in which the process is autocorrelated, the ARMA model perfectly describes the process and there are no modeling errors. In this section we investigate the robustness of the OSLF to violations of these assumptions and discuss steps to ensure that the OSLF is effective at detecting a range of shift sizes.

First, consider the shift time of occurrence. Although the OSLF is optimized for $\tau = 1$, one is generally more interested in the steady-state ARL (i.e., the expected detection lag for shifts that occur after the control chart has been in operation for a while) than the zero-state ARL (the expected detection lag for $\tau = 1$). One might consider optimizing the steady-state ARL of the OSLF, rather than its zero-state ARL; however, this would be substantially more computationally expensive, as well as largely unnecessary. Apley and Chin (2006) demonstrated that the OGLF optimized for shifts occurring at the initial observation is almost equally effective at detecting shifts occurring in the steady state, and, as we demonstrate in the following paragraphs, the same is true for the OSLF.

Note that this is in sharp contrast to the CUSCORE charts of Box and Ramirez (1992) and Luceño (1999). Like the OSLF, the CUSCORE chart can be potentially very effective in detecting a dynamically varying residual mean, because it takes into consideration the dynamics of the fault signature. But the CUSCORE chart is tuned to detect signals occurring at a specific point in time and typically will have radically different zero-state and steady-state ARL performance (Shu, Apley, and Tsung 2002). For example, consider the version of CUSCORE chart for detecting mean shifts in autocorrelated processes that was proposed by Luceño (1999) and analyzed by Shu et al. (2002), Runger and Testik (2003), and Luceño (2004), for which the CUSCORE statistic is defined recursively through

$$S_t = \max\{0, S_{t-1} + \tilde{\mu}_t(e_t - \tilde{\mu}_t/2)\}, \qquad t = 1, 2, 3, \ldots.$$

The CUSCORE chart signals when S_t exceeds a threshold. Luceño (1999) considered two different versions of the CUSCORE chart. In the first version, the feared residual mean $\tilde{\mu}_t$ is taken to be that resulting from a mean shift occurring at $\tau = 1$. In the second version, $\tilde{\mu}_t$ is reinitialized every time the CUSCORE statistic is reset to 0.

Tables 3 and 4 compare the zero-state versus steady-state performance of the two CUSCORE charts (with and without reinitialization), the OSLF, and the optimal EWMA for Examples 15 and 20. For these examples the residual mean functions have pronounced dynamics, so that mismatches between the assumed and actual τ potentially will have the greatest effect. All charts were designed to have a zero-state in-control ARL of 500. The OSLFs and optimal EWMAs in Tables 3 and 4 are exactly those for Examples 15 and 20 of Table 1, optimized

Table 3. Comparison of Steady-State versus Zero-State ARLs for Example 15

	$\mu = 0$ (no sinusoid)		$\mu = .75$ (sinusoid added)	
	Zero-state ARL	*Steady-state ARL*	*Zero-state ARL*	*Steady-state ARL*
CUSCORE without reinitialization	500.9 (.99)	491.7 (1.09)	23.22 (.06)	$\approx\infty$
CUSCORE with reinitialization	500.9 (.98)	490.8 (1.12)	37.30 (.14)	43.12 (.15)
OSLF	500.1 (.98)	496.1 (1.14)	43.30 (.08)	41.78 (.15)
Optimal EWMA	500.3 (.95)	494.2 (1.21)	137.6 (.57)	136.6 (.61)

Table 4. Comparison of Steady-State versus Zero-State ARLs for Example 20

	$\mu = 0$ (no shift)		$\mu = 3$ (shift added)	
	Zero-state ARL	*Steady-state ARL*	*Zero-state ARL*	*Steady-state ARL*
CUSCORE without reinitialization	500.1 (1.02)	652.6 (1.24)	1.33 (.00)	98.2 (.39)
CUSCORE with reinitialization	498.8 (1.16)	497.9 (1.18)	4.16 (.07)	6.09 (.11)
OSLF	500.4 (.99)	495.8 (1.0)	3.21 (.04)	3.41 (.07)
Optimal EWMA	500.7 (.99)	500.2 (1.00)	76.23 (.49)	75.97 (.49)

Table 5. ARLs of the OSLFs for a Range of Shift Sizes for the Process Described by Equation (3)

Shift ($\Delta = \mu/\sigma_X$)	*OSLF* ($\Delta = .5$)	*OSLF* ($\Delta = 1$)	*OSLF* ($\Delta = 2$)	*OSLF* ($\Delta = 3$)
0	370 (1.10)	370 (1.17)	370 (1.16)	370 (1.16)
.5	77.03 (.16)	163.1 (.92)	199.2 (.56)	199.2 (.56)
1	32.67 (.05)	1.67 (.05)	3.56 (.06)	3.56 (.06)
2	14.07 (.02)	1.00 (0)	1.00 (0)	1.00 (0)
3	3.52 (.01)	1.00 (0)	1.00 (0)	1.00 (0)

under the assumption that $\tau = 1$. The control limits for the CUSCORE charts with and without reinitialization in Table 3 were 3.78 and 3.60. The control limits for the CUSCORE charts with and without reinitialization in Table 4 were 4.59 and 2.32. Monte Carlo simulation was used to evaluate the ARLs. For the steady-state ARLs, mean shifts were introduced at $\tau = 100$.

Not surprisingly, the CUSCORE chart without reinitialization is the most effective at detecting shifts occurring at $\tau = 1$ but the least effective at detecting shifts occurring in the steady state. The reason is that it is tuned to detect shifts occurring exclusively at $\tau = 1$ and loses nearly all of its power if the shift occurs in the steady state (its steady-state ARL is worse than that of the EWMA). For this reason, we do not recommend using the CUSCORE chart without reinitialization. Note that the out-of-control steady-state ARL for the CUSCORE chart without reinitialization is $\approx\infty$ in Table 3. The reason is that if the sinusoid is introduced at observation 100, then the feared signal and actual signal are exactly 180 degrees out of phase. Hence the term $\tilde{\mu}_t(e_t - \tilde{\mu}_t/2)$ tends to assume negative values, so that the CUSCORE statistic is reset to 0 on most observations.

We might view the CUSCORE with reinitialization as being tuned to detect mean shifts that have occurred immediately after the CUSCORE statistic was last reset to 0. Consequently, the differences between its zero-state and steady-state ARLs, while still substantial, are much less than for the CUSCORE without reinitialization. The differences between the zero-state and steady-state ARLs of the OSLF are even smaller, so that the OSLF performs better than the CUSCORE with reinitialization for shifts that occur in the steady state. We conclude that the OSLF is almost equally effective at detecting shifts occurring in the steady state, even though it was explicitly optimized for shifts occurring at $\tau = 1$. Apley and Chin (2006) attributed this to the fact that the OGLF and OSLF filters are moving functions of the residuals that move across the residuals one observation at a time. In this respect, the OSLF and OGLF look for mean shifts occurring at all points in time, not any one specific point in time.

Next, consider the shift size μ. The OSLF is optimized for one specific value of μ, whereas in practice we are generally interested in a range of shift sizes. In many situations, an OSLF optimized for one μ may perform poorly in detecting a much larger or much smaller μ. For example, the OSLF for detecting step shifts in iid data was previously shown to be quite similar to an EWMA, and it is well known that an EWMA optimized for a small (large) μ will perform far from optimal for a large (small) μ. In other situations, an OSLF optimized for one μ may perform reasonably well for a range of μ. Each column of Table 5 shows the ARLs over a range of μ for an OSLF optimized for one of the four shift sizes (.5, 1.0, 2.0, or 3.0) considered in the mechanical vibration system described by (3). The parameters of the OSLFs are as shown in Table 2.

If one is interested in a range of μ, then we recommend calculating the OSLF for a few different μ values that span the range of interest. If one of the OSLFs performs adequately over the range, then that OSLF can be used by itself. For example, the OSLF optimized for $\Delta = \mu/\sigma_x = 1.0$ in Table 5 also performs quite well for the larger shifts. Conversely, if no single OSLF performs well over the full range of μ, then we recommend using two or three OSLFs simultaneously, each one optimized for a different μ. This is similar to the practice of using a Shewhart chart in conjunction with a CUSUM (e.g., Lucas 1982) or, more generally, using multiple simultaneous CUSUMS, each optimized for a different μ (e.g., Sparks 2000). Apley and Chin (2006) discussed simple guidelines for designing multiple simultaneous OGLFs, which also apply to the design of multiple simultaneous OSLFs.

The final robustness issue that we consider is robustness to estimation errors in the ARMA model parameters. Tables 6 and 7 show the effects of ARMA parameter errors on the in-control and out-of-control ARLs for Examples 8 and 19. Example 19 was chosen because the residual mean oscillates with quite pronounced dynamics, due to the particular combination of ARMA parameters. Hence one might speculate that errors in the ARMA parameters would have a large effect on the ARL of the OSLF. The residual mean for Example 8 is shown in Figure 1. Tables 6 and 7 also show the robustness of the Shewhart individual chart and the optimal EWMA (OEWMA). As a reference, the in-control and out-of-control ARLs with no parameter errors (taken from Table 1) are shown in the leftmost column.

Table 6. Effects of ARMA Parameter Errors on the ARL of Various Charts for Example 8

Chart	$\hat{\phi} = \phi = .9$	$\hat{\phi} = .9, \phi = .85$	$\hat{\phi} = .9, \phi = .95$
OSLF	13.7 500	12.8 831	16.6 182
OEWMA	29.8 500	27.9 3,656	35.9 122
Shewhart	48.5 500	46.5 476	47.0 442

NOTE: The top numbers are the out-of-control ARLs for a shift of size $\mu = 4.0$, and the bottom numbers are the in-control ARLs.

Table 7. Effects of ARMA Parameter Errors on the ARL of Various Charts for Example 19

Chart	$\hat{\phi}=\phi=.9$ $\hat{\theta}=\theta=-.9$	$\hat{\phi}=.9,\phi=.85$ $\hat{\theta}=\theta=-.9$	$\hat{\phi}=.9,\phi=.95$ $\hat{\theta}=\theta=-.9$	$\hat{\phi}=\phi=.9$ $\hat{\theta}=-.9,\theta=-.85$	$\hat{\phi}=\phi=.9$ $\hat{\theta}=-.9,\theta=-.95$
OSLF	43.3	37.1	44.9	26.1	170
	500	417	571	158	4,613
OEWMA	194	202	137	197	193
	500	1,653	161	540	465
Shewhart	299	285	263	285	276
	500	477	442	469	468

NOTE: The top numbers are the out-of-control ARLs for a shift of size $\mu = 2.0$, and the bottom numbers are the in-control ARLs.

A few points are worth making. First, we cannot generalize that the OSLF is any more affected or any less affected by ARMA parameter errors than the OEWMA. Which chart is more affected depends entirely on the specifics of the example. For the AR(1) process of Example 8, the in-control ARL of the OEWMA is much more affected than that of the OSLF, whereas the out-of-control ARLs of the two charts are roughly equally affected. (The out-of-control ARLs for both charts are relatively robust for this example.) For the ARMA$(1, 1)$ process of Example 19, the OSLF is quite robust to errors in ϕ, whereas the OEWMA is affected rather severely. The exact opposite is true regarding errors in θ, for which the OEWMA is quite robust, whereas the OSLF is severely affected.

The second point is that the Shewhart individual chart is by far the most robust of any of the charts. This is consistent with the findings of Apley and Lee (2003). The third point is that even in the situations in Tables 6 and 7 in which the OSLF is strongly affected by parameter errors, it still performs much better than either the OEWMA or the Shewhart individual chart. For example, for $\theta = -.95$ (as opposed to $\hat{\theta} = -.90$) in Example 19, the out-of-control ARL of the OSLF increases dramatically, to 170.0 from a value of 43.3 with no parameter errors (see Table 7). This is still better than the corresponding out-of-control ARLs of the OEWMA (192.7) and the Shewhart chart (275.5), however. Moreover, the in-control ARL of the OSLF also increases to 4,613, which is desirable.

4. CONCLUSIONS

In this article we have proposed a control charting procedure based on a second-order linear filter applied to the residuals of an ARMA process model. We used a two-dimensional vector Markov chain method to calculate the ARL as a function of the filter parameters and derived an expression for the derivative of the ARL, which was used in a gradient-based algorithm for optimizing the filter parameters.

We have demonstrated through a number of examples that the OSLF can perform substantially better than an optimized EWMA and almost as good as the most general linear filter with optimized impulse response coefficients. One advantage of the OSLF over the OGLF is that it can be optimized using a more computationally efficient algorithm that avoids the need for Monte Carlo simulation. Another advantage is that the OSLF can be more easily implemented recursively using its transfer function form. The OGLF has no equivalent recursive implementation and must be implemented in impulse response form, which requires storage of the entire set of impulse response coefficients.

In situations where the OGLF performs substantially better than the OSLF (e.g., Examples 15 and 28), the OSLF still has some utility. As discussed by Apley and Chin (2006), the optimization algorithm for the OGLF can be sensitive to the initial guess for the design parameters. A reasonable strategy for optimizing the OGLF is to first find the OSLF (the optimization algorithm for which is more stable and robust), and then use the OSLF impulse response coefficients as the initial guess for the OGLF.

One drawback of the OSLF, as well as the OGLF, is that it cannot be designed without the use of a numerical optimization algorithm (available on request). In other words, there are no simple tables, graphs, or heuristic rules that would allow a practitioner to easily design the OSLF without the use of software. One could argue that this drawback is shared by all but the simplest of control charting methods, however. For example, for detecting step mean shifts of known magnitude in iid data, we have familiar tables for selecting the optimal parameters of EWMA and CUSUM charts (Lucas and Saccucci 1990; Montgomery 2005). But no such tables are available for autocorrelated data. The difficulty is that the optimal chart parameters depend strongly on numerous quantities, including the order and parameters of the ARMA model used to represent the autocorrelation, the desired in-control ARL, and the size and nature of the mean shift for which the chart is to be optimized. Even a simple EWMA is very difficult to optimize for autocorrelated data. To the best of our knowledge, if we restrict attention to methods that are sufficiently versatile to be effective in a variety of control charting situations with autocorrelated data, no methods can be optimally designed, or even approximately optimally designed, without the use of software. At the very least, even if the structure of the control chart is determined, software (e.g., Monte Carlo simulation or Markov chain algorithms) is needed to specify control limits that achieve a desired in-control ARL. The ineffectiveness of simple heuristics for guaranteeing even close to optimal designs was illustrated with the PID chart in Table 2.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation (grant DMI-0093580).

APPENDIX: SOME DETAILS ON CALCULATING THE AVERAGE RUN LENGTH AND ITS GRADIENT

In this appendix we provide some details of the approach for calculating the ARL and its gradient (derivative) with respect to the filter parameter vector $\boldsymbol{\zeta}$. The two-dimensional state

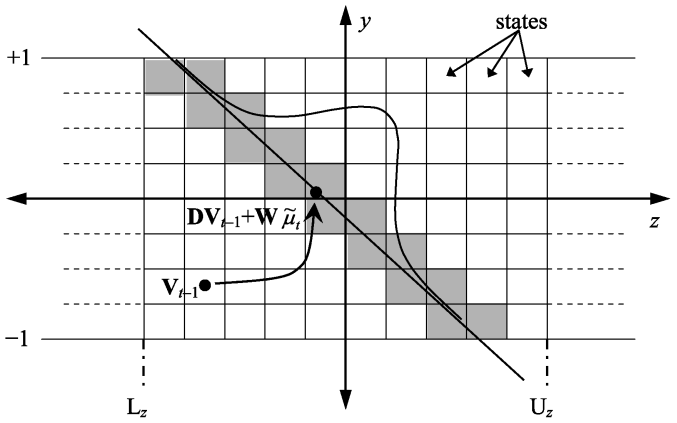

Figure A.1. Two-Dimensional State-Space Discretization in the Markov Chain Approach.

space for $\mathbf{V} = (y, z)^\top$ is discretized into a set of rectangles, as shown in Figure A.1. The range of values for y extends to the lower and upper control limits ± 1. Although the z-axis technically extends out to $\pm\infty$, we may truncate this by defining the upper and lower limits (L_z, U_z) sufficiently wide to ensure that z_t lies between the limits with high probability. Let N_z denote the number of discretized subintervals along the z-axis, and let N_y denote the number of discretized subintervals along the y-axis between ± 1. Therefore, the in-control region consists of $N = N_z \times N_y$ nonabsorbing states. The out-of-control region (y outside the ± 1 interval) is treated as a single absorbing state.

Let $\mathbf{Q}_t$ denote the $N \times N$ transition probability matrix for the nonabsorbing states at time t. The ith row, jth column element $(1 \le i, j \le N)$ of $\mathbf{Q}_t$, denoted by Q_t^{ij}, is defined as $Q_t^{ij} = \Pr\{\mathbf{V}_t \in R_j | V_{t-1} = \mathbf{r}_i\}$, where R_j is the rectangle for state j and $\mathbf{r}_i$ is the centroid of R_i.

Equation (2) implies that $\mathbf{V}_t|\mathbf{V}_{t-1}$ follows a degenerate bivariate normal distribution with mean $\mathbf{DV}_{t-1} + \mathbf{W}\tilde{\mu}_t$ and rank-1 covariance matrix $\mathbf{WW}^\top\sigma^2$. In other words, $\mathbf{V}_t|\mathbf{V}_{t-1}$ is distributed along a one-dimensional line in the two-dimensional state space, as illustrated in Figure A.1. Each Q_t^{ij} can be calculated as the area under the normal density curve for the segment of the distribution line that falls within rectangle R_j. (We provide a more detailed explanation at the end of this appendix.) If the distribution line does not pass through a particular rectangle, then the corresponding element of Q_t^{ij} is exactly 0. Therefore, although $\mathbf{Q}_t$ is an $N \times N$ matrix, each row of $\mathbf{Q}_t$ contains fewer than $\max\{2N_y, 2N_z\}$ nonzero elements. Because $\mathbf{Q}_t$ is a sparse matrix, the computational expense in calculating the ARL is decreased. Jiang (2001) discussed how to take advantage of this sparseness in more detail.

The ARL can be approximated as (Brook and Evans 1972)

$$ARL = \underline{\boldsymbol{\pi}}_0(\mathbf{I} + \mathbf{Q}_1 + \mathbf{Q}_1\mathbf{Q}_2 + \mathbf{Q}_1\mathbf{Q}_2\mathbf{Q}_3 + \cdots)\underline{\mathbf{1}}, \qquad (A.1)$$

where $\underline{\mathbf{1}}$ denotes a column vector of 1's and $\underline{\boldsymbol{\pi}}_0$ denotes the initial state probability vector. In all examples we have optimized the OSLF under the zero-state scenario, which is represented by setting all elements of $\underline{\boldsymbol{\pi}}_0$ equal to 0 except for the single element corresponding to the initial value $\{y_0 = 0, z_0 = 0\}$, which is set equal to 1. Because $\mathbf{Q}_t$ depends on t only through the time-varying mean of the residuals, $\mathbf{Q}_t$ approaches a steady-state value (denoted by $\mathbf{Q}$) as $\tilde{\mu}_t$ approaches a steady-state value. For sufficiently large m, we therefore have $\mathbf{Q} \cong \mathbf{Q}_m \cong \mathbf{Q}_{m+1} \cong \cdots$, and (A.1) becomes

$$ARL = \sum_{n=1}^{m-1} \mathbf{b}_n\underline{\mathbf{1}} + \mathbf{b}_m[\mathbf{I} - \mathbf{Q}]^{-1}\underline{\mathbf{1}}, \qquad (A.2)$$

where $\mathbf{b}_n = \underline{\boldsymbol{\pi}}_0 \prod_{l=1}^{n-1} \mathbf{Q}_l = \mathbf{b}_{n-1}\mathbf{Q}_{n-1}$ can be calculated recursively for $n = 2, 3, \ldots, m$ with $\mathbf{b}_1 = \underline{\boldsymbol{\pi}}_0$. Lu and Reynolds (1999a) have provided further discussion of this steady-state truncation in the Markov chain approach.

To calculate the gradient $\partial ARL/\partial\boldsymbol{\zeta}$, we differentiate (A.1) with respect to each element of $\boldsymbol{\zeta}$. If ζ_k denotes the kth element of $\boldsymbol{\zeta}$, then this gives

$$\begin{aligned}
&\frac{\partial ARL}{\partial \zeta_k} \\
&= \underline{\boldsymbol{\pi}}_0\left[\frac{\partial \mathbf{Q}_1}{\partial \zeta_k} + \left(\frac{\partial \mathbf{Q}_1}{\partial \zeta_k}\mathbf{Q}_2 + \mathbf{Q}_1\frac{\partial \mathbf{Q}_2}{\partial \zeta_k}\right)\right. \\
&\qquad \left. + \left(\frac{\partial \mathbf{Q}_1}{\partial \zeta_k}\mathbf{Q}_2\mathbf{Q}_3 + \mathbf{Q}_1\frac{\partial \mathbf{Q}_2}{\partial \zeta_k}\mathbf{Q}_3 + \mathbf{Q}_1\mathbf{Q}_2\frac{\partial \mathbf{Q}_3}{\partial \zeta_k}\right) + \cdots\right]\underline{\mathbf{1}} \\
&= \underline{\boldsymbol{\pi}}_0\left[\left(\frac{\partial \mathbf{Q}_1}{\partial \zeta_k} + \frac{\partial \mathbf{Q}_1}{\partial \zeta_k}\mathbf{Q}_2 + \frac{\partial \mathbf{Q}_1}{\partial \zeta_k}\mathbf{Q}_2\mathbf{Q}_3 + \cdots\right)\right. \\
&\qquad + \left(\mathbf{Q}_1\frac{\partial \mathbf{Q}_2}{\partial \zeta_k} + \mathbf{Q}_1\frac{\partial \mathbf{Q}_2}{\partial \zeta_k}\mathbf{Q}_3 + \mathbf{Q}_1\frac{\partial \mathbf{Q}_2}{\partial \zeta_k}\mathbf{Q}_3\mathbf{Q}_4 + \cdots\right) \\
&\qquad + \left(\mathbf{Q}_1\mathbf{Q}_2\frac{\partial \mathbf{Q}_3}{\partial \zeta_k} + \mathbf{Q}_1\mathbf{Q}_2\frac{\partial \mathbf{Q}_3}{\partial \zeta_k}\mathbf{Q}_4\right. \\
&\qquad\qquad \left.\left. + \mathbf{Q}_1\mathbf{Q}_2\frac{\partial \mathbf{Q}_3}{\partial \zeta_k}\mathbf{Q}_4\mathbf{Q}_5 + \cdots\right) + \cdots\right]\underline{\mathbf{1}} \\
&= \sum_{n=1}^{m-1} \mathbf{b}_n\frac{\partial \mathbf{Q}_n}{\partial \zeta_k}\mathbf{c}_n + \mathbf{b}_m[\mathbf{I} - \mathbf{Q}]^{-1}\frac{\partial \mathbf{Q}}{\partial \zeta_k}\mathbf{c}_m, \qquad (A.3)
\end{aligned}$$

where $\mathbf{c}_n = [\mathbf{I} + \mathbf{Q}_{n+1} + \mathbf{Q}_{n+1}\mathbf{Q}_{n+2} + \cdots]\underline{\mathbf{1}} = \underline{\mathbf{1}} + \mathbf{Q}_{n+1}\mathbf{c}_{n+1}$ can be calculated recursively for $n = m-1, m-2, \ldots, 1$ with initial condition $\mathbf{c}_m = [\mathbf{I} + \mathbf{Q} + \mathbf{QQ} + \cdots]\underline{\mathbf{1}} = [\mathbf{I} - \mathbf{Q}]^{-1}\underline{\mathbf{1}}$.

The matrices $\mathbf{Q}_t$ and their element-by-element derivative matrices $\partial\mathbf{Q}_t/\partial\zeta_k$ can be calculated as follows. The conditional distribution $\mathbf{V}_t|\mathbf{V}_{t-1} \sim \mathrm{N}_2(\mathbf{DV}_{t-1} + \mathbf{W}\tilde{\mu}_t, \mathbf{WW}^\top\sigma^2)$ derived from (2) is a degenerate bivariate normal distribution along a single line in two-dimensional space, as discussed earlier. Let $y_{j,1} \ge y_{j,2}$ denote the y coordinates of the points at which the distribution line enters and leaves the rectangle R_j for state j (Fig. A.2). The conditional distribution for $\mathbf{V}_t|\mathbf{V}_{t-1}$ implies that $y_t|\mathbf{V}_{t-1} \sim \mathrm{N}(\alpha_1 y_{t-1} + z_{t-1} + \gamma\tilde{\mu}_t, \gamma^2\sigma^2)$. The element Q_t^{ij} of the transition probability matrix, which is shown as the shaded area under the normal density curve in Figure A.2, is therefore given by

$$\begin{aligned}
Q_t^{ij} &= P[y_{j,2} \le y_t \le y_{j,1} | \mathbf{V}_{t-1} = \mathbf{r}_i] \\
&= F\left(\frac{y_{j,1} - (\alpha_1 r_{i,y} + r_{i,z} + \gamma\tilde{\mu}_t)}{\gamma\sigma}\right) \\
&\quad - F\left(\frac{y_{j,2} - (\alpha_1 r_{i,y} + r_{i,z} + \gamma\tilde{\mu}_t)}{\gamma\sigma}\right) \\
&= F(\tilde{y}_{j,1}) - F(\tilde{y}_{j,2}),
\end{aligned}$$

where $r_{i,y}$ ($r_{i,z}$) denotes the y (z) coordinate of $\mathbf{r}_i$, $\tilde{y}_{j,1}$ and $\tilde{y}_{j,2}$ denote the arguments in parentheses, and $F(\cdot)$ denotes the standard normal cumulative distribution function.

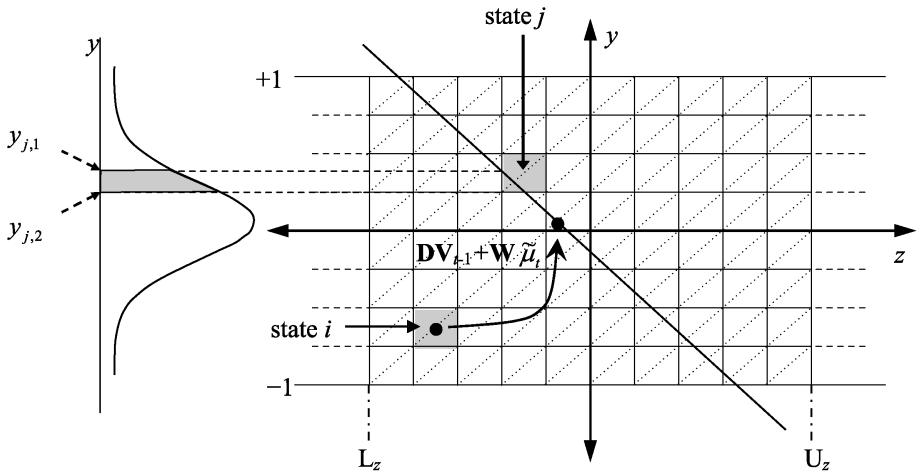

Figure A.2. Illustration of How Q_t^{ij} Is Calculated.

Based on this, the derivative of Q_t^{ij} with respect to each filter parameter is

$$\frac{\partial Q_t^{ij}}{\zeta_k} = f(\tilde{y}_{j,1})\frac{\partial \tilde{y}_{j,1}}{\partial \zeta_k} - f(\tilde{y}_{j,2})\frac{\partial \tilde{y}_{j,2}}{\partial \zeta_k},$$

where $f(\cdot)$ denotes the standard normal probability density function.

Although the discretization depicted in Figure A.1 is appropriate for calculating the ARL, it can be problematic when calculating the derivative of Q_t^{ij} that appears in the expression for the ARL gradient. The problem occurs when $\beta \approx 0$, in which case the distribution line is nearly vertical and may lie entirely within a single vertical column of rectangles [see (2)], without crossing over into a neighboring column. In this case the Markov chain approximation of the derivatives of Q_t^{ij} with respect to β and α_2 are exactly 0, so that the gradient-based optimization algorithm becomes trapped. Because of this, in our algorithm we modify the state-space discretization by partitioning each rectangle into two states through the diagonal lines shown in Figure A.2.

[Received February 2005. Revised December 2005.]

REFERENCES

Apley, D. W., and Chin, C. (2006), "An Optimal Filter Design Approach to Statistical Process Control," *Journal of Quality Technology*, in press.

Apley, D. W., and Lee, H. C. (2003), "Design of Exponentially Weighted Moving Average Control Charts for Autocorrelated Processes With Model Uncertainty," *Technometrics*, 45, 187–198.

Apley, D. W., and Shi, J. (1999), "The GLRT for Statistical Process Control of Autocorrelated Processes," *IIE Transactions*, 31, 1123–1134.

Åström, K. J., and Wittenmark, B. (1990), *Computer-Controlled Systems: Theory and Design* (2nd ed.), Englewood Cliffs, NJ: Prentice-Hall.

Box, G., Jenkins, G., and Reinsel, G. (1994), *Time Series Analysis, Forecasting, and Control* (3rd ed.), Englewood Cliffs, NJ: Prentice-Hall.

Box, G. E. P., and Ramirez, J. (1992), "Cumulative Score Charts," *Quality and Reliability Engineering International*, 8, 17–27.

Brook, D., and Evans, D. A. (1972), "An Approach to the Probability Distribution of CUSUM Run Lengths," *Biometrika*, 59, 539–549.

Chin, C. (2004), "Optimal Filter Design Approaches to Statistical Process Control for Autocorrelated Processes," unpublished doctoral dissertation, Texas A&M University, Dept. of Industrial Engineering.

Jiang, W. (2001), "Average Run Length Computation of ARMA Charts for Stationary Processes," *Communications in Statistics, Part B—Simulation and Computation*, 30, 699–716.

Jiang, W., Tsui, K., and Woodall, W. H. (2000), "A New SPC Monitoring Method: The ARMA Chart," *Technometrics*, 42, 399–410.

Jiang, W., Wu, H., Tsung, F., Nair, V. N., and Tsui, K. (2002), "Proportional Integral Derivative Charts for Process Monitoring," *Technometrics*, 44, 205–214.

Lin, S. W., and Adams, B. M. (1996), "Combined Control Charts for Forecast-Based Monitoring Schemes," *Journal of Quality Technology*, 28, 289–301.

Lu, C., and Reynolds, M. R., Jr. (1999a), "EWMA Control Charts for Monitoring the Mean of Autocorrelated Processes," *Journal of Quality Technology*, 31, 166–188.

——— (1999b), "Control Charts for Monitoring the Mean and Variance of Autocorrelated Processes," *Journal of Quality Technology*, 31, 259–274.

Lucas, J. M. (1982), "Combined Shewhart–Cusum Quality Control Schemes," *Journal of Quality Technology*, 14, 51–59.

Lucas, J. M., and Saccucci, M. S. (1990), "Exponentially Weighted Moving Average Control Schemes: Properties and Enhancements," *Technometrics*, 32, 1–12.

Luceño, A. (1999), "Average Run Lengths and Run Length Probability Distributions for Cuscore Charts to Control Normal Mean," *Computational Statistics and Data Analysis*, 32, 177–195.

——— (2004), "Cuscore Charts to Detect Level Shifts in Autocorrelated Noise," *Quality Technology and Quantitative Management*, 1, 27–45.

Montgomery, D. C. (2005), *Introduction to Statistical Quality Control* (5th ed.), New York: Wiley.

Montgomery, D. C., and Mastrangelo, C. M. (1991), "Some Statistical Process Control Methods for Autocorrelated Data," *Journal of Quality Technology*, 23, 179–193.

Moustakides, G. (1986), "Optimal Stopping Times for Detecting Changes in Distributions," *The Annals of Statistics*, 14, 1379–1387.

Pandit, S. M., and Wu, S. M. (1983), *Time Series and System Analysis With Applications*, New York: Wiley.

Reynolds, M. R., Jr., and Stoumbos, Z. (2001), "Monitoring the Process Mean and Variance Using Individual Observations and Variable Sampling Intervals," *Journal of Quality Technology*, 33, 181–205.

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

Runger, G. C., and Prabhu, S. S. (1996), "A Markov Chain Model for the Multivariate Exponentially Weighted Moving Averages Control Chart," *Journal of the American Statistical Association*, 91, 1701–1706.

Runger, G. C., and Testik, M. C. (2003), "Control Charts for Monitoring Fault Signatures: Cuscore versus GLR," *Quality and Reliability Engineering International*, 19, 387–396.

Shu, L., Apley, D. W., and Tsung, F. (2002), "Autocorrelated Process Monitoring Using Triggered Cuscore Charts," *Quality and Reliability Engineering International*, 18, 411–421.

Sparks, R. S. (2000), "CUSUM Charts for Signalling Varying Location Shifts," *Journal of Quality Technology*, 32, 157–171.

VanBrackle, L. N., III, and Reynolds, M. R., Jr. (1997), "EWMA and CUSUM Control Charts in the Presence of Correlation," *Communications in Statistics, Part B—Simulation and Computation*, 26, 979–1008.