An Optimal Filter Design Approach to Statistical Process Control

DANIEL W. APLEY

Northwestern University, Evanston, IL 60208-3119

CHANG-HO CHIN

Kyung Hee University, Gyeonggi-do, Republic of Korea

Many control charts can be viewed as charting the output of a linear filter applied to process data, with an alarm sounded when the filter output falls outside a set of control limits. We generalize this concept by considering a linear filter in its most general time-invariant form. We provide a strategy for optimizing the filter coefficients in order to minimize the out-of-control ARL, while constraining the in-control ARL to some desired value. The optimal linear filters exhibit a number of interesting characteristics, in particular when the process data are autocorrelated. In many situations, they also substantially outperform an optimally designed exponentially weighted moving average (EWMA) control chart.

Key Words: Autocorrelation; Control Charts; Linear Filtering; Markov Chain Method; Statistical Process Control.

Let $\{x_t : t = 1, 2, \ldots\}$ denote the process data and consider a control chart statistic of the form $y_t = \mathrm{H}(B)x_t$, where $\mathrm{H}(B) = h_0 + h_1 B + h_2 B^2 + \cdots$ is some linear filter in impulse response form, B is the time series backward shift operator, and $\{h_j : j = 0, 1, 2, \ldots\}$ are the impulse response coefficients of the filter. Two simple examples of this are a Shewhart individuals chart and an exponentially weighted moving average (EWMA) chart on x_t. For the Shewhart chart, we have $y_t = x_t$, so that $\mathrm{H}(B) = 1$ is the identity filter. For the EWMA chart with parameter λ, we have $y_t = (1-\lambda)y_{t-1} + \lambda x_t$, so that the filter is $\mathrm{H}(B) = \lambda(1-(1-\lambda)B)^{-1} = \lambda + \lambda(1-\lambda)B + \lambda(1-\lambda)^2 B^2 + \cdots$ (Roberts (1959)).

For autocorrelated data, two additional examples are Shewhart and EWMA charts applied to the residuals of an autoregressive moving-average (ARMA) model of the process (see, e.g., Lu and Reynolds (1999a) or Wardell et al. (1994)). Suppose an autocorrelated process is modeled as

$$x_t = \frac{\Theta(B)}{\Phi(B)} a_t + \mu_t,$$

where μ_t represents the process mean (or alternatively, a deterministic process disturbance to be detected), $a_t \sim \mathrm{NID}(0, \sigma_a^2)$, and $\Phi(B) = (1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)$ and $\Theta(B) = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)$ denote the autoregressive (AR) and moving-average (MA) polynomials of order p and q, respectively. We assume that $\mu_t = 0$ when the process is in control. The residuals are generated via the linear filtering operation, $e_t = \Theta^{-1}(B)\Phi(B)x_t$, which gives (Apley and Shi (1999))

$$e_t = \frac{\Phi(B)}{\Theta(B)} x_t = \frac{\Phi(B)}{\Theta(B)} \left[\frac{\Theta(B)}{\Phi(B)} a_t + \mu_t \right] = a_t + \tilde{\mu}_t, \tag{1}$$

where $\tilde{\mu}_t = \Theta^{-1}(B)\Phi(B)\mu_t$ is a filtered version of the deterministic mean shift μ_t. For a residual-based chart, we may view the $\Theta^{-1}(B)\Phi(B)$ factor in Equation (1) as a linear prefilter to the Shewhart identity filter or the EWMA filter.

Dr. Apley is an Associate Professor in the Department of Industrial Engineering and Management Sciences. His email address is apley@northwestern.edu.

Dr. Chin is an Assistant Professor in the School of Mechanical and Industrial Systems Engineering. His email address is chin@khu.ac.kr.

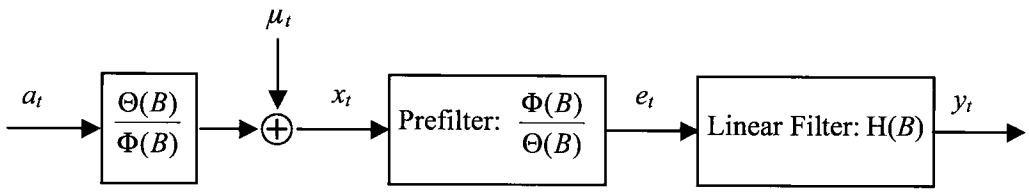


FIGURE 1. Block Diagram Representation of a Control Chart Statistic y_t Generated via a Linear Filtering Operation.

Figure 1 is a block-diagram representation of the generic notion of generating a control chart statistic via a time-invariant (that is, the $\{h_j\}$ do not depend on time t) linear filtering operation, and Table 1 lists the specific forms of linear filters that correspond to a number of different control charts. The inclusion of the whitening prefilter $\Theta^{-1}(B)\Phi(B)$ means that $\mathrm{H}(B)$ is applied to the residuals, and its exclusion means that $\mathrm{H}(B)$ is applied directly to x_t. Table 1 also includes the EWMAST chart of Zhang (1998), the proportional integral derivative (PID) chart of Jiang et al. (2002), and the ARMA(1, 1) chart of Jiang et al. (2000). The latter two were introduced as control charts with more general structure and more flexible design options than the EWMA chart. We note that the usual tabular form of the CUSUM control chart is not a linear filter (neither time varying nor time invariant).

In this paper, we propose a complete generalization of this concept by considering control chart statistics of the form

$$y_t = \mathrm{H}(B)e_t = \sum_{j=0}^{Tr} h_j e_{t-j}, \qquad (2)$$

where Tr is a suitably large truncation time. We treat this as an optimal filter-design problem and present a method for finding the filter coefficients $\{h_j : j = 0, 1, \ldots, Tr\}$ that minimize the out-of-control ARL for a specified mean shift of interest, under the constraint that the in-control ARL equals some desired value. For convenience, we fix the upper and lower control limits (UCL and LCL) on y_t at ± 1 and allow the filter coefficients to be scaled accordingly. We note that y_t is to be calculated and charted for each time period, beginning with $t = 1$. Equation (2) is still valid for $t < Tr$ if we define $e_j = 0$ for $j \leq 0$. One can view this as initializing the filter, similar to how an EWMA is initialized at $t = 0$.

It is entirely for the sake of convenience that we consider filters of the form $y_t = \mathrm{H}(B)e_t$ (applied to the residuals), as opposed to filters of the form $y_t = \mathrm{H}(B)x_t$ (applied to the original data). Because we are assuming a stable, invertible ARMA model (and no model estimation error), any linear filter $\mathrm{H}(B)$ applied to the residuals can be equivalently expressed as the linear filter $\mathrm{H}(B)\Theta^{-1}(B)\Phi(B)$ applied to x_t (i.e., $y_t = \mathrm{H}(B)e_t = \mathrm{H}(B)\Theta^{-1}(B)\Phi(B)x_t$), and vice versa. Consequently, we lose no generality by optimizing over the space of all linear filters applied to e_t rather than to x_t. Considering filters applied to e_t will result in a more computationally efficient optimization algorithm and, as we will demonstrate, yields a more illuminating interpretation of the optimal general linear filter (OGLF) coefficients. Note that, for uncorrelated data, $\Theta(B) = \Phi(B) = 1$, and e_t and x_t coincide.

TABLE 1. Control Charts Based on Linear Filtering

Control chart	Charted statistic	Prefilter	Linear filter
Shewhart on x_t	$y_t = x_t$	No	1
EWMA on x_t	$y_t = (1-\lambda)y_{t-1} + \lambda x_t$	No	$\lambda/[1-(1-\lambda)B]$
Shewhart on e_t	$y_t = e_t = [\Phi(B)/\Theta(B)]x_t$	Yes	1
EWMA on e_t	$y_t = (1-\lambda)y_{t-1} + \lambda e_t$	Yes	$\lambda/[1-(1-\lambda)B]$
ARMA(1,1) chart on x_t	$y_t = [(\theta_0 - \theta B)/(1-\phi B)]x_t$	No	$(\theta_0 - \theta B)/(1-\phi B)$
PID Chart	$y_t = (1-k_I)y_{t-1} - k_P(1-B)y_{t-1}$ $-k_D(1-B)^2 y_{t-1} + (1-B)x_t$	No	$(1-B)/[1-(1-k_I-k_P-k_D)B$ $-(k_P+2k_D)B^2 + k_D B^3]$

The following section discusses a method for optimizing the filter design that is based on a Markov chain approximation of the ARL and its first derivative. We then compare the performance of the OGLF with the optimally designed EWMA and with the PID chart of Jiang et al. (2002). For some examples, such as a step mean shift in i.i.d. data, the OGLF reduces to the optimally designed EWMA. For other examples, the OGLF bears a close resemblance to the combined Shewhart–EWMA scheme of Lucas and Saccucci (1990). For still other examples, especially with highly autocorrelated data, the sequence of optimal filter coefficients $\{h_j\}$ shares certain shape characteristics of the residual mean sequence $\{\tilde{\mu}_j\}$. We discuss these and other interesting characteristics of the OGLF. Following this, we explain why, even though the OGLF is optimized for detecting mean shifts that occur at the initial observation, it is equally effective at detecting shifts that occur at some later point in time in a steady-state analysis. We also discuss guidelines to help ensure that the OGLF is effective at detecting a range of shift sizes. The final section concludes the paper.

Strategy for Optimizing the Filter Design

This section describes concepts that we use in a gradient-based optimization strategy for determining the OGLF design. The inputs to the optimization procedure are (i) the ARMA process model; (ii) the specific mean shift of interest $\{\mu_t : t = 1, 2, \ldots\}$, which includes both magnitude and shape (we consider step, spike, and sinusoidal in the examples); and (iii) the desired in-control ARL. The optimization algorithm calculates the general linear filter (GLF) coefficients that minimize the out-of-control ARL for that mean shift. This requires approximate calculation of the ARL and its gradient (its derivative with respect to the filter coefficients). We use a Markov chain method (Brook and Evans (1972)) as the basis for calculating both. Although y_t in Equation (2) is not a (scalar) Markov process, we approximate it as such:

$$f_{y_t|y_{t-1}}(s_t \mid s_{t-1}) \cong f_{y_t|y_{t-1},y_{t-2},\ldots}(s_t \mid s_{t-1}, s_{t-2}, \ldots) \quad (3)$$

where s_t represents a specific value of y_t and $f_{\cdot|\cdot}$ denotes the conditional density of y_t. The approximation (3) is obviously a crude one for complex-structured GLFs. In spite of this, we have found it to be quite useful in the optimization routine, leading to OGLFs that often perform much better than an optimized EWMA. Furthermore, Chin (2004) and Chin and Apley (2006) have investigated a restricted second-order version of the GLF, which can be optimized using exact Markov chain methods instead of the approximation (3). For a wide variety of examples, they found that an OGLF optimized using the approximation (3) always performed as well as or better than a second-order filter optimized using more exact ARL calculations. Thus, the approximation (3) evidently provides a reasonable relative comparison of the ARLs for different GLFs and a means of approximating the gradient of the ARL. We also use Monte Carlo simulation when more precise ARL calculations are required, such as at the final stage of the optimization to guarantee that the OGLF has the desired in-control ARL.

Because y_t is generated as a linear filtering operation on the Gaussian process a_t, y_t and y_{t-1} have the joint Gaussian distribution

$$\begin{bmatrix} y_t \\ y_{t-1} \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \mu_{t,y} \\ \mu_{t-1,y} \end{bmatrix}, \begin{bmatrix} \sigma_t^2 & v_t \\ v_t & \sigma_{t-1}^2 \end{bmatrix} \right),$$

where $\mu_{t,y} = \sum_{j=0}^{t-1} h_j \tilde{\mu}_{t-j}$ is the mean of y_t, $v_t = \sigma_a^2 \sum_{j=0}^{t-2} h_j h_{j+1}$ is the covariance between y_t and y_{t-1}, and $\sigma_t^2 = \sigma_a^2 \sum_{j=0}^{t-1} h_j^2$ is the variance of y_t. In the preceding summations, the notational convention is that $h_j = 0$ for $j > Tr$. For the bivariate normal random variables $\{y_t, y_{t-1}\}$, the conditional distribution becomes (Johnson and Wichern (2002))

$$y_t \mid y_{t-1} \sim \mathrm{N}\left(\mu_{t,y} + \frac{v_t(y_{t-1} - \mu_{t-1,y})}{\sigma_{t-1}^2}, \sigma_t^2 - \frac{v_t^2}{\sigma_{t-1}^2} \right). \quad (4)$$

In the Markov chain approach, we discretize the in-control region $(-1, 1)$ for y_t into N subintervals of length $\delta = 2/N$, and the out-of-control region (y_t outside the $(-1, 1)$ interval) is treated as a single absorbing state. This is illustrated in Figure 2, where A_j denotes the subinterval representing state j, and $\alpha_j = \mathrm{LCL} + (j - 1/2)\delta$ is the midpoint of A_j. Let $\mathbf{Q}_t$ denote the state transition probability matrix at time t, the ith row, jth column element of which is defined as $(1 \le i, j \le N)$

$$\begin{aligned} Q_t^{ij} &= \Pr\{y_t \in A_j \mid y_{t-1} = \alpha_i\} \\ &= \Pr\{\alpha_j - \delta/2 < y_t \le \alpha_j + \delta/2 \mid y_{t-1} = \alpha_i\}. \quad (5) \end{aligned}$$

We can approximate ARL $\approx \boldsymbol{\pi}_0(\mathbf{I} + \mathbf{Q}_1 + \mathbf{Q}_1\mathbf{Q}_2 + \mathbf{Q}_1\mathbf{Q}_2\mathbf{Q}_3 + \cdots)\mathbf{1}$, where $\mathbf{1}$ is a column vector of ones, and $\boldsymbol{\pi}_0$ is a row vector of which the jth element is $\Pr\{y_0 \in A_j\}$ (see Brook and Evans (1972)).

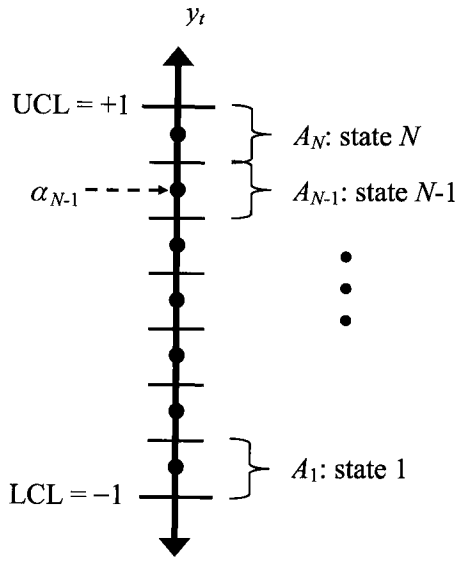


FIGURE 2. Discretized One-dimensional State Space in the Markov Chain Approximation.

Throughout, we take y_0 to be identically zero. A convenient alternative expression for the ARL is

$$\text{ARL} = \sum_{p=1}^{m-1} \mathbf{b}_p \mathbf{1} + \mathbf{b}_m [\mathbf{I} - \mathbf{Q}]^{-1} \mathbf{1}, \qquad (6)$$

where m is large enough to approximate $\mathbf{Q}_t$ by its steady-state value (denoted by $\mathbf{Q}$) for $t \geq m$, and $\mathbf{b}_p = \boldsymbol{\pi}_0 \prod_{l=1}^{p-1} \mathbf{Q}_l$. See Lu and Reynolds (1999a) for additional discussion of this truncated representation of the ARL. Because the filter $\text{H}(B)$ is a moving function of the residuals, with $\{h_j\}$ decaying to zero (this is always true because the filter is truncated; however, even with large truncation times we have observed the filter coefficients decaying to zero for every example that we have encountered), and the residuals are independent, $\mathbf{Q}_t$ will converge to a steady-state value as long as the mean of the residuals converges to a constant steady-state value. We recommend taking m to be the truncation time Tr added to the time that it takes for the mean of the residuals to settle down to a steady-state value. Appendix B provides guidelines for selecting Tr. Based on Equation (6), Chin (2004) has shown that the derivative of the ARL with respect to each filter coefficient h_j $(j = 0, 1, 2, \ldots)$ is given by

$$\frac{\partial \text{ARL}}{\partial h_j} = \sum_{p=1}^{m-1} \mathbf{b}_p \frac{\partial \mathbf{Q}_p}{\partial h_j} \mathbf{c}_p + \mathbf{b}_m [\mathbf{I} - \mathbf{Q}]^{-1} \frac{\partial \mathbf{Q}}{\partial h_j} \mathbf{c}_m, \quad (7)$$

where $\mathbf{c}_p = [\mathbf{I} + \mathbf{Q}_{p+1} + \mathbf{Q}_{p+1}\mathbf{Q}_{p+2} + \cdots]\mathbf{1}$ and $\partial \mathbf{Q}_p / \partial h_j$ is the element-by-element derivative matrix. Expressions for $\mathbf{Q}_p$ and $\partial \mathbf{Q}_p / \partial h_j$ are provided in Appendix A. For computational purposes, $\mathbf{b}_p$ can be calculated recursively via $\mathbf{b}_p = \mathbf{b}_{p-1}\mathbf{Q}_{p-1}$ with initial condition $\mathbf{b}_1 = \boldsymbol{\pi}_0$, and $\mathbf{c}_p$ can be calculated recursively backward in time via $\mathbf{c}_p = \mathbf{1} + \mathbf{Q}_{p+1}\mathbf{c}_{p+1}$ with initial condition $\mathbf{c}_m = [\mathbf{I} + \mathbf{Q} + \mathbf{Q}\mathbf{Q} + \cdots]\mathbf{1} = [\mathbf{I} - \mathbf{Q}]^{-1}\mathbf{1}$.

Because of the large number of filter coefficients, incorporating the gradient information of Equation (7) into the optimization routine substantially reduces the computational expense involved in searching for the optimal filter. The optimization algorithm coded in Matlab™ is available on request from the authors. Some details of the algorithm are described in Appendix B, and further details can be found in Chin (2004).

Discussion and Examples

Comparison with the PID Chart

In this section, we compare the OGLF with the PID chart of Jiang et al. (2002). The PID chart, the form of which is listed in Table 1, can be viewed as a time-invariant linear filtering operation with three independent design parameters (k_P, k_I, and k_D). In this respect, it is the closest conceptually to the OGLF. We illustrate with the same autocorrelated process considered in Jiang et al. (2002) and Jiang et al. (2000), which is a mechanical vibration system originally considered in Pandit and Wu (1983). The estimated process model for the vibration signal was the ARMA(2, 1) model (see Jiang et al. (2000)),

$$x_t - 1.44x_{t-1} + .60x_{t-2} = a_t + .52a_{t-1}, \qquad (8)$$

with $\sigma_a = 2.21$ and $\sigma_x = 9.13$. Note that, for an ARMA process, σ_x can be calculated from σ_a and the ARMA parameters using, for example, the impulse-response method (Box et al. (1994)). For notational simplicity, suppose the observations are scaled so that $\sigma_a = 1$ ($\sigma_x = 4.13$). Using this as the model, we conducted Monte Carlo simulations to compare the ARL performances of the OGLF, a Shewhart individuals chart on the residuals of Equation (8), and the PID chart. Table 2 shows the results. The in-control ARL was 370 for all charts. For the out-of-control ARLs, we added a step mean shift of magnitude μ occurring at observation number one (i.e., $\mu_t = 0$ for $t < 1$ and $\mu_t = \mu$ for $t \geq 1$). We used the same values of $\Delta = \mu/\sigma_x$ as in Jiang et al. (2002), which are shown in the first column of Table 2. Table 2 includes the results for four different OGLFs, each one optimized for one of the four different mean shift

TABLE 2. ARLs of the OGLFs, the Residual-Based Shewhart Chart, and the PID Charts. The Standard Errors Are in Parentheses

Shift ($\Delta = \mu/\sigma_X$)	OGLF ($\Delta = 0.5$)	OGLF ($\Delta = 1$)	OGLF ($\Delta = 2$)	OGLF ($\Delta = 3$)	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 $\{k_P, k_I, k_D\} = \{-.8, 0, .5\}$
0	370	370	370	370	370	370	370	370
	(0.68)	(0.73)	(0.74)	(0.74)	(0.74)	(0.73)	(0.73)	(0.72)
.5	61.26	90.56	200	200	200	141	351	118
	(0.15)	(0.26)	(0.56)	(0.56)	(0.56)	(0.27)	(0.72)	(0.22)
1	1.90	1.40	3.56	3.56	3.56	44.9	118	37.3
	(0.00)	(0.01)	(0.06)	(0.06)	(0.06)	(0.08)	(0.53)	(0.06)
2	1.00	1.00	1.00	1.00	1.00	11.6	1.00	10.9
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.02)	(0.00)	(0.01)
3	1.00	1.00	1.00	1.00	1.00	5.44	1.00	5.60
	(0.00	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)

sizes. The parameters for the three PID charts were taken directly from Table 1 of Jiang et al. (2002), which they had selected according to their heuristic design guidelines. More specifically, they had chosen 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 had argued, should translate to reasonably good ARL performance for that size mean shift. The ARLs shown in Table 2 are zero-state ARLs calculated using 250,000 replicates.

For all mean shifts in Table 2, the OGLF optimized for that shift (the value of Δ for which each OGLF was optimized is shown in parentheses) performs substantially better than the residual-based Shewhart chart or any of the PID charts, except for the cases where both charts detect the shift nearly perfectly (ARL = 1.00). The explanation for why the OGLF performs better than the other charts in Table 2 lies in the shape of the filter and in the mean of the residuals, which are shown in Figure 3. The summation in Equation (2) can be viewed as a measure of how closely the past residuals "correlate" with the time-reversed version of $\{h_j\}$ for the OGLF (because $\{h_j\}$ is deterministic, by "correlation" we mean the summation of the product of the two sequences). Figure 3(e) illustrates this for the $\Delta = 0.5$ case by showing the time-reversed OGLF (scaled for visualization purposes) superimposed on the mean of the residuals, three time steps after the occurrence of the mean shift. At this point in time, the time-reversed OGLF correlates quite highly with the residual mean. We would expect the resulting OGLF statistic y_t to be large, with a reasonably high probability of exceeding its UCL. At the time steps immediately preceding and immediately following this (i.e., two and four time steps following the occurrence of the shift), the OGLF has high negative correlation with the residual mean and we would expect the charted statistic to be large and negative. Because the control limits are two sided, this also is likely to cause an alarm. We discuss this in more detail later (see Figure 12, for example).

A similar explanation applies for the case $\Delta = 1.0$, the OGLF $\{h_j\}$ for which is shown in Figure 3(c). Figure 3(d) shows the OGLF $\{h_j\}$ for $\Delta = 2$ and $\Delta = 3$, both of which coincide with the Shewhart filter ($h_0 = 0.33$ and $h_j = 0$ for $j \geq 1$). This is reasonable because a mean shift of size $\mu = 2\sigma_x = 8.26\sigma_a$ causes an initial spike of 8.26 standard deviation units in the mean of the residuals, which can be detected with probability ≈ 1.0.

Performance Improvement over the Optimally Designed EWMA

In this section, we compare the OGLF to an optimized residual-based EWMA chart of the form

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

with control limits ± 1, where $0 < \lambda \leq 1$ is the EWMA parameter and k is a scaling constant (to account for the fixed control limits). For each example, the EWMA design parameters k and λ are chosen using the same criteria as for the OGLF: Minimize the out-of-control ARL while providing a desired in-

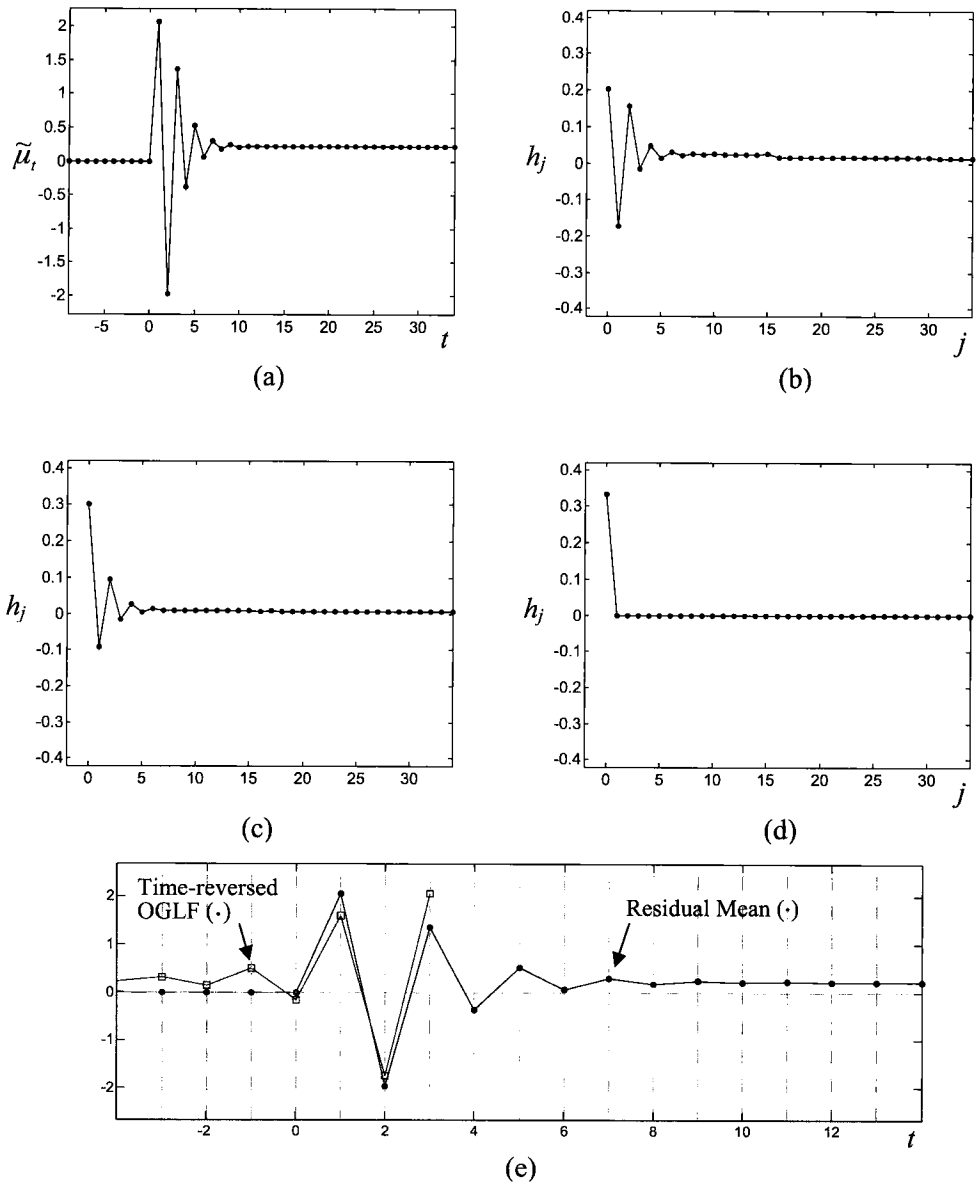


FIGURE 3. Mass–Spring–Dashpot System: (a) Residual Mean for $\Delta = 0.5$; (b) OGLF $\{h_j\}$ for $\Delta = 0.5$; (c) OGLF $\{h_j\}$ for $\Delta = 1$; (d) OGLF $\{h_j\}$ for $\Delta = 2$ and $\Delta = 3$; and (e) the OGLF for $\Delta = 0.5$ Applied to the Residual Mean Three Time Steps after the Occurrence of the Shift.

control ARL. Henceforth, we will refer to this as the optimal EWMA, with the understanding that it is optimal over the class of all EWMA charts, and not over the class of all control charts in general. Note that Shewhart individuals charts are a special case of the EWMA with $\lambda = 1$.

Because the ARL performance of all charts depends heavily on the form and magnitude of the residual mean and on the ARMA model describing the process, we considered a broad combination of scenarios in the 32 examples listed in Table 3. The different models considered are all ARMA(1, 1) of the form $x_t - \phi x_{t-1} = a_t - \theta a_{t-1}$, special cases of which include AR(1) [$\theta = 0$ in Examples 5–12] and i.i.d. [$\phi = \theta = 0$ in Examples 1–4 and 13–16]. Without loss of generality, we will assume $\sigma_a = 1$ for the remainder of the paper. We consider three different types of mean shifts—step, spike, and sinusoidal—and a range of mean shift sizes that depends on the spe-

TABLE 3. ARL Comparison of the OGLF and the Optimal EWMA

	Time series model		Shift		OGLF	Optimal EWMA			% reduction
No	ϕ	θ	Type	Size μ	ARL	λ	k	ARL	R
1	0	0	Step	0.5	28.82 (0.03)	0.047	0.11672	28.82 (0.03)	0
2				1.5	5.45 (0.01)	0.242	0.21791	5.45 (0.01)	0
3				3	1.86 (0.00)	0.676	0.30670	1.86 (0.00)	0
4				4	1.21 (0.00)	0.887	0.32161	1.21 (0.00)	0
5	0.9	0	Step	0.5	355.31 (0.57)	0.002	0.05271	355.31 (0.57)	0
6				1.5	130.64 (0.18)	0.007	0.06540	130.64 (0.18)	0
7				3	46.91 (0.10)	0.021	0.08866	49.43 (0.07)	5
8				4	13.72 (0.06)	0.038	0.10802	29.78 (0.05)	54
9	0.9	0	Spike	0.5	495.39 (0.98)	1.00	0.32360	497.12 (1.00)	0
10				1.5	422.01 (0.98)	1.00	0.32360	454.46 (0.99)	7
11				3	82.72 (0.54)	1.00	0.32360	177.83 (0.76)	53
12				4	6.72 (0.14)	1.00	0.32360	28.70 (0.32)	77
13	0	0	Sinusoid	S_1	15.79 (0.02)	1.00	0.32360	124.20 (0.42)	85
14				S_2	30.69 (0.04)	1.00	0.32363	226.61 (0.68)	82
15				S_3	32.90 (0.04)	0.608	0.29861	178.47 (0.57)	76
16				S_4	10.61 (0.01)	0.616	0.29965	26.31 (0.05)	60
17	0.9	−0.9	Step	0.5	447.66 (0.75)	0.002	0.05271	447.66 (0.75)	0
18				1.5	139.26 (0.54)	0.003	0.05565	255.72 (0.39)	46
19				2	41.54 (0.36)	0.004	0.05838	194.09 (0.28)	79
20				3	3.12 (0.03)	1.00	0.32360	76.23 (0.49)	96

TABLE 3. *Continued*

No	Time series model ϕ	θ	Shift Type	Size μ	OGLF ARL	Optimal EWMA λ	k	ARL	% reduction R
21	0.9	0.5	Step	0.5	205.04 (0.30)	0.004	0.05839	205.58 (0.30)	0
22				1.5	50.28 (0.07)	0.021	0.08874	50.28 (0.07)	0
23				3	10.77 (0.03)	0.120	0.16616	10.77 (0.03)	0
24				4	2.74 (0.01)	0.304	0.23735	2.88 (0.01)	5
25	0.9	0.5	Spike	0.5	497.47 (0.99)	1.00	0.32363	497.61 (0.99)	0
26				1.5	461.86 (0.99)	1.00	0.32360	469.74 (0.99)	2
27				3	208.77 (0.80)	1.00	0.32360	259.67 (0.87)	20
28				4	50.75 (0.41)	1.00	0.32360	86.10 (0.56)	41
29	0.5	0.9	Step	0.5	7.86 (0.00)	0.223	0.21121	7.90 (0.00)	0
30				1.5	3.03 (0.00)	0.498	0.28170	3.05 (0.00)	0
31				3	1.59 (0.00)	0.877	0.32123	1.59 (0.00)	0
32				4	1.18 (0.00)	0.991	0.32362	1.18 (0.00)	0

cific 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 3. S_1, S_2, and S_3 are sinusoidal functions with amplitude 0.75 and periods of 2, 4, and 8 time steps, respectively. S_4 has amplitude 1.5 and period 8 time steps. Note that, with the sinusoidal shifts, the mean of the residuals does not settle down to a steady-state value. For this situation, in the context of the OGLF optimization algorithm, one can select m large enough so that the steady-state term in Equation (6) essentially disappears. Alternatively, noting that $\mathbf{Q}_t$ is cyclostationary in the steady state, one could develop a cyclostationary version of Equation (6) for maximum computational efficiency.

As in the simulations of the previous section, we calculate the zero-state ARL using 250,000 Monte Carlo replications. For the out-of-control scenarios, the mean shift was added at time step one. The in-control ARL is 500 in all cases. The results are shown in Table 3, where the percentage reduction $R = (\text{ARL}_{\text{EWMA}} - \text{ARL}_{\text{OGLF}})/\text{ARL}_{\text{EWMA}}$ measures the extent to which the OGLF outperforms the optimized EWMA. For all 32 examples, the optimized parameters of the EWMA are listed. The OGLF coefficients for each example are plotted in Appendix C.

Providing that the optimization algorithm converges to the globally optimal GLF, the OGLF cannot perform worse than the optimal EWMA because the EWMA is a special case of the GLF. This can be observed in Table 3, where the OGLF performance ranges anywhere from comparable with the EWMA to substantially better, depending on the specific example. For a step mean shift in i.i.d. data (Examples

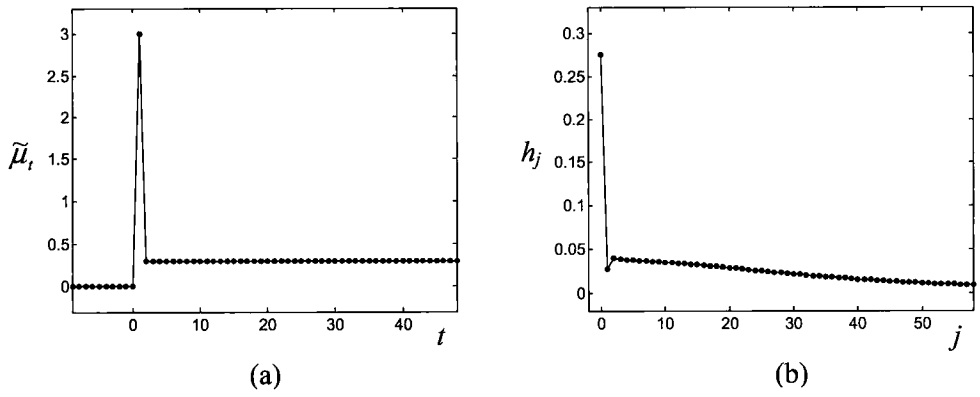


FIGURE 4. Residual Mean (a) and $\{h_j\}$ (b) for Example 7.

1–4), the OGLF and EWMA performances are identical. In fact, as will be discussed in the following section, the OGLF impulse responses in these examples are virtually identical to the EWMA impulse responses. In contrast, in one extreme case (Example 20), the ARL for the OGLF is 3.1, versus 76 for the optimal EWMA. The following section provides some explanations and interpretations of these observations.

Table 3 does not contain many models in which the ARMA parameters are both small (e.g. with magnitude less than 0.5). The reason is that, for step mean shifts in these situations, the OGLF turns out quite similar to an EWMA (see Examples 1–4, 21–24, and 29–32 in Appendix C), in shape and in performance. We discuss this phenomenon in the context of Examples 1–4 for detecting step mean shifts in i.i.d. data ($\phi = \theta = 0$) in the next section.

Optimal Filter Characteristics

For the i.i.d. processes with a step mean shift (Examples 1–4 in Table 3), the OGLFs are very similar to the EWMAs. If one were to superimpose the impulse responses of the optimal EWMAs on top of $\{h_j\}$ for the OGLFs for Examples 1–4 shown in Appendix C, they would be virtually indistinguishable. Evidently, out of the class of all time-invariant linear filters, a simple EWMA is optimal or very nearly so. Note that, for any time-invariant linear filter, such as an EWMA, the impulse-response coefficients can be calculated in a simple manner by recognizing that they are precisely the output response of the filter with an input sequence that consists of a single unit-magnitude impulse (i.e., spike) at time zero. Hence, the impulse-response coefficients can be calculated by simulating the output of the filter for an input sequence defined as $u_0 = 1$, and $u_j = 0$ for $j = 1, 2, 3, \ldots$. The impulse-response coefficient h_j is the simulated output at observation j. Additional details of how to calculate the impulse-response coefficients for a linear filter can be found in Box et al. (1994).

For the AR(1) process with a step mean shift (Examples 5–8 in Table 3), the mean of the residuals settles down to a small steady-state value after an initial spike at the first time step following the shift, as illustrated in Figure 4(a) (see Wardell et al. (1994), Superville and Adams (1994), and Apley and Shi (1999), for further discussion of this phenomenon and its implications). For mean shifts $\mu = 0.5$ and $\mu = 1.5$, the initial spike in the residuals is not large enough to substantially improve detection, and we must rely primarily on the small steady-state value of the residual mean. Correspondingly, the OGLF converges to something that is very close to the optimal EWMA, having small λ, which can be seen from the plots of $\{h_j\}$ in Appendix C for Examples 5 and 6.

In contrast, for the larger mean shifts with $\mu = 3$ and $\mu = 4$ in Examples 7 and 8, the initial spike in the residual mean is large enough to affect detection performance. In these cases, the shape of the OGLF is quite interesting, as illustrated in Figure 4(b) for $\mu = 3$. The OGLF shown in Figure 4(b) has a close connection to a combined Shewhart–EWMA procedure, discussions of which can be found in Lucas and Saccucci (1990) and Reynolds and Stoumbos (2001). The OGLF for this example is very closely approxi-

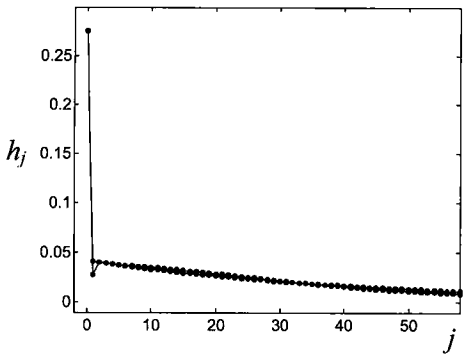


FIGURE 5. Comparison of the OGLF $\{h_j\}$ for Example 7 and Its Combined Shewhart–EWMA Approximation of Equation (9). The OGLF has Slightly Lower $\{h_1\}$.

mated by the filter

$$y_t = \left[\frac{0.275 - 0.228B}{1 - 0.977B}\right] e_t = 0.233e_t + \frac{0.042}{1 - 0.977B} e_t. \tag{9}$$

This is evident from Figure 5, in which we superimpose the impulse response of Equation (9) and the OGLF $\{h_j\}$ for Example 7. The term $0.233e_t$ in Equation (9) represents a Shewhart individuals chart on the residuals. The term on the far right represents an EWMA of the residuals with EWMA parameter $\lambda = 1 - 0.977 = 0.023$. Thus, Equation (9) is a weighted sum of the Shewhart individuals chart statistic and an EWMA statistic. If the weighted sum falls outside a control limit, the OGLF signals. In comparison, the combined Shewhart–EWMA procedure of Lucas and Saccucci (1990) signals if *either* the Shewhart chart *or* the EWMA chart signals. The relative weighting of the two component charts in the combined Shewhart–EWMA procedure is controlled by choice of their respective control limits. For the OGLF, the relative weighting has been determined automatically, in order to minimize the out-of-control ARL.

Although the two approaches are slightly different, we would expect them to behave similarly. In particular, they are both designed to detect large shifts (such as the initial spike in the residual mean) quickly and to eventually detect small shifts (such as the small steady-state shift in the residual mean) with as short an ARL as possible. Lu and Reynolds (1999b) have demonstrated the effectiveness of combined Shewhart–EWMA schemes for detecting mean shifts in autocorrelated processes. The OGLF strategy can be viewed as a way of optimizing the EWMA parameter and the relative weighting of the Shewhart and EWMA component charts. In this example, the optimized EWMA parameter $\lambda = 0.023$ is smaller than what one might normally consider using.

In Examples 9–12, which consider spike mean shifts in an AR(1) process, the relative performance improvement of the OGLF is dramatic. The reason becomes apparent if we compare the shapes of the residual mean and the OGLF $\{h_j\}$ shown in Figure 6. The mean of the residuals oscillates twice before settling down to zero. Because an EWMA with small λ would not be effective at detecting the oscillating mean, the optimal EWMA is essentially a Shewhart individuals chart, as indicated in Table 3. The OGLF in Figure 6(b) is more effective, however, because it is tuned to have high correlation with the residual mean.

The results for i.i.d. processes with a sinusoidal mean shift (Examples 13–16) are as one would expect, with the OGLF outperforming the optimal EWMA by a wide margin. Once again, the optimal

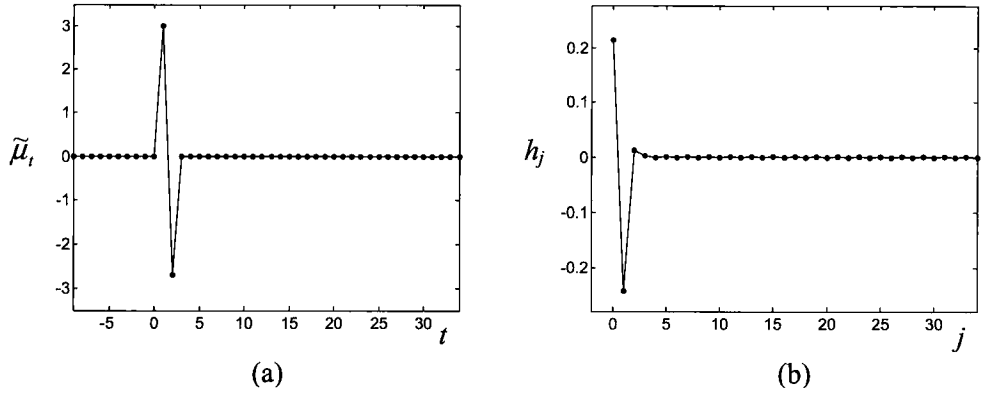


FIGURE 6. Residual Mean (a) and OGLF $\{h_j\}$ (b) for Example 11.

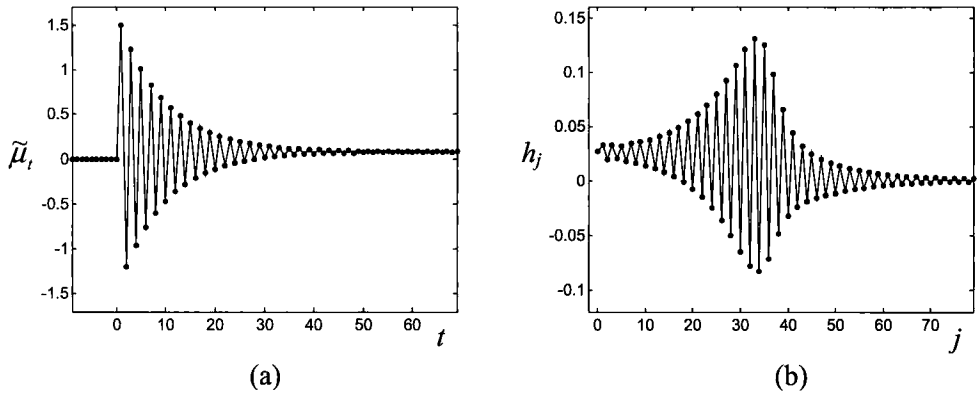


FIGURE 7. Residual Mean (a) and OGLF $\{h_j\}$ (b) for Example 18.

EWMAs are very similar to Shewhart individuals charts. As shown in Appendix C, the OGLF $\{h_j\}$'s appear to be exponentially decaying sinusoids of the same frequency as the sinusoidal process disturbance. This obviously results in high correlation between the OGLF $\{h_j\}$ and the disturbance every half period.

Examples 17–20, for step mean shifts in an ARMA(1, 1) process, with $\phi = 0.9$ and $\theta = -0.9$, are interesting in that the shape of the OGLF depends strongly on the size of the mean shift (see Appendix C). Note that the residual mean, shown in Figure 7(a), oscillates in a manner similar to the residual mean shown in Figure 3(a) for the mechanical vibration system. For $\mu = 0.5$ (Example 17), the shift is so small that no method works very well and the OGLF is virtually the same as the optimal EWMA. For $\mu = 2$ and $\mu = 3$ (Examples 19 and 20), the shift is large enough that the oscillating OGLFs shown in Appendix C work reasonably well. For the midrange mean shift $\mu = 1.5$ (Example 18), however, the OGLF has the unusual shape shown in Figure 7(b). The $\{h_j\}$'s start off quite small, slowly growing in magnitude until they peak at approximately time step 34, and then slowly decaying in magnitude. The fact that the $\{h_j\}$'s start off small means that the OGLF sacrifices some level of early detectability (which would be quite small regardless of the method that is used) in return for better detection at a later point and, overall, a lower ARL. In Figure 8, which helps clarify this, we decompose the OGLF via $h_j = h_{1,j} + h_{2,j}$, where $h_{1,j}$ consists of the first 34 coefficients ($h_{1,j} = h_j$ for $0 \le j \le 33$, and $h_{1,j} = 0$ otherwise) and $h_{2,j}$ consists of the remaining coefficients. Figure 9 superimposes the time-reversed $h_{1,j}$ (scaled for visualization purposes) on top of the residual mean 34 time steps following the occurrence of the mean shift. The two are almost perfectly correlated, which (roughly) maximizes the probability of detecting by the 34th time step. The remaining OGLF component, $h_{2,j}$, the shape of which is similar to the OGLFs from Examples 19 and 20, evidently further reduces the ARL by improving the eventual detection in the event that the shift is not detected within 34 time steps.

One might wonder whether the unusual shape of the OGLF in Figure 8 is an artifact of inaccura-

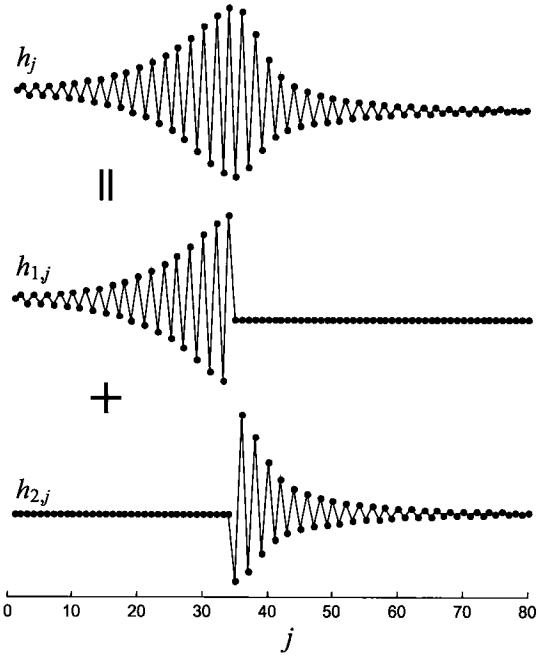


FIGURE 8. Decomposition of the OGLF $\{h_j\}$ of Example 18 into its First 34 Coefficients $\{h_{1,j}\}$ and the Remaining Coefficients $\{h_{2,j}\}$.

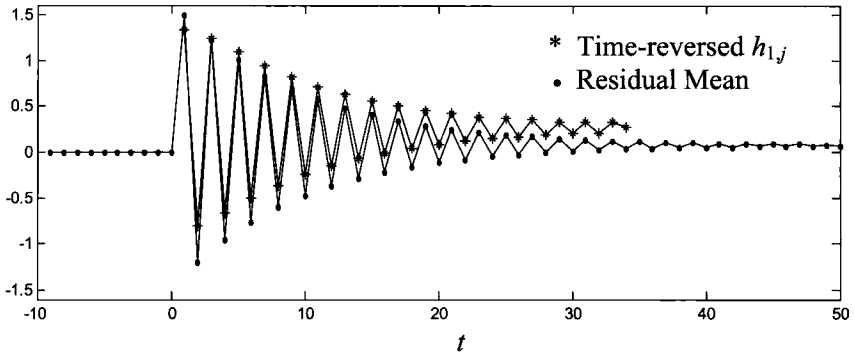


FIGURE 9. Illustration of the Nearly Perfect Correlation Between the Time-Reversed OGLF Component $\{h_{1,j}\}$ and the Residual Mean $\tilde{\mu}_t$ at 34 Time Steps Following the Occurrence of the Shift.

cies or approximations in the numerical optimization algorithm and whether the more intuitively shaped OGLFs from Example 17 or Example 19 might do a better job of detecting the mean shift of Example 18. These two examples differ from Example 18 only in the size of the shift (Example 17 is smaller; Example 19 is larger). Additional simulation reveals that the ARLs for detecting the 1.5-magnitude shift of Example 18 are 258 and 164 using the OGLFs from Examples 17 and 19, respectively. Similarly, the ARL using only the $h_{2,j}$ component shown in Figure 8 (with the leading zeros removed and $h_{2,j}$ scaled by 1.316, so as to provide an in-control ARL of 500) is 168. These are all substantially worse than the ARL of 139 that results from using the OGLF shown in Figure 8. We conclude that both $h_{1,j}$ and $h_{2,j}$ are important components of the OGLF from Example 18 and that they are not the result of a misbehaved optimization algorithm.

A similar phenomenon occurs in Examples 25–28 for spike mean shifts in an ARMA(1, 1) process with $\phi = 0.9$ and $\theta = 0.5$. The mean of the residuals and the OGLF for $\mu = 4$ (Example 28) are shown in Figure 10. The time-reversed OGLF $\{h_j\}$ is almost perfectly correlated with the residual mean five time steps following the occurrence of the shift. Under the larger shifts ($\mu = 3$ and $\mu = 4$), this results in substantially shorter ARLs than for the optimal EWMA, which, in these cases, are Shewhart individuals charts. Note that correlating the residuals with a function that coincides with the residual mean under the assumed mean shift is reminiscent of the GLRT (Apley and Shi (1999)) or CUSCORE (Box and Ramírez (1992), Luceño (1999), Shu et al. (2002), Runger and Testik (2003)) procedures.

The Effects of Incorrectly Specified Shift Time-of-Occurrence, Shift Size, and Shift Direction

In this section, we demonstrate that, even though the OGLF is optimized under the assumption that the shift occurs on the initial observation, it is just

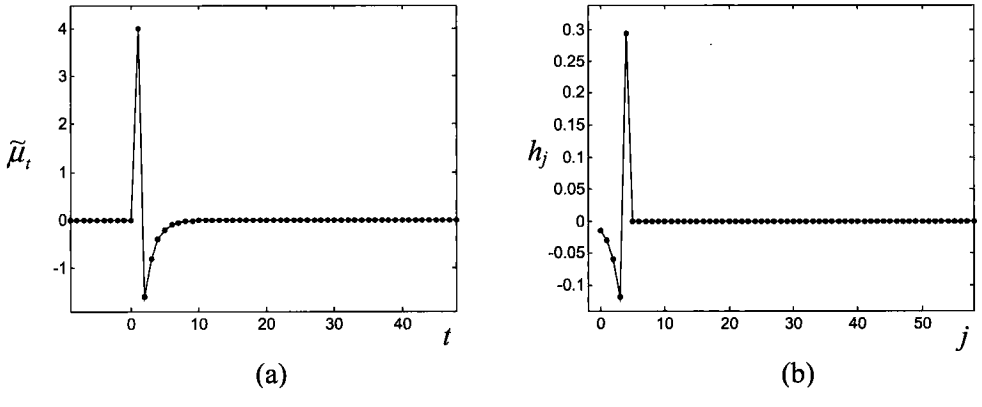


FIGURE 10. Residual Mean (a) and OGLF $\{h_j\}$ (b) for Example 28.

as effective (or very nearly so) at detecting shifts that occur at some later, unknown point in time. We also illustrate why an OGLF optimized for a positive shift ($\mu > 0$) is inherently two sided and its ARL performance for detecting a negative shift ($\mu < 0$) is exactly the same as for detecting a positive shift. Regarding an incorrectly specified shift size, the OGLF will often lack robustness, much like other charts that are tuned for a specific size shift. We briefly discuss strategies for ensuring the effective detection of a range of shift sizes.

Shift Time of Occurrence

The optimization of the OGLF is carried out under the assumption that the shift occurs on the initial observation, and, correspondingly, all of the out-of-control simulation results that we have presented so far were for shifts introduced at the first observation. In practice, shifts generally occur at some later, unknown point in time. A natural question is whether the OGLF performance will suffer if the shift time of occurrence does not coincide with the first observation. Using common control charting terminology, the question is whether the steady-state ARL (the expected detection lag for shifts introduced after the control chart has reached steady state) is substantially worse than the zero-state ARL (the expected detection lag for shifts introduced at the time of the first observation). In the following, we demonstrate that the steady-state performance of the OGLF is very similar to its zero-state performance.

For many control charts designed for detecting step shifts in i.i.d. data (e.g., CUSUM, Shewhart, or EWMA) there is little or no difference between the steady-state and the zero-state ARLs. In contrast, some charts that are designed to detect a signal that varies dynamically (such as the mean of the residuals of an autocorrelated process) can be extremely sensitive to whether the shift occurs on the first observation or at some later point in time. For example, consider the CUSCORE charts for detecting mean shifts in autocorrelated processes that were investigated in Luceño (1999, 2004), Shu et al. (2002), and Runger and Testik (2003). The one-sided upper CUSCORE statistic in these papers is defined recursively via

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

and the chart signals when S_t exceeds a threshold. Luceño (1999) considered two different versions of the CUSCORE chart. In the first version, one takes the residual mean $\tilde{\mu}_t$ to be that which results from a feared mean shift occurring at the initial observation. Consequently, it is tuned to detect a mean shift occurring exclusively at the *initial* observation, and its performance for detecting mean shifts that occur at a later point in time will often be dramatically worse (Shu et al. (2002)). In the second version of the CUSCORE considered in Luceño (1999), $\tilde{\mu}_t$ is reinitialized every time the CUSCORE statistic is reset to zero. We denote the CUSCORE with reinitialization by $S_{r,t}$ and the CUSCORE with no reinitialization by $S_{n,t}$. In essence, the CUSCORE with reinitialization is tuned to detect mean shifts that have occurred immediately after the CUSCORE statistic was last reset to zero. The two-sided CUSCORE versions for detecting both positive and negative shifts would run an upper one-sided CUSCORE in conjunction with a lower one-sided CUSCORE in which $\tilde{\mu}_t$ is replaced by $-\tilde{\mu}_t$. Because the EWMA and OGLF are inherently two-sided, in the following discussion, we consider the two-sided versions of the CUSCORE charts.

Box and Ramírez (1992) discussed a number of CUSCORE strategies, one of which (their centered CUSCORE implemented as a sequence of Wald-like sequential tests), when applied to detecting mean shifts in ARMA processes, is precisely the preceding CUSCORE $S_{r,t}$. Box and Ramírez (1992) also considered the following CUSCORE statistic:

$$C_t = \sum_{j=1}^{t} \tilde{\mu}_j e_j. \tag{10}$$

We will refer to C_t as the running CUSCORE, to distinguish it from the CUSCORES $S_{n,t}$ and $S_{r,t}$. Whereas the Wald-like CUSCORE $S_{r,t}$ of Box and Ramírez (1992) was to be used as a control chart with control limits, they suggested plotting the running CUSCORE C_t versus t as an informal exploratory tool. In order to illustrate a fundamental difference between the CUSCORE procedures and the OGLF, however, we shall consider plotting C_t with symmetric control limits that sound an alarm if $|C_t|$ exceeds some fixed threshold. Regarding $\tilde{\mu}_t$, if one had prior knowledge of the exact point in time at which the shift might occur (or a small set of potential change points), then $\tilde{\mu}_t$ would be the residual mean for a shift occurring at that point in time. In the subsequent discussion, we assume no such prior knowledge is available and we take $\tilde{\mu}_t$ in the running CUSCORE C_t to be the residual mean for a feared shift occurring at the initial observation.

For the situation of Example 20, Table 4 compares the zero-state versus steady-state performances of the three CUSCORE charts ($S_{r,t}$, $S_{n,t}$, and C_t),

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
Running CUSCORE C_t	499.3	1202	1.17	233.6
	(2.89)	(4.52)	(0.00)	(0.74)
CUSCORE $S_{n,t}$ without reinitialization	500.8	655.9	1.45	139.3
	(2.95)	(2.87)	(0.00)	(0.80)
CUSCORE $S_{r,t}$ with reinitialization	500.3	502.3	3.43	3.75
	(2.57)	(2.61)	(0.07)	(0.08)
OGLF	500.4	495.8	3.12	3.44
	(1.94)	(1.86)	(.03)	(.05)
Optimal EWMA	500.7	500.2	76.23	75.97
	(.99)	(1.00)	(.49)	(.49)

the OGLF, and the optimal EWMA. Tables 5 and 6 show analogous results for Examples 16 and 8, respectively. We illustrate with these examples because the residual means have pronounced features (either oscillation or a spike). Consequently, mismatches between the assumed and actual shift times of occurrence will have the greatest potential adverse effect on the CUSCOREs and OGLF. All charts were designed to have a zero-state in-control ARL of 500. The OGLFs and optimal EWMAs are exactly those from Table 3, having been optimized for the specified shifts occurring at the initial observation. The control limits for C_t, $S_{n,t}$, and $S_{r,t}$ in Table 4 were 6.03, 2.93, and 5.29, respectively. Likewise, the control limits for the three CUSCORE charts in Table 5 were 22.95, 4.89, and 4.60, respectively, and the control limits for the three CUSCORE charts in Table 6 were 9.55, 3.99, and 5.17, respectively. Monte Carlo simulation was used to evaluate the ARLs. For the steady-state ARLs in Tables 4 and 6, mean shifts were introduced at observation number 100. For the steady-state ARLs in Table 5, the times of occurrence were randomly drawn from a uniform distribution over the interval [97, 103]. This was to prevent the

TABLE 5. Comparison of Steady-State Versus Zero-State ARLs for Example 16

	$\mu = 0$ (no shift)		$\mu = 3$ (shift added)	
	Zero-state ARL	Steady-state ARL	Zero-state ARL	Steady-state ARL
Running CUSCORE C_t	500.2	421.9	21.75	123.8
	(2.31)	(2.37)	(0.02)	(0.84)
CUSCORE $S_{n,t}$ without reinitialization	499.4	486.4	9.84	80.6
	(2.42)	(2.81)	(0.03)	(0.43)
CUSCORE $S_{r,t}$ with reinitialization	500.9	494.6	11.64	13.22
	(2.1)	(1.95)	(0.03)	(0.03)
OGLF	500.3	490.2	11.64	11.31
	(1.98)	(2.06)	(0.02)	(0.03)
Optimal EWMA	499.8	496.5	26.15	26.00
	(1.98)	(2.17)	(0.11)	(0.11)

TABLE 6. Comparison of Steady-State Versus Zero-State ARLs for Example 8

	$\mu = 0$ (no shift)		$\mu = 3$ (shift added)	
	Zero-state ARL	Steady-state ARL	Zero-state ARL	Steady-state ARL
Running CUSCORE C_t	500.2	484.0	1.66	52.56
	(2.17)	(1.98)	(0.02)	(0.14)
CUSCORE $S_{n,t}$ without reinitialization	499.9	474.7	4.74	25.57
	(1.94)	(2.02)	(0.06)	(0.10)
CUSCORE $S_{r,t}$ with reinitialization	499.8	494.5	8.11	13.86
	(1.91)	(1.88)	(0.05)	(0.08)
OGLF	500.3	488.2	13.72	15.59
	(1.78)	(2.01)	(0.06)	(0.12)
Optimal EWMA	500.1	477.6	29.78	28.21
	(1.72)	(2.34)	(0.05)	(0.11)

coefficients of the CUSCOREs without reinitialization from being either consistently in phase or consistently out of phase with the sinusoidal shift.

The CUSCORE charts without reinitialization (both C_t and $S_{n,t}$) usually detect shifts occurring at the initial observation extremely well. However, they perform miserably for detecting a shift that occurs at some later point in time. Clearly, for processes in which the feared signal has pronounced dynamics, CUSCORE charts without reinitialization should not be used unless one knows a priori when the shift will occur. The CUSCORE $S_{r,t}$ with reinitialization is much less affected by the shift time of occurrence than the CUSCOREs without reinitialization and often performs quite well in the steady state. The OGLF is even less affected by shift time of occurrence. In fact, for Example 8, the OGLF steady-state ARL is slightly better than its zero-state ARL (11.31 versus 11.64). The overall steady-state performances of the reinitialized CUSCORE $S_{r,t}$ and the OGLF appear comparable: In some examples, the CUSCORE performs better; while in others, the OGLF performs better, although the differences are typically not extreme. Design of the OGLF would involve the additional computational task of calculating the OGLF coefficients. Design of the CUSCORE requires a Monte Carlo simulation [or a Markov chain analysis as described in Luceño (1999)] to select the proper control limits. Note that small differences between the zero-state and steady-state *in-control* ARLs in Tables 4–6 are attributable to different initial conditions (initialized at zero, versus what can be viewed as being drawn randomly from the steady-state distribution). The large differences in Table 4 for the CUSCORE charts without reinitialization are due to the shape of the feared signal $\tilde{\mu}_t$. Because $\tilde{\mu}_t$ decays to small values in the steady state for this example, the CUSCORE coefficients become quite small, resulting in a large steady-state in-control ARL.

Given that the OGLF was optimized for a shift occurring at the initial observation, one might find it surprising that it detects shifts occurring in the steady state with virtually the same effectiveness as it detects shifts occurring at the initial observation. Figures 11 and 12 illustrate the reason by contrasting the OGLF with the running CUSCORE C_t, which is a form of linear filter, sometimes referred to as a matched filter. Figure 11 shows a sequence of residuals for the ARMA process of Example 20, with a mean shift added at observation 50. Figure 12(a) shows the running CUSCORE C_t of Equation (10) applied to the residuals of Figure 11. The control limits for C_t were ± 6.03. Figure 12(b) shows the results of the OGLF for the same example, with the filter coefficients $\{h_j\}$ as given in Example 20 of Appendix C. The remaining panels of the figure illustrate the calculation of the chart statistics for specific values of the time index. Figure 12(c) illustrates the calculation of the running CUSCORE C_t at $t = 15$ by showing the first 15 residuals and the first 15 coefficients of $\{\tilde{\mu}_j\}$. The CUSCORE summation in Equation (10) for $t = 15$ is the sum of the product of these two signals, which is the value $C_{15} = 3.19$ plotted in panel (a). Similarly, Figure 12(d) illustrates the cal-

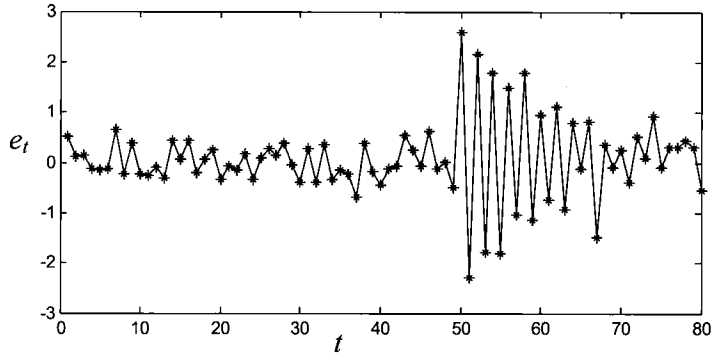


FIGURE 11. Residuals from Example 20 with a Mean Shift Added at Observation 50. σ_a Has Been Reduced to More Clearly Illustrate the Oscillating Residual Mean.

culation of the OGLF chart statistic for observation 15 by showing the same 15 residuals and the first 15 coefficients $\{h_{15-j} : j = 1, 2, \ldots, 15\}$ of the OGLF in reverse order. The OGLF summation in Equation (2) for $t = 15$ is the sum of the product of these two signals, which is the value $y_{15} = -0.03$ plotted in panel (b). The other panels on the left show the calculations for the CUSCORE for other times, while those on the right show the corresponding calculations for the OGLF. Both charts have control limits that are symmetric about zero. Hence, either positive or negative excursions of the charted statistic may trigger an alarm.

A key point of Figure 12 is that the OGLF is a *time-invariant* linear filter (its coefficients slide forward one time period as each new observation is obtained), whereas the running CUSCORE C_t is not. Because the coefficients of the time-invariant OGLF slide across the residuals as each new observation is obtained, the OGLF is able to detect shifts that occur in the steady state nearly as well as those that occur on the initial observation. The right panels of Figure 12 illustrate this. The OGLF at time $t = 53$, for example, is the summation of the product of the two signals $\{e_j\}$ and $\{h_{53-j}\}$ shown in Figure 12(h). Because these two signals have high negative correlation, $y_{53} = -1.31$ is large and negative and falls below the LCL. For the OGLF at time $t = 56$ illustrated in Figure 12(j), the OGLF coefficients are shifted over three time periods relative to those for $t = 53$. Because the correlation between the two signals $\{e_j\}$ and $\{h_{56-j}\}$ shown in Figure 12(j) is large and positive, the OGLF value $y_{56} = 1.46$ is large and positive and exceeds the UCL. This also explains the oscillatory behavior of the OGLF statistic seen in Figure 12(b).

In contrast, the CUSCORE is not a *time-invariant* linear filter (although it is a time-varying linear filter), so that its coefficients do not slide across the residuals as each new observation is taken. The first CUSCORE coefficient $\tilde{\mu}_1$ always multiplies the first residual e_1 in Equation (10), and similarly for the remaining coefficients and residuals. As illustrated in the left panels of Figure 12, by the time the shift occurs at $t = 50$, there is very little correlation between the CUSCORE coefficients and the residuals. Hence, the running CUSCORE C_t has difficulty detecting shifts that occur in the steady state.

Although one could attempt to optimize the OGLF under the steady-state scenario, this would be substantially more computationally expensive than our approach of optimizing under the zero-state scenario. The preceding discussion indicates that even if one is more interested in steady-state performance measures, it is sufficient to use the simpler approach of optimizing under the zero-state scenario.

Although none of the versions of CUSCORE charts that we have discussed satisfy the definition of a time-invariant linear filter, Box and Luceño (1997, pp. 245–246) present a different version, in which one charts $\sum_{j=t-k}^{t} \tilde{\mu}_{j-(t-k)+1} e_j$, which is a time-invariant linear filter. One must select a specific value for k, and the interpretation is that at each time period t, one looks for a mean shift that may have occurred at the specific time $t - k$. This is similar to the GLRT of Vander Wiel (1996) and Apley and Shi (1999), except that for the GLRT, one looks for shifts that may have occurred at each time $\{t - N, t - N + 1, \ldots, t\}$ within a moving window of length $N + 1$ time periods. For Examples 25–28 (see the earlier discussion surrounding Figure 10) and 10–

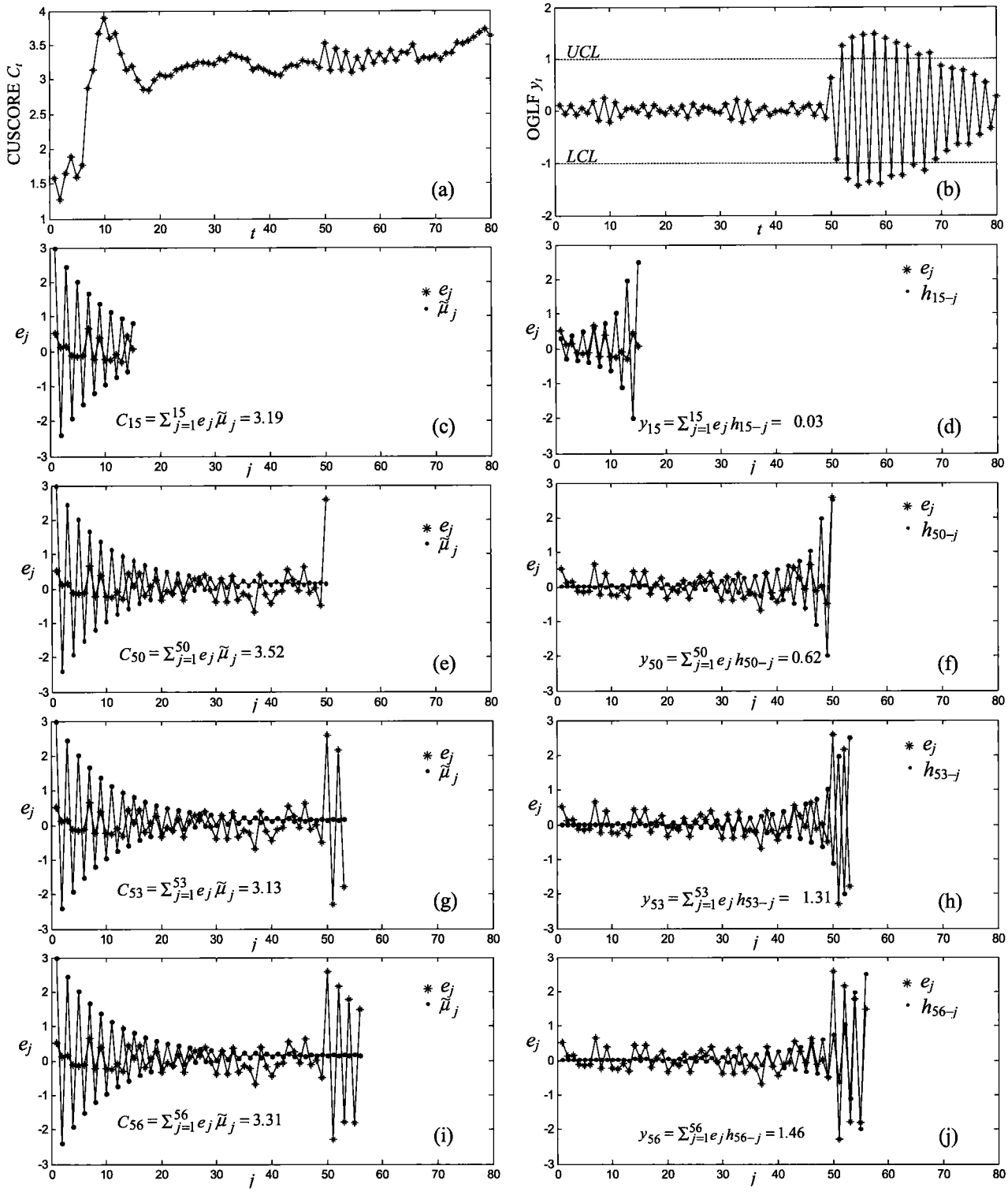


FIGURE 12. Illustration of the Difference Between the OGLF (Right Panels) and the Running CUSCORE C_t (Left Panels) Applied to the Residuals Shown in Figure 11. The OGLF Is a Time-Invariant Linear Filter that Moves across the Residuals and Is Thus Able to Detect Shifts that Occur at Times for Which It Was Not Explicitly Optimized. The OGLF $\{h_j\}$ Have Been Scaled for Visual Convenience.

12, the OGLF was nearly identical to the Box and Luceño CUSCORE with $k = 4$ and $k = 1$, respectively. For any example in which the OGLF reduces to a Shewhart individual chart, the OGLF coincides trivially with the Box and Luceño CUSCORE with $k = 0$. For most of the other examples that we have considered, however, the OGLF was very different from the Box and Luceño CUSCORE for any choice of k. The fact that the OGLF can be very different from the Box and Luceño CUSCORE is quite clear from the OGLF $\{h_j\}$ plotted in Appendix C for Examples 17–20. Because the only difference between these four examples is the magnitude of the shift, the CUSCORE coefficients $\{\tilde{\mu}_j\}$ would be identical in all four examples (except for a scale factor, which can be absorbed into the control limits). In contrast, the OGLF coefficients $\{h_j\}$ are markedly different in all four examples.

Shift Direction

An OGLF optimized for detecting a shift of size $+\mu$ has identical performance for detecting a shift of size $-\mu$ in the opposite direction. This follows by the symmetry of the control limits (± 1) and the antisymmetry of Equation (2) with respect to the residuals. Hence, the OGLF is inherently two sided. This two sidedness also improves performance in situations in which the residual mean oscillates. For example, the OGLF in Figure 12(b) almost signals first at $t = 51$ because of violating the lower control limit. Although we omit the results here, we have observed a similar phenomenon for the CUSCORE with reinitialization. For Examples 16 and 20, the one-sided version of the CUSCORE S_t with reinitialization did not perform nearly as well as the two-sided version documented in Tables 4 and 5.

Unknown Shift Size

The optimization of the OGLF was also carried out for a specific shift size μ. Because shift sizes are generally not known a priori, we would like the chart to be effective at detecting a range of shift sizes. In this situation, one might select a single shift size of particular interest, perhaps somewhere in the middle to the lower end of the range of values that one is interested in detecting, and optimize the OGLF based on that. This is analogous to the common CUSUM design practice of selecting the CUSUM parameters to be optimal for a single shift size of particular interest. For certain types of processes, this may provide reasonable performance over a range of shift sizes. Consider the example from Table 2. If we had optimized the OGLF for the small- to medium-sized mean shifts ($\Delta = 0.5$ or 1.0), its performance would also be quite reasonable for the larger shifts ($\Delta = 2.0$ and 3.0), as is evident from Table 2.

As another example, consider the two sinusoids S_3 and S_4 from Examples 15 and 16, respectively, which differ only in amplitude (magnitude). The amplitude of S_3 is half that of S_4. In spite of this, additional Monte Carlo simulation reveals that the OGLF optimized for S_3 has almost as short an ARL for detecting S_4 as does the OGLF optimized for S_4 (11.68 versus 10.61). The OGLF optimized for S_4 has an ARL of 45.53 for detecting S_3 (compared with an ARL of 32.90 for the OGLF optimized for S_3). In fact, the OGLF $\{h_j\}$ for Examples 15 and 16 are quite similar, as seen in Appendix C. Both appear to be exponentially decaying sinusoids of the same frequency as the anticipated sinusoid, and they only differ in their rate of decay. Consequently, they both provide reasonable detection of the two sinusoids with different amplitudes.

This suggests an approach for ensuring effective detection of a range of shift sizes: Calculate the OGLF for a few different shift sizes that span the range of interest, plot their impulse response coefficients as in Appendix C, and inspect them to see how similar they are. If they are all quite similar (e.g., the OGLFs for Examples 9–12 or the OGLFs for Examples 25–28), then we can conclude that using the OGLF designed for a midrange-size shift will provide reasonable performance for the other shift sizes.

On the other hand, suppose that the process is such that the OGLFs differ substantially for different shift sizes. For example, the OGLFs for step shifts in i.i.d. processes (Examples 1–4) are quite different. This is not surprising, given that the OGLFs for step shifts in i.i.d. processes are close to EWMAs, and EWMAs tuned to detect small (large) shifts perform very poorly for detecting large (small) shifts. In this situation, we recommend using multiple simultaneous OGLFs, each one optimized for a different shift size. Similar to what was suggested by Sparks (2000), who discussed the design of multiple simultaneous CUSUMs, we suggest that each of the individual OGLF charts be designed to have the same in-control ARL. The ARL of the individual charts should be chosen somewhat larger than the overall desired in-control ARL. Monte Carlo simulation can be used at the final stage to scale up/down all OGLF coefficients, so that the combined procedure has the correct desired in-control ARL, similar to what is de-

scribed as the last step of the optimization algorithm in Appendix B.

A potential alternative to running parallel OGLFs would be to use a linear combination of two or more OGLFs, each optimized for a different size shift (reminiscent of the OGLF in Figure 4, which can be viewed as a linear combination of an EWMA and a Shewhart chart). Another potential alternative would be to use a single OGLF that minimizes some weighted combination of the ARLs for a set of mean shifts, perhaps a form of expected loss function with prior probabilities assigned to each shift size. This topic warrants further research.

Conclusions

In this paper, we have treated control charting as an optimal filter-design problem. We considered control chart statistics that are the output of a general time-invariant linear filter applied to the data, and we directly optimized the filter impulse response coefficients to minimize the out-of-control ARL for a specified mean shift, while constraining the in-control ARL to a desired value. The primary focus of the paper was on efficiently performing the optimization and on interpreting the characteristics of the OGLFs, which often assume shapes that are unusual but intuitively reasonable.

One conclusion is that, for step mean shifts in i.i.d. data, the OGLF happens to coincide with a simple EWMA filter. This is not surprising given the large body of empirical evidence indicating 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*. Our observations add to the large body of evidence supporting the view that it is difficult to improve on the performance of an EWMA for detecting step mean shifts in i.i.d. data.

The OGLF may be quite different from, and also perform much better than, an optimized EWMA in other situations, in particular when the data are autocorrelated. For step mean shifts in AR(1) processes with $\phi \approx 0.9$, which are relatively common models for industrial process data, the OGLF relates closely to a combined Shewhart–EWMA procedure. For situations in which the residual mean experiences pronounced dynamics, the OGLF $\{h_j\}$ tend to inherit some of the shape characteristics of the residual mean. In these situations, the OGLF typically has the largest performance advantage over an optimized EWMA.

Appendix A
Expressions for Q_p and Its Derivative

This appendix provides expressions for the elements of the state-transition matrix $\mathbf{Q}_p$ and their derivatives with respect to the GLF coefficients used in Equation (7). From Equations (4) and (5), each element of $\mathbf{Q}_p$ is given by

$$\begin{aligned} Q_p^{ij} &= F\left(\frac{\sigma_{p-1}^2\left(\alpha_j + \frac{\delta}{2} - \mu_{p,y}\right) - v_p(\alpha_i - \mu_{p-1,y})}{\sigma_{p-1}(\sigma_{p-1}^2\sigma_p^2 - v_p^2)^{1/2}}\right) \\ &\quad - F\left(\frac{\sigma_{p-1}^2\left(\alpha_j - \frac{\delta}{2} - \mu_{p,y}\right) - v_p(\alpha_i - \mu_{p-1,y})}{\sigma_{p-1}(\sigma_{p-1}^2\sigma_p^2 - v_p^2)^{1/2}}\right) \\ &= F(c_{i,j,p}) - F(d_{i,j,p}), \end{aligned}$$

where $F(\cdot)$ denotes the standard normal cumulative distribution function and $c_{i,j,p}$ and $d_{i,j,p}$ denote the arguments in parentheses. The derivative of Q_p^{ij} with respect to h_r is therefore

$$\frac{\partial Q_p^{ij}}{\partial h_r} = f(c_{i,j,p})\frac{\partial c_{i,j,p}}{\partial h_r} - f(d_{i,j,p})\frac{\partial d_{i,j,p}}{\partial h_r},$$

where $f(\cdot)$ denotes the standard normal density. Expressions for $\partial c_{i,j,p}/\partial h_r$ and $\partial d_{i,j,p}/\partial h_r$ can be calculated based on the definitions of the various quantities preceding Equation (4).

Appendix B
Some Details of the Optimization Algorithm

The constrained optimization problem is to select the GLF coefficients $\{h_j : j = 0, 1, \ldots\}$ that minimize $\mathrm{ARL}(h_0, h_1, \ldots \mid \mu_t)$ subject to the constraint $\mathrm{ARL}(h_0, h_1, \ldots \mid 0) = \mathrm{ARL}_d$, where $\mathrm{ARL}(\cdot \mid \mu_t)$ denotes the ARL of the GLF chart for a process with mean $\{\mu_t : t = 1, 2, \ldots\}$ and ARL_d denotes the desired in-control ARL. We use an iterative search procedure with search direction determined via the gradient projection method (Luenberger (1984)) depicted in Figure B.1.

The steps to determine the locus of points (in the space of GLF coefficients) over which we search at iteration number i $(i = 1, 2, \ldots)$ are outlined as follows.

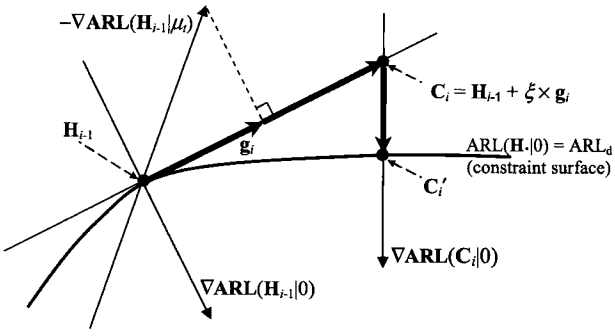


FIGURE B.1. Depiction of the Gradient Projection Search Procedure at Iteration i.

1. Compute the gradients $\nabla\mathbf{ARL}(\mathbf{H}_{i-1} \mid \mu_t)$ and $\nabla\mathbf{ARL}(\mathbf{H}_{i-1} \mid 0)$ using Equation (7), where $\mathbf{H}_{i-1}$ denotes the vector of GLF coefficients at iteration $i-1$.
2. Compute the initial search direction $\mathbf{g}_i$ as the projected gradient

$$\mathbf{g}_i = -\nabla\mathbf{ARL}(\mathbf{H}_{i-1} \mid \mu_t) + \left(\frac{\nabla\mathbf{ARL}(\mathbf{H}_{i-1} \mid \mu_t) \cdot \nabla\mathbf{ARL}(\mathbf{H}_{i-1} \mid 0)}{|\nabla\mathbf{ARL}(\mathbf{H}_{i-1} \mid 0)|^2}\right) \times \nabla\mathbf{ARL}(\mathbf{H}_{i-1} \mid 0).$$

3. Define a step size ξ and compute a new candidate set $\mathbf{C}_i$ of GLF coefficients via

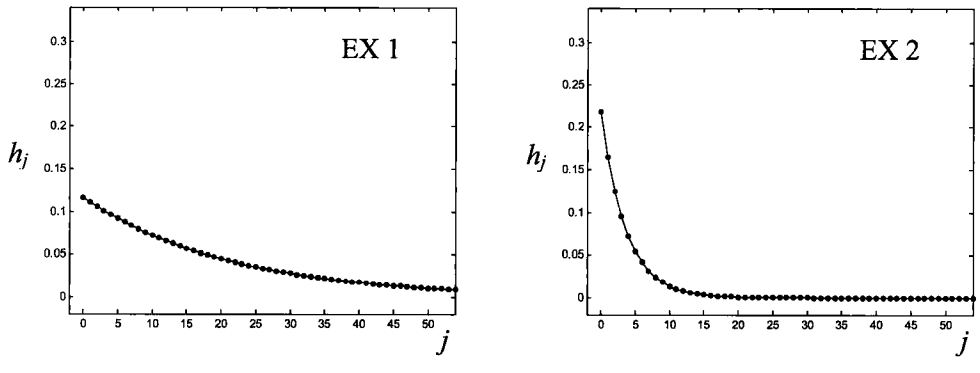
$$\mathbf{C}_i = \mathbf{H}_{i-1} + \xi \times \mathbf{g}_i.$$

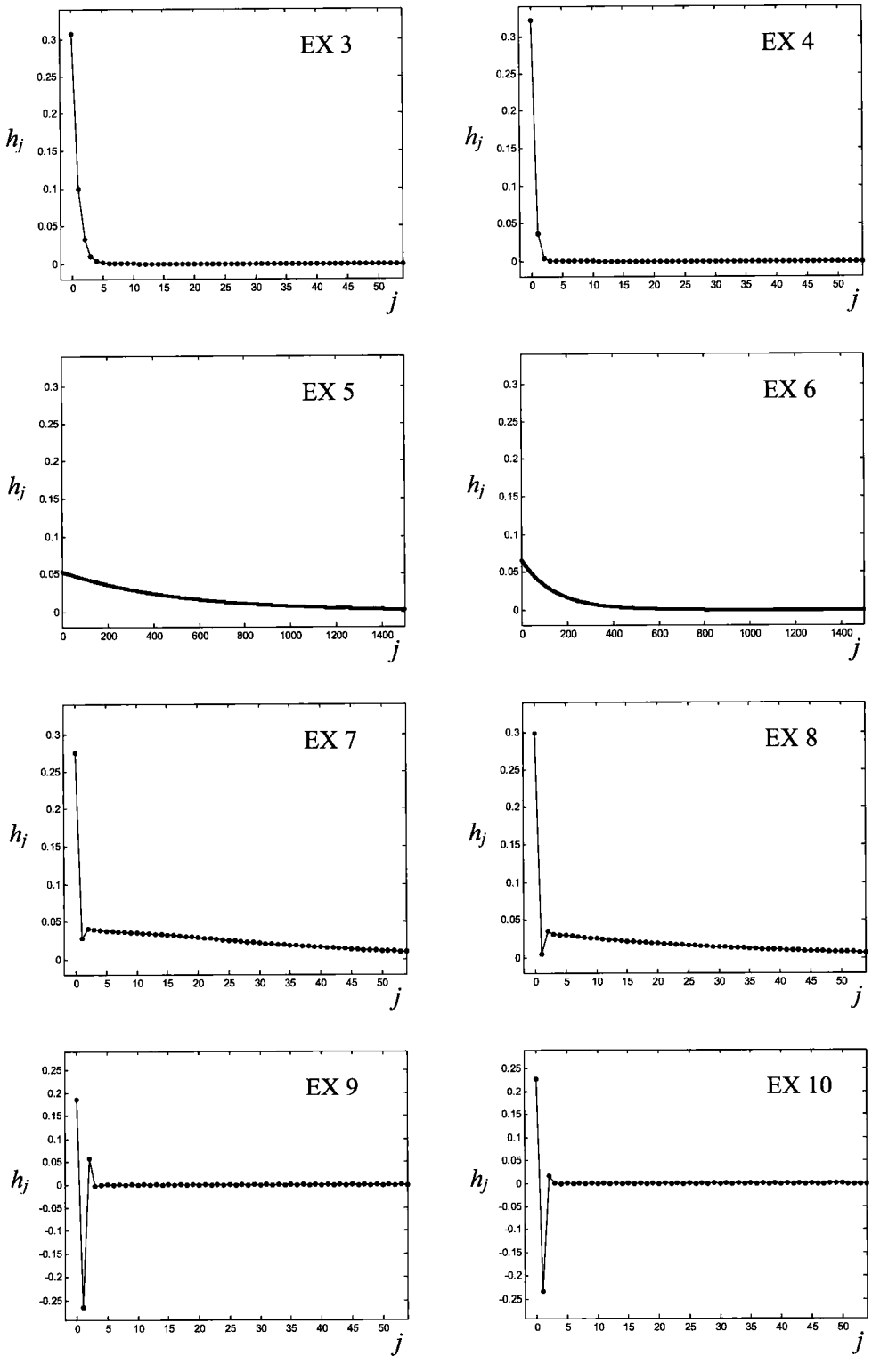
   Because ARL($\mathbf{C}_i \mid 0$) will not necessarily equal ARL$_d$, we must next do the following.
4. Iteratively search along the gradient direction $\pm\nabla\mathbf{ARL}(\mathbf{C}_i \mid 0)$ for a new candidate set $\mathbf{C}_i'$ of GLF coefficients until we satisfy the constraint ARL($\mathbf{C}_i' \mid 0$) = ARL$_d$ (see Figure B.1).
5. Repeat Steps 3 and 4 for a set of step sizes (say $\xi/2$, $\xi/4$, etc.), producing a set of $\mathbf{C}_i'$ along the feasible search path, each satisfying ARL($\mathbf{C}_i \mid 0$) = ARL$_d$.
6. Evaluate ARL($\mathbf{C}_i \mid \mu_t$) for each $\mathbf{C}_i'$ from Step 5 and choose $\mathbf{H}_i$ to be the $\mathbf{C}_i'$ that minimizes ARL($\mathbf{C}_i \mid \mu_t$).

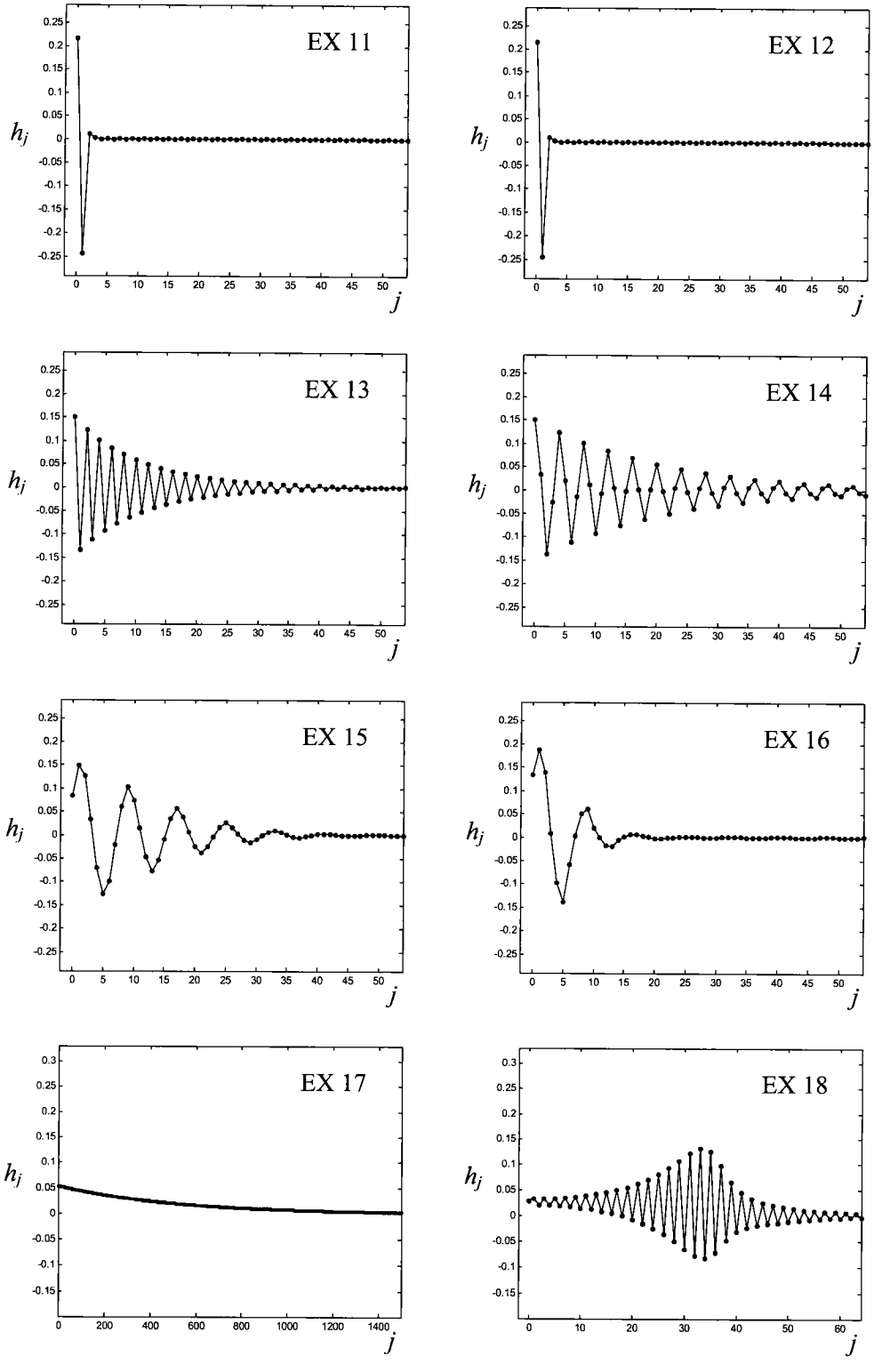
Steps 1–6 constitute a single iteration of the search procedure in the optimization algorithm. The iterations are repeated until no further reduction in ARL($\mathbf{H}_i \mid \mu_t$) is achieved (per some suitable stopping criterion), at which point the optimization algorithm terminates. Because the Markov chain approach for calculating the ARL provides only a rough approximation, the final step is to scale all GLF coefficients by a single constant (i.e., $h_j \to ch_j$ for $j = 0, 1, \ldots$ and some constant c) so that the in-control ARL equals ARL$_d$ according to a Monte Carlo simulation with a sufficiently large number of replicates.

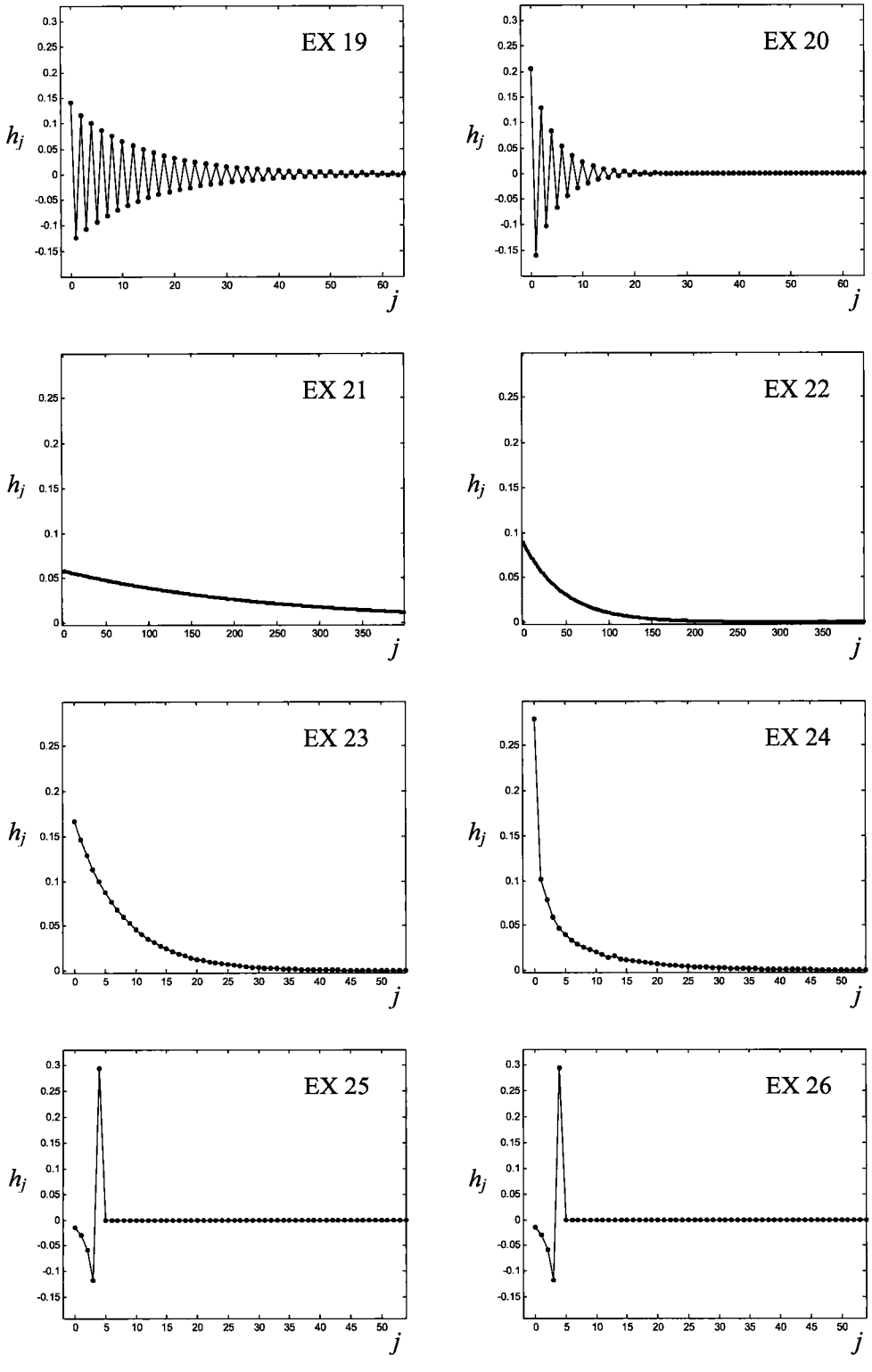
In order to select the filter truncation time Tr, we recommend the following simple procedure. First, select a tentative value of $Tr = 100$ and find the OGLF for that value. Then inspect the OGLF coefficients to determine whether h_j has decayed to negligible values for $j \approx Tr$. If so, keep the OGLF for $Tr = 100$. If not, increase Tr to a larger value based on when it appears the OGLF coefficients may decay to negligible values, and repeat. Note that $Tr = 100$ was sufficient for all examples shown in Appendix C, except Examples 5, 6, 17, 21, and 22. Note also that the OGLF closely resembles an EWMA for these examples, which we anticipate will be the case whenever large Tr values are required. Consequently, as an alternative to using an OGLF with an extremely large Tr value, we recommend simply using an EWMA.

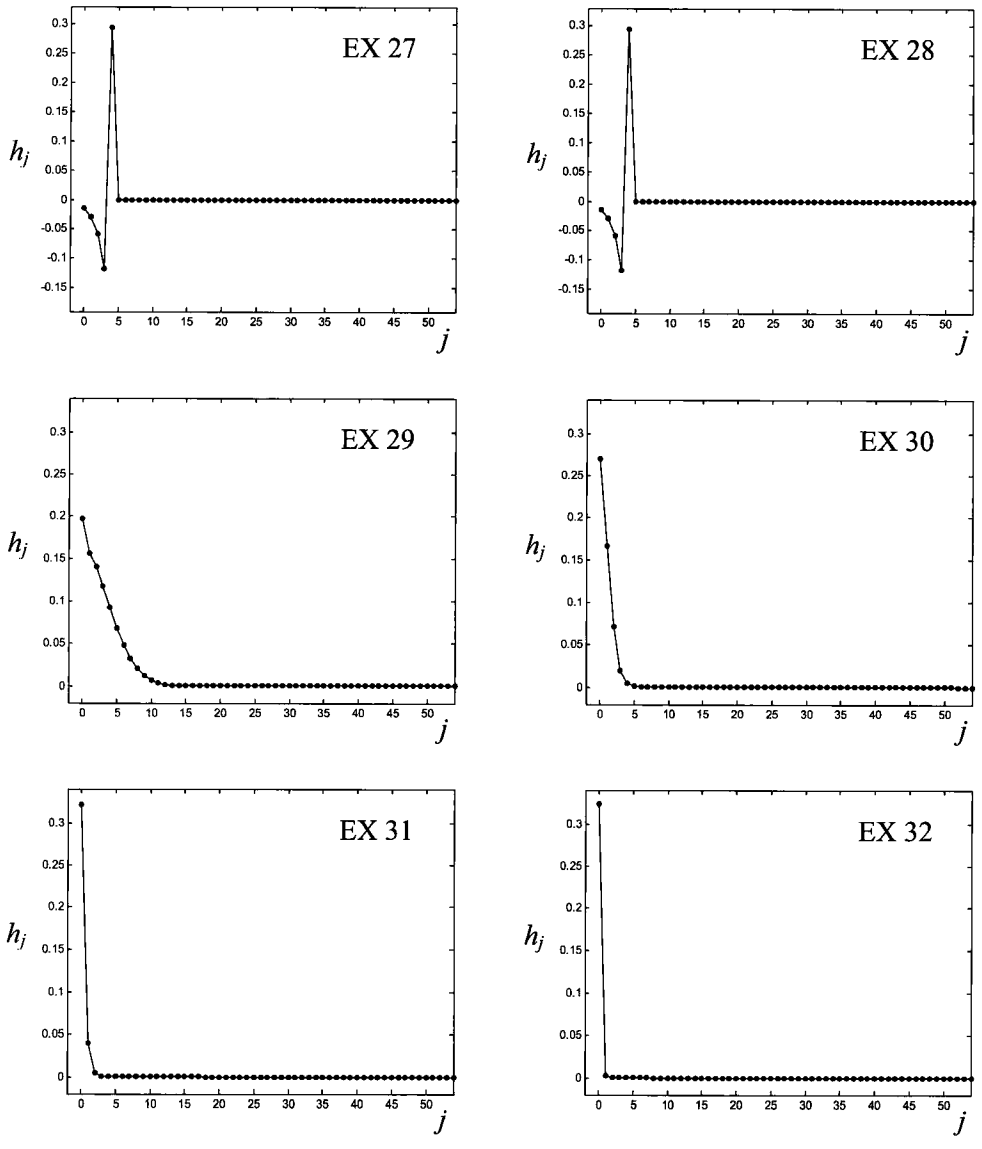
Appendix C
Illustration of the OGLFs for the 32 Examples in Table 3











References

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

BOX, G. E. P.; JENKINS, G.; and REINSEL, G. (1994). *Time Series Analysis, Forecasting, and Control*, 3rd ed., Prentice-Hall, Englewood Cliffs, NJ.

BOX, G. E. P. and LUCEÑO, A. (1997). *Statistical Control by Monitoring and Feedback Adjustment*. Wiley, New York, NY.

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

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

CHIN, C. (2004). "Optimal Filter Design Approaches to Statistical Process Control for Autocorrelated Processes". Unpublished Ph.D. Dissertation, Texas A&M University.

CHIN, C. and APLEY, D. W. (2006). "Optimal Design of Second-Order Linear Filters for Control Charting". *Technometrics* 48(3), pp. 337–348.

JIANG, W.; TSUI, K. and WOODALL, W. H. (2000). "A New SPC Monitoring Method: The ARMA Chart". *Technometrics* 42, pp. 399–410.

JIANG, W.; WU, H.; TSUNG, F.; NAIR, V. N.; and TSUI, K. (2002). "Proportional Integral Derivative Charts for Process Monitoring". *Technometrics* 44, pp. 205–214.

JOHNSON, R. A. and WICHERN, D. W. (2002). *Applied Multivariate Statistical Analysis*, 5th ed. Prentice Hall, Upper Saddle River, NJ.

LU, C. and REYNOLDS, M. R., JR. (1999a). "EWMA Control Charts for Monitoring the Mean of Autocorrelated Processes". *Journal of Quality Technology* 31, pp. 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, pp. 259–274.

LUCAS, J. M. and SACCUCCI, M. S. (1990). "Exponentially Weighted Moving Average Control Schemes: Properties and Enhancements". *Technometrics* 32, pp. 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, pp. 177–195.

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

LUENBERGER, D. G. (1984). *Linear and Nonlinear Programming*, 2nd ed. Addison-Wesley, Reading, MA.

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

PANDIT, S. M. and WU, S. M. (1983). *Time Series and System Analysis with Applications.* John Wiley & Sons, New York, NY.

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, pp. 181–205.

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

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

SHU, L.; APLEY, D. W.; and TSUNG, F. (2002). "Autocorrelated Process Monitoring Using Triggered CUSCORE Charts". *Quality and Reliability Engineering International* 18, pp. 411–421.

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

SUPERVILLE, C. R. and ADAMS, B. M. (1994). "An Evaluation of Forecast-Based Quality Control Schemes". *Communications in Statistics: Simulation and Computation* 23, pp. 645–661.

VANDER WIEL, S. A. (1996), "MONITORING PROCESSES THAT WANDER USING INTEGRATED MOVING AVERAGE MODELS". *Technometrics* 38, PP. 139–151.

WARDELL, D. G.; MOSKOWITZ, H.; and PLANTE, R. D. (1994). "Run-Length Distributions of Special-Cause Control Charts for Correlated Processes". *Technometrics* 36, pp. 3–17.

ZHANG, N. F. (1998). "A Statistical Control Chart for Stationary Process Data". *Technometrics* 40, pp. 24–38.