자기상관 데이터의 통계적 공정관리를 위한 선형 필터 기법

A Linear Filtering Method for Statistical Process Control with Autocorrelated Data

진창호*, Daniel W. Apley**

* 경희대학교 기계·산업시스템공학부 (chin@khu.ac.kr)
** Department of Industrial Engineering and Management Sciences, Northwestern University
(apley@northwestern.edu)

Abstract

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 in order 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.

1. INTRODUCTION

Many common control charting methods are based on linear filtering in the following sense. The statistic to be charted is calculated as the output of a linear filter applied to the sequence of process observations $\{x_t\colon t = 1, 2, 3, \ldots\}$. An alarm is sounded at observation number t if y_t falls outside a set of control limits, where $\{y_t\colon t = 1, 2, 3, \ldots\}$ denote the sequence of control chart statistics. A classic example of this 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, an EWMA chart and a Shewhart individual chart 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 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 (2004) proposed a compete generalization of the concept of a control chart based on linear filtering. They considered a control chart statistic of the form $y_t = \mathrm{H}(B)x_t$, where $\mathrm{H}(B) = h_0 + h_1B + h_2B^2 + \ldots$ is a general linear filter (GLF) in impulse response form, with B denoting the time-series backshift operator and $\{h_j\colon j = 0, 1, 2, \ldots\}$ denoting the impulse response coefficients. They treated this as an optimal filter design problem and developed a method for finding the filter impulse response coefficients 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 (i.i.d.) 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 (2004) directly optimized the design of a GLF of the form $y_t = \mathrm{H}(B)e_t$, where e_t denotes the residuals of an ARMA process model (see Section 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 stable and invertible.

One disadvantage of the method of Apley and Chin (2004) 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 zero). In order to avoid these drawbacks, we propose as a control chart statistic a second-order linear filter (SLF) of the form

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

where α_1, α_2, β 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 Eq. (1) is a special case of the GLF considered in Apley and Chin (2004), in which the filter is second-order and applied to the residuals, as opposed to x_t.

As in Apley and Chin (2004), we focus on optimizing the design of the filter. Specifically, we develop an approach for selecting the SLF parameters α_1, α_2, β, and γ in order to minimize the out-of-control ARL under the constraint that the in-control ARL equals some desired value. In Sections 2 and 3 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 the filter design optimization is one aspect that distinguishes this work from the work on ARMA and PID charts. The heuristic design procedures suggested in Jiang et al. (2000) and Jiang et al. (2002) for the ARMA and PID charts are somewhat ambiguous and may result in control charts that perform far from optimal.

Another difference between this work and the ARMA chart of Jiang et al. (2000) is that our SLF is applied to the residuals. In contrast, 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 following 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 4. Indeed, for many of the examples that we will consider in Section 4, the performance of our optimized SLF almost equals that of the most general linear filter optimized in Apley and Chin (2004).

2. ARL CALCULATION

Throughout this paper, 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 i.i.d. 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, respectively. The model residuals (i.e., the one-step-ahead prediction errors) are generated via 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. Thus, the residuals are an independent sequence of Gaussian random variables with variance σ^2 and time-varying mean $\tilde{\mu}_t$.

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 $t \le 0$, and $\mu_t = \mu$ for $t > 0$), while simultaneously constraining the in-control ARL to some desired value. In order to accomplish this, we express the ARL as a function of the filter parameters using the following variation of the Markov chain approach of Brooks and Evans (1972). Define the vector $V_t = (y_t, z_t)^{\mathrm{T}}$, where $z_t = \alpha_2 y_{t-1} - \gamma\beta e_t$, and note that V_t can be written as

$$V_t = \begin{bmatrix} y_t \\ z_t \end{bmatrix} = \begin{bmatrix} \alpha_1 & 1 \\ \alpha_2 & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} \gamma \\ -\gamma\beta \end{bmatrix} e_t$$
$$= DV_{t-1} + We_t \qquad (2)$$

where $D = \begin{bmatrix} \alpha_1 & 1 \\ \alpha_2 & 0 \end{bmatrix}$ and $W = \begin{bmatrix} \gamma \\ -\gamma\beta \end{bmatrix}$.

Because V_t is a two-dimensional vector Markov process, two-dimensional Markov chain methods similar to those used in 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. Although Jiang (2001) suggests using a three-dimensional vector Markov chain representation for a similar process, we have invoked the observable canonical form (Åström and Wittenmark 1990) of the filter dynamics in order to yield the two-dimensional representation in Eq. (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 in Apley and Chin (2004) 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 in Runger and Prabhu (1996), VanBrackle and Reynolds (1997), and Jiang (2001) is easily applied to the present situation as follows. The two-dimensional state space for $V = (y, z)^T$ is discretized into a set of rectangles, as shown in Figure 1. The range of values for y extends to the upper and lower control limits ±1. Although the z-axis technically extends out to ±∞, we may truncate this by defining the upper and lower limits (L_z, U_z) wide enough 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. The in-control region therefore 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 Q_t denote the $N \times N$ transition probability matrix for the nonabsorbing states at time t. The i^{th} row, j^{th} column element ($1 \le i,j \le N$) of Q_t, denoted Q_t^{ij}, is defined as $Q_t^{ij} = Pr\{V_t \in R_j \mid V_{t-1} = r_i\}$, where R_j is the rectangle for state j, and r_i is the centroid of R_i.

Eq. (2) implies that $V_t|V_{t-1}$ follows a degenerate bivariate normal distribution with mean $DV_{t-1} + W\tilde{\mu}_t$ and rank-1 covariance matrix $WW^T\sigma^2$. In other words, $V_t|V_{t-1}$ is distributed along a one-dimensional line in the two-dimensional state space, as illustrated in Figure 1. Each Q_t^{ij} can be calculated as the area under the normal density

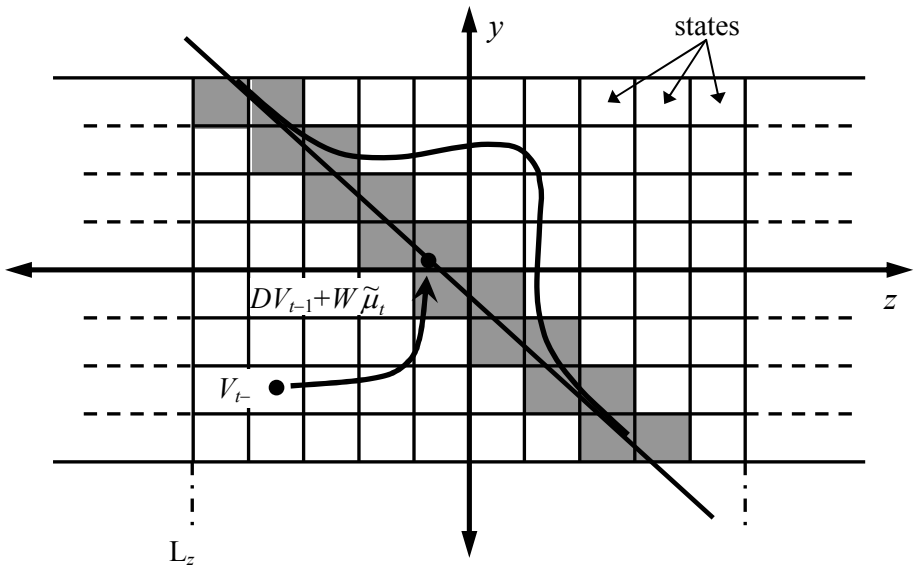


FIGURE 1. Two-dimensional State Space Discretization in the Markov Chain Approach.

In other words, $V_t|V_{t-1}$ is distributed along a one-dimensional line in the two-dimensional state space, as illustrated in Figure 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. If the distribution line does not pass through a particular rectangle, then the corresponding element of Q_t^{ij} is exactly zero. Therefore, although Q_t is an $N\times N$ matrix, each of row of Q_t contains less than $max\{2N_y,2N_z\}$ nonzero elements. Because Q_t is a sparse matrix, the computational expense in calculating the ARL is lessened. Jiang (2001) discusses in more detail how to take advantage of this sparseness.

Let Q_t denote the $N\times N$ transition probability matrix for the nonabsorbing states at time t. The i^{th} row, j^{th} column element ($1 \le i,j \le N$) of Q_t, denoted Q_t^{ij}, is defined as $Q_t^{ij} = Pr\{V_t \in R_j \mid V_{t-1} = r_i\}$, where R_j is the rectangle for state j, and r_i is the centroid of R_i.

Eq. (2) implies that $V_t|V_{t-1}$ follows a degenerate bivariate normal distribution with mean $DV_{t-1} + W\,\tilde{\mu}_t$ and rank-1 covariance matrix WW^{T} [2]. In other words, $V_t|V_{t-1}$ is distributed along a one-dimensional line in the two-dimensional state space, as illustrated in Figure 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. If the distribution line does not pass through a particular rectangle, then the corresponding element of Q_t^{ij} is exactly zero. Therefore, although Q_t is an $N\times N$ matrix, each of row of Q_t contains less than $max\{2N_y,2N_z\}$ nonzero elements. Because Q_t is a sparse matrix, the computational expense in calculating the ARL is lessened. Jiang (2001) discusses in more detail how to take advantage of this sparseness.

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

$$\text{ARL} = \underline{\pi}_0 (\mathrm{I} + Q_1 + Q_1Q_2 + Q_1Q_2Q_3 + \cdots)\underline{1}, \qquad (3)$$

where $\underline{1}$ denotes a column vector of ones and $\underline{\pi}_0$ denotes the initial state probability vector. In all examples, we consider only the zero-state ARL. This is represented by setting all elements of $\underline{\pi}_0$ equal to zero except for the single element corresponding to the initial value $\{y_0 = 0, z_0 = 0\}$, which is set equal to one. Because Q_t depends on t only via the time varying mean of the residuals, Q_t approaches a steady-state value (denoted Q) as $\tilde{\mu}_t$ approaches a steady-state value. For sufficiently large m we therefore have $Q \cong Q_m \cong Q_{m+1} \cong \cdots$, and Eq. (3) becomes

$$\text{ARL} = \sum_{n=1}^{m-1} b_n \underline{1} + b_m[I - Q]^{-1}\underline{1}, \qquad (4)$$

where $b_n = \underline{\pi}_0 \prod_{l=1}^{n-1} Q_l = b_{n-1}Q_{n-1}$ can be calculated recursively for $n = 2, 3, \ldots, m$ with $b_1 = \underline{\pi}_0$. Lu and Reynolds (1999a) provide further discussion of this steady-state truncation in the Markov chain approach.

3. OPTIMAL FILTER DESIGN STRATEGY

We briefly describe the strategy for optimizing the vector of SLF parameters $\zeta = [\alpha_1\ \alpha_2\ \beta\ \gamma]^{\mathrm{T}}$. 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 $\partial\text{ARL}/\partial$, an expression for which we develop in this section.

Let ζ_k denote the k^{th} element of ζ. Differentiating Eq. (3) gives

$$\begin{aligned}
&\frac{\partial ARL}{\partial \zeta_k} \\
&= \underline{\pi}_0 \left[\begin{array}{l} \dfrac{\partial Q_1}{\partial \zeta_k} \\ +\left(\dfrac{\partial Q_1}{\partial \zeta_k} Q_2 + Q_1 \dfrac{\partial Q_2}{\partial \zeta_k} \right) \\ +\left(\dfrac{\partial Q_1}{\partial \zeta_k} Q_2 Q_3 + Q_1 \dfrac{\partial Q_2}{\partial \zeta_k} Q_3 + Q_1 Q_2 \dfrac{\partial Q_3}{\partial \zeta_k} \right) + \Lambda \end{array} \right] \underline{1} \\
&= \underline{\pi}_0 \left[\begin{array}{l} \left(\dfrac{\partial Q_1}{\partial \zeta_k} + \dfrac{\partial Q_1}{\partial \zeta_k} Q_2 + \dfrac{\partial Q_1}{\partial \zeta_k} Q_2 Q_3 + \Lambda \right) \\ +\left(Q_1 \dfrac{\partial Q_2}{\partial \zeta_k} + Q_1 \dfrac{\partial Q_2}{\partial \zeta_k} Q_3 + Q_1 \dfrac{\partial Q_2}{\partial \zeta_k} Q_3 Q_4 + \Lambda \right) \\ +\left(Q_1 Q_2 \dfrac{\partial Q_3}{\partial \zeta_k} + Q_1 Q_2 \dfrac{\partial Q_3}{\partial \zeta_k} Q_4 + Q_1 Q_2 \dfrac{\partial Q_3}{\partial \zeta_k} Q_4 Q_5 + \Lambda \right) + \Lambda \end{array} \right] \underline{1} \\
&= \sum_{n=1}^{m-1} b_n \frac{\partial Q_n}{\partial \zeta_k} c_n + b_m [I - Q]^{-1} \frac{\partial Q}{\partial \zeta_k} c_m \qquad (5)
\end{aligned}$$

where m and b_n are as defined in Section 2, and $c_n = [\mathrm{I} + Q_{n+1} + Q_{n+1}Q_{n+2} + \cdots]\underline{1} = \underline{1} + Q_{n+1}c_{n+1}$ can be calculated recursively for $n = m-1, m-2, \ldots, 1$ with initial condition $c_m = [\mathrm{I} + Q + QQ + \cdots]\underline{1} = [I - Q]^{-1}\underline{1}$.

Using Eqs. (4) and (5) for the ARL and its gradient, we have coded in MATLAB a straightforward gradient-based algorithm for optimizing the SLF parameters, which is available from the authors upon request. Note that numerical evaluation of $\partial \mathrm{ARL}/\partial\zeta$ in Eq. (5) involves roughly the same computational expense as evaluation of the ARL.

4. DISCUSSION AND EXAMPLES

4.1 Comparison with the Optimal EWMA and the OGLF

In this section we compare the optimal SLF (OSLF) with the OGLF of Apley and Chin (2004) and with an optimized residual-based EWMA. The parameters of all three charts are optimized to minimize the out-of-control ARL while constraining the in-control ARL to equal 500. 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. Note that the EWMA can be written as a first-order linear filter $y_t = \mathrm{H}(B)e_t$, where $\mathrm{H}(B) = (1 - (1 - \lambda)B)^{-1}g$, 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 performances of all three charts depend 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 (2004) considered, which are represented by the 28 examples listed in Table 1. The process models are all ARMA(1,1) 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 i.i.d. Without loss of generality, we assume $\sigma = 1$ for the remainder of the paper. We also consider three different types of mean shifts – step, spike, and sinusoidal – and a range of mean shift sizes that depends 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 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, respectively. S_4 has amplitude 1.5 and a period of eight observations.

Table 1 lists the out-of-control ARL values for all three charts for the 28 examples. All ARL values are zero-state values, 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, where of interest, we will show the impulse response coefficients for the OGLFs. The OGLF impulse response coefficients for all 28 examples can be found in Apley and Chin (2004).

As shown in Apley and Chin (2004), the OGLF reduces to the simple-structured EWMA for the case of step mean shifts in i.i.d. 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).

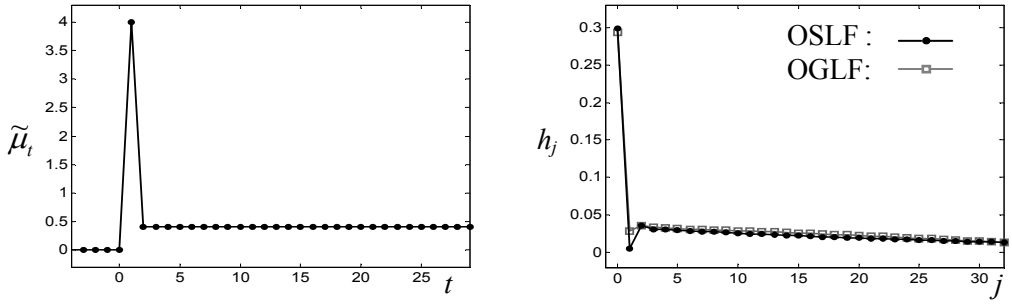
In many of the examples listed in Table 1, a substantial performance improvement can be achieved by increasing the complexity of the filter from the first-order EWMA to the SLF and the GLF. In most of the examples where the EWMA and OGLF performances differ most, the performance of the OSLF also is much 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.

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 2(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) have recommended using a combined Shewhart-EWMA

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	124.20 (.42)	−.558	.322	.326	.1506	15.79 (.02)
14				S_2	30.69 (.04)	1.000	.3236	226.61 (.68)	−.026	−.903	−.243	.1494	30.69 (.04)
15				S_3	32.90 (.04)	.608	.2986	178.47 (.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.



(a) (b)

Figure 2. Residual Mean (a) and Impulse Responses of the OSLF and the OGLF (b) for Example 8.

scheme (see Lucas and Saccucci 1990 and Reynolds and Stoumbos 2001 for additional discussion of combined Shewhart-EWMA charts). The shapes of the OSLF and OGLF impulse response functions shown in Figure 2(b) for Example 8 indicate that they have a close correspondence to a combined Shewhart-EWMA scheme. Note that the plots in Figure 2(b) were obtained by expressing both the OGLF and the OSLF in their impulse response forms $y_t = \mathrm{H}(B)e_t$, with $\mathrm{H}(B) = h_0 + h_1B + h_2B^2 + \ldots$ Note also 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:

$$y_t = .298\left[\frac{1-.847B}{1-.863B-.105B^2}\right]e_t$$
$$= \left[\frac{.264}{1-.108B}\right]e_t + \left[\frac{.034}{1-.971B}\right]e_t$$
$$\cong .264e_t + \frac{.034}{1-.971B}e_t$$

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 2(b). Whereas the typical combined Shewhart-EWMA scheme charts the two statistics separately but simultaneously, the OSLF combines them together into a single statistic y_t. In spite of this difference, we would expect the two charts to behave similarly. The Shewhart component is effective at detecting the initial spike in the residuals, and the EWMA component with small λ is effective at 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, in order to minimize the ARL.

The OSLFs in Examples 5—7 can be similarly viewed as 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 sustained shift in the residual mean.

The OSLF and the OGLF are also quite similar for the AR(1) processes with $\phi = .9$ and spike mean shift (Examples 9 through 12 in Table 1), and both outperform the optimal EWMA for large mean shifts. Figure 3(a) shows the residual mean for Example 12, and Figure 3(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 3. The residual mean oscillates above and below zero on the first two observations following the shift. Both the OSLF and the OGLF are tuned to detect this oscillation, in the sense that their impulse response coefficients also oscillate.

5. CONCLUSIONS

In this paper, we 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 we derived an expression for the derivative of the ARL that is used in a gradient-based algorithm for optimizing the filter parameters.

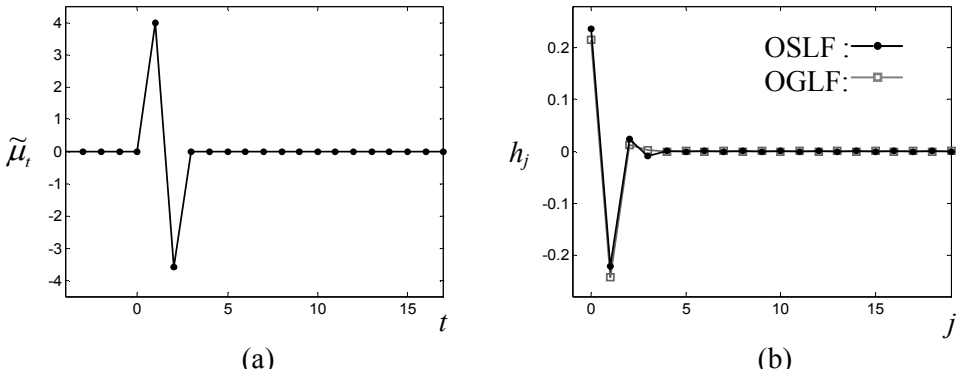


Figure 3. Residual Mean (a) and Impulse Responses of the OSLF and the OGLF (b) for Example 12.

We have demonstrated with 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, using its recursive form. The OGLF has no equivalent recursive form 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 in Apley and Chin (2004), 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.

REFERENCES

Apley, D. W., and Chin, C. (2004), "An Optimal Filter Design Approach to Statistical Process Control," Submitted for Publication.

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.

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 Ph.D. 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–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.

Lu, C., and Reynolds, M. R., Jr. (1999b), "Control Charts for Monitoring the Mean and Variance of Autocorrelated Processes," *Journal of Quality Technology*, 31, 259–274.

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

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

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.

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