통계적 공정 관리를 위한 일반 선형 필터의 최적 설계

진창호* • Daniel W. Apley**
* 경희대학교 기계 • 산업시스템공학부
** Associate Professor, Industrial Engineering and Management Department,
Northwestern University

Optimal Filter Design Approach to Statistical Process Control

Chang-Ho Chin* · Daniel W. Apley**
* School of Mechanical and Industrial Systems Engineering, Kyung Hee University
* Industrial Engineering and Management Sciences, Northwestern University, U.S.A.

Abstract: Many control charting methods for both i.i.d and autocorrelated data can be viewed as charting the output of a linear filter applied to the process data. We propose a generalization of this concept, in which the filter parameters are optimally selected to minimize the out-of-control ARL while constraining the in-control ARL to some desired value. A number of interesting characteristics of the optimal filters are discussed.

1. Introduction

To this date, no control chart consistently outperforms the others, because the performance of control charts is substantially influenced by the original process. Many design methodologies have been proposed to optimally select the sample size, sampling interval, control limits, and parameters of the control chart according to the underlying process. In most researches, however, the basic structure of the control chart has not been a subject of investigation. As a result, the inherent characteristics of the charted statistics generated by the fixed structure of control charts have limited improvement in control chart design and performance. To minimize the limitation, therefore, we propose a control charting scheme, which we call the general linear filter (GLF), based on generalizing the concept of linear filters for control charts that many control charting schemes for both independently, identically distributed (i.i.d.) and autocorrelated data can be viewed as charting the output of a linear filter applied to the process data. This generalization is one of the main contributions of this research. In addition, an optimal design methodology for the proposed control charting scheme is developed.

To explain the linear filtering of control charts, let $y_t = \mathrm{H}(B)x_t$ denote the charted statistic, where t is a time index; x_t is the original process data; and $\mathrm{H}(B) = h_0 + h_1B + h_2B^2 + \ldots$ is a linear filter in impulse response form with B denoting the time-series backshift operator. Two simple examples of this are a Shewhart individual chart and an EWMA chart on x_t. For the shewhart chart, $y_t = x_t$, with $\mathrm{H}(B) = 1$ as 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 $\mathrm{H}(B) = (1 - (1 - \lambda)B)$-1 $\lambda = \lambda + \lambda(1 - \lambda)B + \lambda(1 - \lambda)^2B^2 + \ldots$. Residual-based Shewhart and Exponentially Weighted Moving Average (EWMA) charts can be viewed similarly if x_t is assumed to follow an Autoregressive

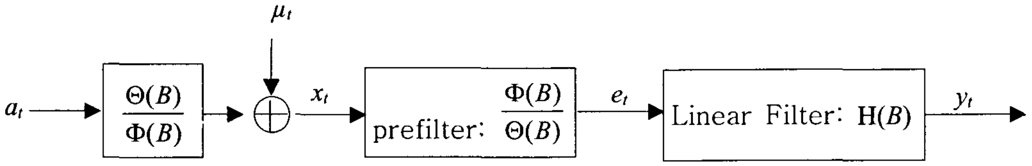
μ_t
a_t
$\frac{\Theta(B)}{\Phi(B)}$
x_t
prefilter: $\frac{\Phi(B)}{\Theta(B)}$
e_t
Linear Filter: H(B)
y_t

Figure 1. Block Diagram Representation of a Linear Filtering Operation.

Table 1. Control Charts Based on Linear Filtering

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

Note: In the PID chart, x_t and y_t are a disturbance and a PID-based residual, respectively.

Moving Average (ARMA) process model, plus (potentially) an additive deterministic mean shift, μ_t of the form

$$x_t = \frac{\Theta(B)}{\Phi(B)} a_t + \mu_t, \qquad (1)$$

where t is a time index; a_t is an i.i.d. Gaussian process with mean 0 and variance σ_a^2 denoted $a_t \sim NID(0, \sigma_a^2)$; $\Phi(B) = (1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)$ 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. $\mu_t = 0$ for the in-control process and $\mu_t \neq 0$ for the out-of-control process. The model residuals (i.e., the one-step-ahead prediction errors) are generated via the linear filtering operation

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

$$= a_t + \frac{\Phi(B)}{\Theta(B)} \mu_t = a_t + \tilde{\mu}_t, \qquad (2)$$

where $\tilde{\mu}_t = \Phi(B)/\Theta(B)\mu_t$ is just the filtered version of the deterministic mean shift μ_t. We may view this $\Phi(B)/\Theta(B)$ in Equation (2) as a linear prefilter to the Shewhart or EWMA filter, as shown in Figure 1 and Table 1. Table 1 also includes the Proportional Integral Derivative (PID) chart of Jiang et al. (2002), which reduces to a third-order filter on x_t without a prefilter. The ARMA(1,1) chart of Jiang et al. (2000) is a first-order filter on x_t with no prefilter.

With the whitening prefilter, therefore, the dynamic structure of control charts can be generally expressed by the following model

$$\begin{aligned} y_t &= \mathrm{H}(B)e_t \\ &= h_0 e_t + h_1 e_{t-1} + h_2 e_{t-2} + \ldots\ldots + h_{Tr} e_{t-Tr} \\ &= \sum_{j=0}^{Tr} h_j e_{t-j} \end{aligned}, \qquad (3)$$

where H(B) is the GLF in design and Tr is a truncation time large enough to approximate $h_j \cong 0$ for $j > Tr$. We use the residual-based model in Equation (3) because it is more convenient to work with and there is no loss of generality if the ARMA model is stable and invertible.

Based on the model in Equation (3), in this paper, we treat the optimal design problem of control charts as an optimal filter design problem. The impulse response coefficients of the GLF are selected to minimize the out-of-control Average Run Length (ARL) subject to the in-control ARL, equaling some specified value. Section 2 discusses the calculation of ARL for the GLF based on the Markov chain method and the gradient-based numerical optimization strategy. In Section 3, performance comparison between the optimal general linear filters (OGLF) and other control charts is given. Section 4 presents the main conclusions of this research.

2. Optimization Strategy for Filter Design

We use a gradient-based numerical optimization strategy, which requires the calculation of the ARL and its derivative. The Markov chain approach (Brook and Evans 1972) is used to compute the ARL. Since the y_t in Equation (3) does not have the Markov property, we approximate the distribution of y_t as that of a one-dimensional Markov process:

$$f_{y_t|y_{t-1}}(s_t|s_{t-1}) \cong f_{y_t|y_{t-1},y_{t-2},\ldots}(s_t|s_{t-1},s_{t-2},\cdots) \quad (4)$$

where s_t is a specific state at timestep t and f is the conditional probability distribution function of y_t given the previous state(s). The approximation of the Markov property of the charted statistic y_t causes some discrepancy between the approximated ARL and the actual one. However, as will be demonstrated in the results presented later in this paper, the Markov approximation still provides a reasonable relative ARL comparison between two different filters in the optimization procedure. We also use Monte Carlo simulation when more precise ARL calculations are require, such as to guarantee that the final OGLF really does have 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

$$\begin{bmatrix} y_t \\ y_{t-1} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_{t,y} \\ \mu_{t-1,y} \end{bmatrix}, \begin{bmatrix} \sigma_t^2 & v_t \\ v_t & \sigma_{t-1}^2 \end{bmatrix}\right) , \quad (5)$$

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 of 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. Then, the conditional distribution of y_t given y_{t-1} is (Johnson and Wichern 1998)

$$N\left(\hat{\mu}_t + \frac{v_t(y_{t-1} - \mu_{t-1,y})}{\sigma_{t-1}^2}, \sigma_t^2 - \frac{v_t^2}{\sigma_{t-1}^2}\right) . \quad (6)$$

We set the control limits for y_t at ±1 without loss of generality, because all of the impulse response coefficients of the GLF will be scaled by the same value to account for this. The in-control region (y_t inside the ±1 interval) is discretized into N equal subintervals of length $\delta = 2/N$, and the out-of -control regions are treated as a single absorbing state. In Figure 2, A_j indicates the subinterval for state j, and a_j = LCL + (j − 1/2)δ is the midpoint of A_j. For the Markov chain approach, the i^{th} row, j^{th} column element ($1 \leq i,j \leq N$) of the transition probability matrix at time t for the nonabsorbing states, denoted Q_t^{ij}, is defined as

$$Q_t^{ij} = Pr\{y_t \in A_j \mid y_{t-1} = a_i\}. \quad (7)$$

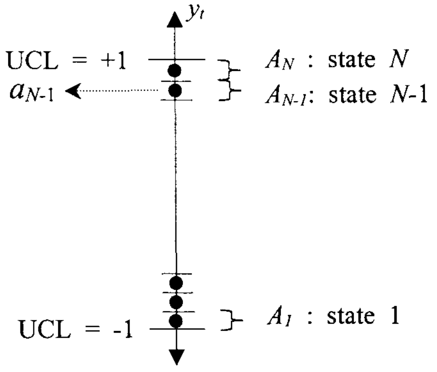

Figure 2. One-dimensional State Space Discretized for the Markov Chain Approach.

Following the analytical expressions of Apley and Chin (2004), in this paper, the ARL and its derivative with respect to filter coefficient h_j are approximated as

$$\text{ARL} = \sum_{p=1}^{m-1} b_p \underline{1} + b_m[I - Q]^{-1}\ \underline{1} \quad \text{and} \quad (8)$$

$$\frac{\partial ARL}{\partial h_j} = \sum_{p=1}^{m-1} b_p \frac{\partial Q_p}{\partial h_j} c_p + b_m [I - Q]^{-1} \frac{\partial Q}{\partial h_j} c_m , \quad (9)$$

where m is a sufficiently large integer such that Q_t approaches a steady state value $Q \cong Q_m \cong Q_{m+1} \cong \cdots$; $b_p = \underline{\pi}_0 \prod_{l=1}^{p-1} Q_l = b_{p-1}Q_{p-1}$ with initial condition $b_1 = \underline{\pi}_0$, and c_p = [I + Q_{p+1} + $Q_{p+1}Q_{p+2}$ +⋯] $\underline{1}$ = $\underline{1}$ + $Q_{p+1}c_{p+1}$ with initial condition c_m = [I + Q + QQ +⋯] $\underline{1}$ = [I −

$Q]^{-1}\ \underline{1}$ can be calculated recursively, respectively.

In the optimal design procedure, the impulse response coefficients of the GLF are determined to optimally detect a specified mean shift for the underlying process. The information required to implement the optimization algorithm includes the ARMA model for the underlying process, the magnitude and type (e.g., a constant step shift, ramp shift, sinusoid, etc.) of the mean shift of particular interest, a reasonable starting point of the optimization search such as the Shewhart chart or the EWMA chart, and the desired in-control ARL. The optimization search starts from the user-specified starting point and continues in the direction of the gradient to reduce the out-of-control ARL until it reaches an optimal solution. Since the optimization algorithm has numerous filter coefficients to search, the utilization of the gradient information improves the optimization routine remarkably. Our analytical expressions in Equations (8) and (9) for the calculation of the ARL and its derivatives based on the approximation of the Markov property in Equation (4) substantially reduce the computational time, thereby facilitating the practical implementation.

3. Performance Improvement over the Optimal EWMA

In this section, the residual-based EWMA chart with control limits ±1 is defined as

$$y_t = (1 - \lambda)y_{t-1} + ke_t, \qquad (10)$$

where $0 < \lambda \leq 1$ is a constant; k is an EWMA scaling constant; and the residual e_t is the filtered version of x_t as shown in Equation (2). This section compares the performance of the optimal EWMA with the OGLF to show how much the charting performance is improved by enhancing the design flexibility – the design degree of freedom in the filter design. Each impulse response coefficient of the GLF is individually selected, whereas the impulse response of the EWMA is determined by only one parameters, λ because k is adjusted to provide the desired in-control ARL. In this sense, the GLF is more flexible in design than the EWMA.

The detection capability of Residual-based charts including the GLF significantly depends on the form and magnitude of the residual mean. For comparison, therefore, 28 examples of various processes (i.i.d., AR(1), ARMA(1,1)) and mean shifts (step, spike, sinusoidal) are chosen to generate 7 different type of residual means, according to which the examples are divided into 7 groups. Each group consists of 4 examples with different mean shift sizes. Mean shifts are assumed to occur at time t = 1. The step mean shift is defined as $\mu_t = 0$ for $t < 1$ and $\mu_t = \mu$ for $t \geq 1$, where μ_t is a process mean at time t. The spike mean shift is defined as $\mu_t = 0$ for $t < 1$, $\mu_t = \mu$ for $t = 1$, and $\mu_t = 0$ for $t \geq 2$. S_1, S_2, and S_3 in Table 2 indicate the sinusoidal mean shifts with an amplitude of $.75\sigma_a$ and a period of 2, 4, 8 timesteps, respectively. S_4 has an amplitude of $1.5\sigma_a$ and a period of 8. For the 28 examples in Table 2, the GLF and the EWMA are optimally designed to minimize the out-of-control ARL while constraining the in-control ARL to 500. Table 2 shows the ARL values obtained based on a simulation with the 250,000 replications with the simulation standard errors shown in parentheses.

The numerical results for all of the 28 examples in Table 2 show that the OGLF outperforms or performs comparably with the optimal EWMA in every case. The ARL improvement tends to become more substantial as the magnitude of the mean shift increases. For some examples with a large mean shift, the EWMA converges to the Shewhart chart with λ = 0 in Equation (10) since the Shewhart chart is the most effective form of EWMA for detecting large mean shifts. In many cases, however, the ARL performance of the Shewhart chart is also significantly worse than that of the OGLF. This is because the Shewhart chart focuses only on the most recent observation, whereas the GLF is designed to consider the transient dynamics and the steady state value as well.

The performance of the OGLF for sinusoidal mean shifts is examined by amplitude and period. The OGLF detects sinusoidal mean shifts faster with shorter periods and/or larger amplitudes. To sum up, the OGLF outperforms the optimal EWMA in 17 of the 28 examples, and the reduction in the out-of-control ARL over the optimal EWMA reaches 96%.

4. Conclusion

In this paper, the concept of linear filters for control chars is generalized and the GLF is proposed as a control charting scheme. In addition, we have developed a methodology to optimally design a GLF in accordance with the statistical optimization criterion of minimizing the out-of-control ARL while constraining the in-control ARL to some desired value. The ARL performance of the OGLF is compared with those of the residual-based Shewhart chart and the optimal EWMA. The optimal linear filters substantially outperform the existing control charts in situations where their lower order model structures are an obstacle to optimization. Especially with large mean shifts, the improvement is remarkable.

No one chart consistently outperforms the others. However, the significance of GLF is based on their structural flexibility which allows the derivation of a linear filter that outperforms, or performs comparably to, existing control charts such as the residual-based Shewhart chart, EWMA chart, and PID charts. Because of the relationship between the impulse response coefficients and the residual means, the flexibility of the filter structure plays a key role in determining its performance. Therefore, additional flexibility from a higher order filter guarantees better detection capability for more kinds of mean shifts. In other words, in this capacity, the performance the OGLF is superior to other existing charts.

In the optimization procedure, the Monte Carlo simulation is used to make up for the inaccuracy in the ARL that is due to the rough approximation of the Markov property of the GLF. It significantly increases the computational expense. Apley and Chin (2004) provide an alternative approach to

Table 2. Comparison of the Optimal General Linear Filter (OGLF) and the Optimal EWMA

	Time Series Model		*Shift*		*OGLF*	*Optimal EWMA*		
No	θ_1	ϕ_1	*Type*	*Size* (μ/σ_a)	ARL_1	$(1-\lambda)$	h	ARL_1
1	*0*	*0*	Step	.5	28.82 (.03)	.953	.11672	28.82 (.03)
2	*0*	*0*	Step	1.5	5.45 (.01)	.758	.21791	5.45 (.01)
3	*0*	*0*	Step	3	1.86 (.00)	.324	.30670	1.86 (.00)
4	*0*	*0*	Step	4	1.21 (.00)	.113	.32161	1.21 (.00)
5	0	.9	Step	.5	355.31 (.57)	.998	.05271	355.31 (.57)
6	0	.9	Step	1.5	130.64 (.18)	.993	.06540	130.64 (.18)
7	0	.9	Step	3	46.91 (.10)	.979	.08866	49.43 (.07)
8	0	.9	Step	4	13.72 (.06)	.962	.10802	29.78 (.05)
9	0	.9	Spike	.5	495.39 (.98)	0	.32360	497.12 (1.00)
10	0	.9	Spike	1.5	422.01 (.98)	0	.32360	454.46 (.99)
11	0	.9	Spike	3	82.72 (.54)	0	.32360	177.83 (.76)
12	0	.9	Spike	4	6.72 (.14)	0	.32360	28.70 (.32)
13	0	0	Sinusoid	S_1	15.79 (.02)	0	32360	124.20 (.42)
14	*0*	*0*	Sinusoid	S_2	30.69 (.04)	0	.32363	226.61 (.68)
15	*0*	*0*	Sinusoid	S_3	32.90 (.04)	.392	.29861	178.47 (.57)
16	*0*	*0*	Sinusoid	S_4	10.61 (.01)	.384	.29965	26.31 (.05)
17	-.9	.9	Step	.5	447.66 (.75)	.998	.05271	447.66 (.75)
18	-.9	.9	Step	1.5	139.26 (.54)	.997	.05565	255.72 (.39)
19	-.9	.9	Step	2	41.54 (.36)	.996	.05838	194.09 (.28)
20	-.9	.9	Step	3	3.12 (.03)	0	.32360	76.23 (.49)
21	.5	.9	Step	.5	205.04 (.30)	.996	.05839	205.58 (.30)
22	.5	.9	Step	1.5	50.28 (.07)	.979	.08874	50.28 (.07)
23	.5	.9	Step	3	10.77 (.03)	.88	.16616	10.77 (.03)
24	.5	.9	Step	4	2.74 (.01)	.696	.23735	2.88 (.01)
25	.5	.9	Spike	.5	497.47 (.99)	0	.32363	497.61 (.99)
26	.5	.9	Step	1.5	461.86 (.99)	0	.32360	469.74 (.99)
27	.5	.9	Step	3	208.77 (.80)	0	.32360	259.67 (.87)
28	.5	.9	Step	4	50.75 (.41)	0	.32360	86.10 (.56)

Note: ARL_1 is the ARL in the out-of-control process.

reduce this weakness of the OGLF and enable it to provide comparable charting performance in many cases.

REFERENCES

Apley, D. W., and Chin, C. (2004), "Optimal Design of 2nd-order Linear Filters for Statistical Process Control," *To be submitted for Publication.*

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

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.

Johnson, R. A., and Wichern, D. W. (1998), *Applied Multivariate Statistical Analysis* (4th ed.), New Jersey: Prentice Hall.