OPTIMAL FILTER DESIGN APPROACHES TO STATISTICAL PROCESS CONTROL FOR AUTOCORRELATED PROCESSES

A Dissertation

by

CHANG-HO CHIN

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2004

Major Subject: Industrial Engineering

OPTIMAL FILTER DESIGN APPROACHES TO STATISTICAL PROCESS CONTROL FOR AUTOCORRELATED PROCESSES

A Dissertation

by

CHANG-HO CHIN

Submitted to Texas A&M University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Approved as to style and content by:

Daniel W. Apley (Co-Chair of Committee)	Yu Ding (Co-Chair of Committee)
Way Kuo (Member)	Dean W. Wichern (Member)
	Mark L. Spearman (Head of Department)

August 2004

Major Subject: Industrial Engineering

ABSTRACT

Optimal Filter Design Approaches to Statistical Process Control

for Autocorrelated Processes. (August 2004)

Chang-Ho Chin, B.S., Korea University;

M.S., Texas A&M University

Co-Chairs of Advisory Committee: Dr. Daniel W. Apley
Dr. Yu Ding

Statistical Process Control (SPC), and in particular control charting, is widely used to achieve and maintain control of various processes in manufacturing. A control chart is a graphical display that plots quality characteristics versus the sample number or the time line. Interest in effective implementation of control charts for autocorrelated processes has increased in recent years. However, because of the complexities involved, few systematic design approaches have thus far been developed.

Many control charting methods can be viewed as the charting of the output of a linear filter applied to the process data. In this dissertation, we generalize the concept of linear filters for control charts and propose new control charting schemes, the general linear filter (GLF) and the 2^{nd}-order linear filter, based on the generalization. In addition, their optimal design methodologies are developed, where the filter parameters are optimally selected to minimize the out-of-control Average Run Length (ARL) while constraining the in-control ARL to some desired value. The optimal linear filters are compared with other methods in terms of ARL performance, and a number of their

interesting characteristics are discussed for various types of mean shifts (step, spike, sinusoidal) and various ARMA process models (i.i.d., AR(1), ARMA(1,1)).

Also, in this work, a new discretization approach for substantially reducing the computational time and memory use for the Markov chain method of calculating the ARL is proposed. Finally, a gradient-based optimization strategy for searching optimal linear filters is illustrated.

To Almighty God

who testified to his words:

“Cast your burden upon the LORD, and He will sustain you;

He will never allow the righteous to be shaken.”

(Psalms 55:22)

ACKNOWLEDGEMENTS

First and foremost, I give my heartfelt thanks to God for transforming me into a real Christian through hardship. The procedure of pursuing the Ph.D. degree gave me chances not only to acquire new knowledge, but also to realize His good purpose for my life. I admit that living with His purpose is the only way to really live. Throughout my life, I will worship and glorify Him.

I would like to express sincere appreciation to Dr. Daniel Apley for his generosity and masterful guidance. His strict adherence to the scientific method inspired me to strive for a high level of rigor through the course of my research. I will never forget his sharp insights and poignant comments. Special thanks also go to Dr. Yu Ding for providing advice and help in time of need. Often, his involvement went beyond anyone's expectations. I cannot thank them both enough. I would also like to extend thanks to Dr. Kuo and Dr. Wichern who served on my advisory committee.

I am deeply grateful to my parents for their steadfast and devoted love. My father has been my real mentor and closest friend. His heart-warming advice refreshed and encouraged me when I was exhausted. My mother has been a love nest in which I could always find peace. I love and respect them. I should also like to express my gratitude to my grandmother and sisters for their love. I thank God for forming us into one family.

I also owe special thanks to all the Christians in the Vision Mission Church for their prayers and concern.

Lastly, words are not adequate to express my deepest thanks and love to my wife, Hanna, for giving me comfort and joy in the middle of despair and frustration. Hanna, I want to tell you that through the difficulties of the last years, I have been convinced that you are the suitable helper that God prepared for me.

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF FIGURES

FIGURE Page

FIGURE Page

LIST OF TABLES

CHAPTER I

INTRODUCTION

The control chart is a primary Statistical Process Control (SPC) tool that promotes process stability and quality improvement by means of detecting process shifts that require corrective action. As graphical monitors, control charts generally contain a centerline at the target value and two other horizontal lines called control limits at the plus and minus deviation points from the centerline. If a point plotted on the control chart falls outside the control limits, the process is declared not to be in a state of control.

I.1 Generalization of the Concept of Linear Filters

Many common control charts for autocorrelated data are based on linear filtering. 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 $\mathrm{H}(B) = (1 - (1 - \lambda)B)^{-1}\lambda$. More examples are given in Section II.2.

Therefore, many control charting methods can be viewed as the charting of the output of a linear filter applied to the process data. This concept of linear filters for

---

This dissertation follows the style and the format of *Technometrics*.

control charts is generalized in this dissertation, which is one of the main contributions of this research. In addition, based on this generalization, new control charting schemes are proposed and their optimal design methodologies are developed.

I.2 Overview of Control Chart Design

Since the advent of Shewhart charts, many control charts have been developed to monitor, control, and improve processes. Steady efforts have also been made to optimally design them with respect to statistical criteria. As a result, optimal design methodologies have been proposed for simple control charts such as Shewhart charts on independent and identical distributed (i.i.d.) observations. Artiles-León, David, and Meeks (1996) described a methodology to find the optimal control limits of $\bar{x}$ control charts with supplementary stopping rules that minimizes the out-of-control Average Run Length (ARL) for a fixed in-control ARL. Parkhideh and Parkhideh (1998) developed a model to optimally design a flexible zone individual chart based on the desired in-control and out-of-control ARL values.

Tables, plots, and crude heuristics are available for designing more complicated charts such as the Exponentially Weighted Moving Average (EWMA) chart, the Autoregressive Moving Average (ARMA) chart (Jiang, Tsui, and Woodall 2000), and the Proportional Integral Derivative (PID) chart (Jiang, Wu, Tsung, Nair, and Tsui 2002). Crowder (1989) provided the plots of optimal smoothing parameters and control limit constants to aid the design of EWMA charts. A table containing a list of optimal parameters of EWAM control schemes is offered to facilitate its design in Lucas and

Saccucci (1990). Lin and Adams (1996) proposed construction guidelines for the combined exponentially weighted moving average-Shewhart (CES) control chart. VanBrackle and Reynolds (1997) generated tables to aid in adjusting the control limits of EWMA and Cumulative sum (CUSUM) charts in order to provide a reasonable false alarm rate in the presence of correlation. Jiang et al. (2000) and Jiang et al. (2002) developed informal procedures to determine the appropriate parameter values of the ARMA(1,1) chart and the PID chart based on two signal-to-noise ratios. Design procedures for EWMA charts with estimated parameters were developed by Jones (2002).

I.3 Control Charts in the Presence of Correlation

Conventional control charts are based on the assumption that the observations are independently and identically distributed (i.i.d.) over time. With increasing automation, however, inspection rates have increased. Consequently, data are more likely to be autocorrelated, which can significantly deteriorate control charting performance. Johnson and Bagshaw (1974) and Bagshaw and Johnson (1975) discussed the effect of serial correlation on the performance of CUSUM charts, and Harris and Ross (1991) investigated the impact of serial correlation on the performance of EWMA and CUSUM charts.

Numerous control chart modifications have been proposed for monitoring autocorrelated processes. One approach is to monitor the original autocorrelated data using conventional control charts with modified control limits (Johnson and Bagshaw

1974; Vasilopoulos and Stamboulis 1978; Zhang 1998). Another common approach is to apply conventional control charts with normal control limits to the uncorrelated residuals of an appropriate ARMA model (Alwan and Roberts 1988; Runger, Willemain, and Prabhu 1995; Lin and Adams 1996; Apley and Shi 1999). In addition, Jiang et al. (2000) and Jiang et al. (2002) proposed the ARMA(1,1) chart and the PID chart, respectively, for use with autocorrelated data.

In contrast to the aforementioned modifications, few design procedures have been developed for autocorrelated data. Although for Exponentially Weighted Moving Average (EWMA) charts on i.i.d. data, tables do exist (Lucas and Saccucci 1990) that provide the optimal EWMA parameters that minimize the out-of-control ARL under some specified constraint on the in-control ARL, no such tables exist for autocorrelated data. This is because the optimal EWMA filter parameter depends on many factors, including the details of the ARMA process model. The filter design problem is even more complex for the ARMA(1,1) chart of Jiang et al. (2000) and the PID chart of Jiang et al. (2002) because more filter parameters must be selected.

I.4 Performance Measurement

In the design procedure, the ARL, which is defined as the average number of samples plotted before the first alarm sounds, is a popular measure used for evaluating the performance of control charts. The integral equation method and the Markov chain method are widely used to calculate the ARL. Crowder (1987) and VanBrackle and Reynolds (1997) used the integral equation method originally developed by Page (1954)

to evaluate the performance of EWMA charts and CUSUM charts. Yang and Makis (1997) derived integral equations for the ARLs of conventional control charts applied to process residuals. Brook and Evans (1972) originally proposed the Markov chain method to calculate the ARL of CUSUM schemes; many others used this method to evaluate the performance of existing charts such as EWAM charts and CUSUM charts (Crosier 1986; Reynolds, Amin, and Arnold 1990; Runger and Prabhu 1996; VanBrackle and Reynolds 1997; Jiang 2001). Reynolds (1995) proposed a unified treatment of the two methods.

Applying the integral equation method and the Markov chain method to a one-sided CUSUM is illustrated below. For a one-sided CUSUM scheme to detect positive shifts, we plot

$$S_t = \max\{0, S_{t-1} + X_t - K\}, \tag{1.1}$$

where t is a time index; X_t is a sample statistic at timestep t; and K is the reference value. Page (1954) proposed the integral equation to calculate the ARL for this scheme as

$$L(s) = 1 + L(0)F(k-s) + \int L(x)dF(k+x-s), \tag{1.2}$$

where $L(x)$ is the ARL of the CUSUM chart after it is reset at x and F is the cumulative distribution function of the sample statistic. The solution to the integral equation can be obtained by the Gauss-Legendre quadrature (Kantorovich and Krylov 1964; Baker 1977).

The Markov chain method described by Brook and Evans (1972) enables us to calculate the ARL for a continuous CUSUM scheme by discretizing its state space into N_{mc} subintervals. The discrete CUSUM has the transition probability matrix in the partitioned form

$$R = \begin{pmatrix} Q & (I-Q)\underline{1} \\ \underline{0}^T & \underline{1} \end{pmatrix}, \tag{1.3}$$

where the submatrix Q contains the transition probabilities for non-absorbing states; I is the identity matrix; and $\underline{1}$ is a column vector of ones. Let the $\underline{L}_r$ be a vector of length N_{mc} whose elements represent the probabilities of a run length r starting from N_{mc} non-absorbing states, respectively. Then,

$$L_1 = (I-Q)\underline{1}, \tag{1.4}$$

and

$$L_r = QL_{r\text{-}1} = Q^{r\text{-}1}L_1. \tag{1.5}$$

Let ARL(N_{mc}) be the ARL for the discrete scheme with N_{mc} subintervals. Using the ARL(N_{mc}) for several values of N_{mc}, the extrapolation to the asymptotic ARL is obtained by fitting

$$\text{ARL}(N_{mc}) = \text{asymptotic ARL} + \text{B}/N_{mc}^{\,2} + \text{C}/N_{mc}^{\,4}, \tag{1.6}$$

by least squares. This approximation (Equation (1.6)) is usually used to improve the accuracy of the Markov chain method.

I.5 Outline of the Dissertation

The purpose of this research is to generalize the concept of linear filters for control charts and develop optimal design methodologies for linear filters that have a design structure which is flexible enough to include existing control charts as special cases. The linear filters are optimally designed with respect to a statistical criterion such that the out-of-control ARL is minimized while constraining the in-control ARL to some specific value.

In Chapter II, control charting schemes based on linear filtering are described in some detail. Based on a generalization of this concept, we propose a general control charting scheme for autocorrelated data, the general linear filter (GLF). An optimal design methodology for the GLF is developed with respect to the aforementioned statistical optimization criterion. Optimal GLFs are compared with other methods in terms of ARL performance and a number of interesting characteristics are discussed for various types of mean shifts and various ARMA process models.

Chapter III presents a 2^{nd}-order linear filter as a control charting scheme. It is less versatile than the GLF, but is much more efficient to implement and its performance is almost as good as the GLF's in many cases. The performance of 2^{nd}-order linear filters is analyzed for various examples and their characteristics are also discussed.

Chapter IV develops a new discretization method for the Markov chain method which substantially reduces memory use and computational time in implementation. The developed approach is compared with conventional approaches in terms of accuracy and computational expense.

Chapter V illustrates a gradient-based optimization strategy. A flowchart is presented to show the overall strategy. Section V.2 explains the gradient-based search in detail. Section V.3 discusses the selection of the search starting point and its impact on the convergence to the optimal linear filters. Finally, Chapter VI concludes this dissertation by describing the contributions made here as well as listing some directions for future research.

CHAPTER II

OPTIMAL DESIGN OF GENERAL LINEAR FILTERS FOR STATISTICAL PROCESS CONTROL

II.1 Introduction

In spite of the extensive research mentioned in Sections I.2 and I.3, to this date no control chart consistently outperforms the others, because the performance of control charts is substantially influenced by the original process. In the design procedure, the sample size, sampling interval, control limits, and parameters of the control chart can be optimally selected according to the underlying process. However, the basic structure of the control chart has not been a subject of investigation. In other words, the inherent characteristics of the charted statistics generated by the fixed structure of control charts have limited improvement in control chart design and performance.

Therefore, in this chapter, we propose a control charting scheme, which we call the general linear filter (GLF), that is based on generalizing the concept of linear filters. As mentioned above, many control charting schemes for both i.i.d. and autocorrelated data can be viewed as charting the output of a linear filter applied to the process data. The GLF is expressed in the general form of linear filters and is flexibly designed to measure the underlying process without incurring limitations that are due to the inherent characteristics of a fixed structure. Furthermore, we develop a statistical design methodology for the selection of optimal filter parameter values. In Section II.2, we generalize the linear filtering operation of control charting schemes and propose the

general linear filter (GLF). Section II.3 discusses the calculation of the ARL for the GLF based on the Markov chain method and the gradient-based numerical optimization strategy for optimal design. In Section II.4, a performance comparison between optimal general linear filters (OGLF) and other control charts is presented and the interesting characteristics of the OGLF are illustrated. Section II.5 presents the chapter summary.

II.2 General Linear Filter (GLF)

As discussed in Section I.1, the Shewhart individual chart and the EWMA chart on i.i.d are based on linear filtering. Residual-based Shewhart and EWMA charts can be viewed similarly if x_t is assumed to follow an ARMA process model, plus (potentially) an additive deterministic mean shift, μ_t of the form

$$x_t = \frac{\Theta(B)}{\Phi(B)} a_t + \mu_t, \tag{2.1}$$

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

$$e_t = \frac{\Phi(B)}{\Theta(B)} x_t = \frac{\Phi(B)}{\Theta(B)} \left[\frac{\Theta(B)}{\Phi(B)} a_t + \mu_t \right] = a_t + \frac{\Phi(B)}{\Theta(B)} \mu_t = a_t + \tilde{\mu}_t, \tag{2.2}$$

where $\tilde{\mu}_t = \Phi(B)/\Theta(B)\mu_t$ is just the filtered version of the deterministic mean shift μ_t.

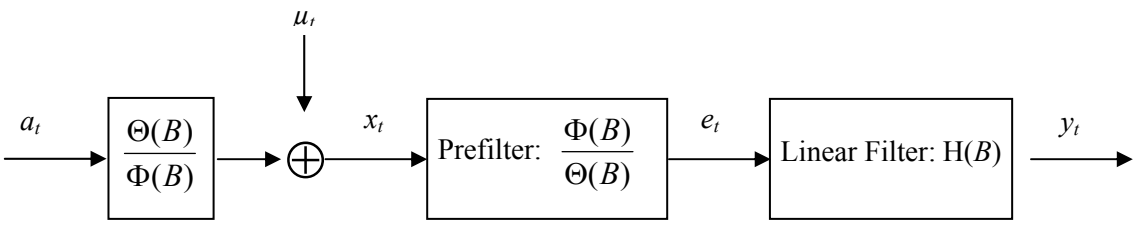


Figure 2.1. Block Diagram Representation of a Linear Filtering Operation.

We may view this $\Phi(B)/\Theta(B)$ in Equation (2.2) as a linear prefilter to the Shewhart or EWMA filter, as shown in Figure 2.1 and Table 2.1. Table 2.1 also includes the 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

$$y_t = \mathrm{H}(B)e_t = h_0 e_t + h_1 e_{t-1} + h_2 e_{t-2} + \ldots\ldots + h_{Tr} e_{t-Tr} = \sum_{j=0}^{Tr} h_j e_{t-j}\,, \qquad (2.3)$$

where H(B) is the general linear filter (GLF) in design and Tr is a truncation time large enough to approximate $h_j \cong 0$ for $j > Tr$. Based on the model in Equation (2.3), we treat the 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 ARL subject to the in-control ARL, equaling some specified value.

Table 2.1. 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	$\dfrac{\lambda}{1-(1-\lambda)B}$
Shewhart on e_t	$y_t = e_t = \dfrac{\Phi(B)}{\Theta(B)} x_t$	Yes	1
EWMA on e_t	$y_t = (1-\lambda)\, y_{t-1} + \lambda\, e_t$	Yes	$\dfrac{\lambda}{1-(1-\lambda)B}$
ARMA(1,1) chart on x_t	$y_t = \dfrac{\theta_0 - \theta B}{1-\phi B} x_t$	No	$\dfrac{\theta_0 - \theta B}{1-\phi B}$
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	$\dfrac{1-B}{\left[\begin{array}{l}1-(1-k_I-k_P-k_D)B \\ \quad -(k_P+2k_D)B^2 + k_D B^3\end{array}\right]}$

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

II.3 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, which is denoted ARL_0 in the in-control process and ARL_1 in the out-of-control process. Since the y_t in Equation (2.3) does not have the Markov property, we approximate y_t as a one-dimensional Markov process:

$$f_{y_t|y_{t-1}}(s_t|s_{t-1}) \cong f_{y_t|y_{t-1},y_{t-2},\ldots\ldots}(s_t|s_{t-1},s_{t-1},\cdots\cdots), \tag{2.4}$$

where s_t is a specific state at timestep t and f is the conditional probability distribution function of y_t given the previous states. The approximation of the Markov property of the charted statistic y_t increases the discrepancy between the approximated ARL and the

actual one. However, the approximated ARL can still be used to compare the performance of two different linear filters in the optimization procedure, because it is proportionate to the actual ARL. In cases requiring the precise value of the ARL, rather than a relative magnitude, the Monte Carlo simulation is used to make up for the inaccuracy in the ARL and always guarantee that the final OGLF really does have the desired in-control ARL.

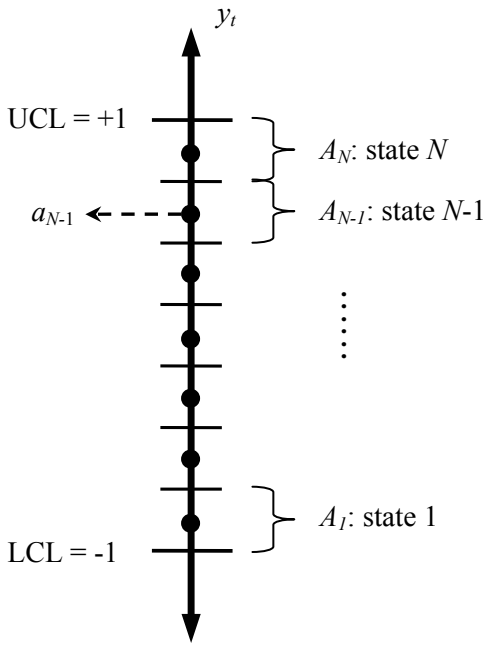


Figure 2.2. One-dimensional State Space Discretized for the Markov Chain Approach.

y_t and y_{t-1} has a joint Gaussian distribution as

$$\begin{bmatrix} y_t \\ y_{t-1} \end{bmatrix} \sim N\left(\begin{bmatrix} \hat{\mu}_t \\ \hat{\mu}_{t-1} \end{bmatrix}, \begin{bmatrix} \sigma_t^2 & v_t \\ v_t & \sigma_{t-1}^2 \end{bmatrix} \right), \tag{2.5}$$

where $\hat{\mu}_t = \sum_{j=0}^{t-1} h_j \tilde{\mu}_{t-j}$ 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 with y_{t-1} fixed is (Johnson and Wichern 1998)

$$N\left(\hat{\mu}_t + \frac{v_t (y_{t-1} - \hat{\mu}_{t-1})}{\sigma_{t-1}^2}, \sigma_t^2 - \frac{v_t^2}{\sigma_{t-1}^2} \right). \tag{2.6}$$

Since all of the impulse response coefficients of the GLF can be evenly scaled, we set the control limits for y_t at ±1 without loss of generality. 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 2.2, A_j indicates the subinterval for state j, and a_j = LCL + $(j - 1/2)\delta$ is the midpoint of A_j. For the Markov chain approach, the i^{th} row, j^{th} column element ($1 \le i,j \le N$) of the transition probability matrix at time t for the nonabsorbing states, denoted Q_t^{ij}, is defined as

$$\begin{aligned} Q_t^{ij} &= Pr\{y_t \in A_j \mid y_{t-1} = a_i\} \\ &= Pr\{a_j - \delta/2 < y_t \le a_j + \delta/2 \mid y_{t-1} = a_i\}. \end{aligned} \tag{2.7}$$

Then, the ARL can be approximated as (Brook and Evans 1972)

$$\text{ARL} = \underline{\pi}_0 (\text{I} + Q_1 + Q_1 Q_2 + Q_1 Q_2 Q_3 + \cdots) \underline{1}, \tag{2.8}$$

where $\underline{1}$ denotes a column vector of ones and $\underline{\pi}_0$ denotes the initial state probability vector. The elements of $\underline{\pi}_0$ are all zero, except for a single element of one that corresponds to the initial value for y_t (typically zero). Because Q_t approaches a steady state value as the mean of residuals settles down to a steady-state value, we have $Q \cong Q_m \cong Q_{m+1} \cong \cdots$ for a sufficiently large m. Thus, Equation (2.8) becomes

$$\text{ARL} = \sum_{p=1}^{m-1} b_p \underline{1} + b_m [I - Q]^{-1} \underline{1}, \tag{2.9}$$

where $b_p = \underline{\pi}_0 \prod_{l=1}^{l-1} Q_l = b_{p-1} Q_{p-1}$ can be calculated recursively for $p = 1, 2, \ldots, m$ with $b_1 = \underline{\pi}_0$. Additional discussion of this truncation for the one-dimensional Markov chain case can be found in Lu and Reynolds (1999). The optimization algorithm uses the following analytical expression for the derivative of the ARL with respect to the filter parameters, which we developed for more effective implementation. Let h_j denote the $(j+1)^{\text{th}}$ filter coefficient. Based on Equation (2.8), it can be shown that

$$\begin{aligned} \frac{\partial ARL}{\partial h_j} &= \underline{\pi}_0 \left[\frac{\partial Q_1}{\partial h_j} + \left(\frac{\partial Q_1}{\partial h_j} Q_2 + Q_1 \frac{\partial Q_2}{\partial h_j} \right) + \left(\frac{\partial Q_1}{\partial h_j} Q_2 Q_3 + Q_1 \frac{\partial Q_2}{\partial h_j} Q_3 + Q_1 Q_2 \frac{\partial Q_3}{\partial h_j} \right) + \cdots \right] \underline{1} \\ &= \sum_{p=1}^{m-1} b_p \frac{\partial Q_p}{\partial h_j} c_p + b_m \left[I - Q \right]^{-1} \frac{\partial Q}{\partial h_j} c_m , \end{aligned} \tag{2.10}$$

where $c_p = [\text{I} + Q_{p+1} + Q_{p+1}Q_{p+2} + \cdots]\underline{1} = \underline{1} + Q_{p+1}c_{p+1}$ can be calculated recursively for $p = m, m-1, \ldots, 1$ with initial condition $c_m = [\text{I} + Q + QQ + \cdots]\underline{1} = [I - Q]^{-1}\underline{1}$.

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 preliminary information required to implement the optimization algorithm includes the ARMA model for the underlying process, the magnitude and type of the mean shift of particular interest, a reasonable initial guess of the optimal general linear filter (OGLF) such as the Shewhart chart or the EWMA chart, and the desired in-control ARL. The optimization search starts from the user-specified initial guess of the OGLF 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.

II.4 Discussion and Examples

II.4.1 Comparison with the PID Chart

To compare the GLF with other existing control charts, we consider the spring-mass-dashpot system in Pandit and Wu (1983). The dynamics of the mechanical system can be described by the ARMA(2,1) process (Jiang et al. 2000)

$$X_t - 1.4385X_{t-1} + .6000X_{t-2} = a_t + .5193a_{t-1}, \qquad (2.11)$$

where $\hat{\sigma}_X$ = 9.130 and $\hat{\sigma}_a$ = 2.212. On the assumption that Equation (2.11) is the perfect model for the process, the model residuals are i.i.d. A zero-state ARL performance comparison of the OGLF with the PID charts, the EWMAST chart (= P

chart) of Zhang (1998), and the residual-based Shewhart chart is shown in Table 2.2, where the zero-state ARL is the ARL of a process starting from zero. The parameters of the PID charts are taken directly from Table 1 of Jiang et al. (2002): these were appropriately determined based on the authors' heuristic algorithm. Some elaborate design method might improve the performance of the PID chart by finding a better set of parameters, but such a method does not exist. Note that our zero-state ARL values in Table 2.2 differ somewhat from the steady-state ARL values shown in Table 1 of Jiang et al. (2002). All of the charts under consideration are designed to provide an in-control ARL of 370.

Table 2.2. ARLs of the OGLF, the Residual-based Shewhart Chart, and the PID charts

Shift ($\Delta=\mu/\sigma_X$)	*OGLF (LCL,UCL)=(-1,+1)* *ARL*	*Residual-based Shewhart chart (L=3.000)*	*P* K_P*=-.8 (L=2.596)*	*PI* (K_P,K_I)*=(-.3, 1.8) (L=2.978)*	*PD* (K_P,K_D)*=(-.8,.5) (L=2.531)*
0	370	370	370	370	370
	(.68)	(.74)	(.73)	(.73)	(.72)
.5	61.26	200	141	351	118
	(.15)	(.56)	(.27)	(.72)	(.22)
1	1.40	3.56	44.9	118	37.3
	(.01)	(.06)	(.08)	(.53)	(.06)
2	1.00	1.00	11.6	1.00	10.9
	(.00)	(.00)	(.02)	(.00)	(.01)
3	1.00	1.00	5.44	1.00	5.60
	(.00)	(.00)	(.01)	(.00)	(.00)

Note: the simulation standard errors are shown in parentheses.

The step mean shift is assumed to occur at $t = 1$. 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. Figure 2.3 shows the fault signature for $\Delta = .5$ and the impulse response coefficients (h_j) of the OGLF for each mean shift, where the fault signature is defined as the time-varying mean of the residuals (Apley and Shi 1999). h_j indicates how the past and present residuals, e_{t-j}, affect

the present statistic y_t as shown in Equation (2.3). In this example, the OGLF outperforms or performs comparably with the other charts. For shifts Δ = .5 and 1, the OGLFs are oscillating filters around $\tilde{\mu}_t$ = .025 and zero, respectively, which improves the ARL performance of the best PID chart significantly. The impulse response of the OGLF, which is highly correlated to the fault signature, promotes a larger magnitude of the charted statistic y_t in Equation (2.3) than those of the Shewhart chart or the PID chart for the first several timesteps. This results in the higher detection capability of the OGLF and, thereby, causes the huge reduction in the out-of-control ARL.

Figure 2.3(a) and (b) show the fault signature and impulse response of the OGLF for Δ = .5. As the timestep moves forward, the fault signature and the impulse response coefficients show a positive correlation and a negative correlation by turns. Hence, the charted statistic y_t in Equation (2.3) comes out to be a large value each time even if the sign changes in turn. See the plots (e) and (f) of Figure 2.3, where the OGLF is scaled for illustration purpose so that the largest impulse response coefficient is equal to the largest mean of the residuals. A similar explanation is given for the OGLF for Δ = 1. The OGLF simply reduces to the Shewhart chart for Δ = 2 and 3 as shown in Figure 2.3(d).

Note the first coefficient of the OGLF increases as the mean shift size increases. As the OGLF for Δ = 1 also takes advantage of the high correlation with the fault signature, it tries to detect a mean shift at an earlier stage of occurrence with the larger first coefficient than that for Δ = .5. The OGLFs for Δ = 2 and 3 have an even larger first coefficient, since the first spike of the fault signature is large enough to be detected at the first timestep.

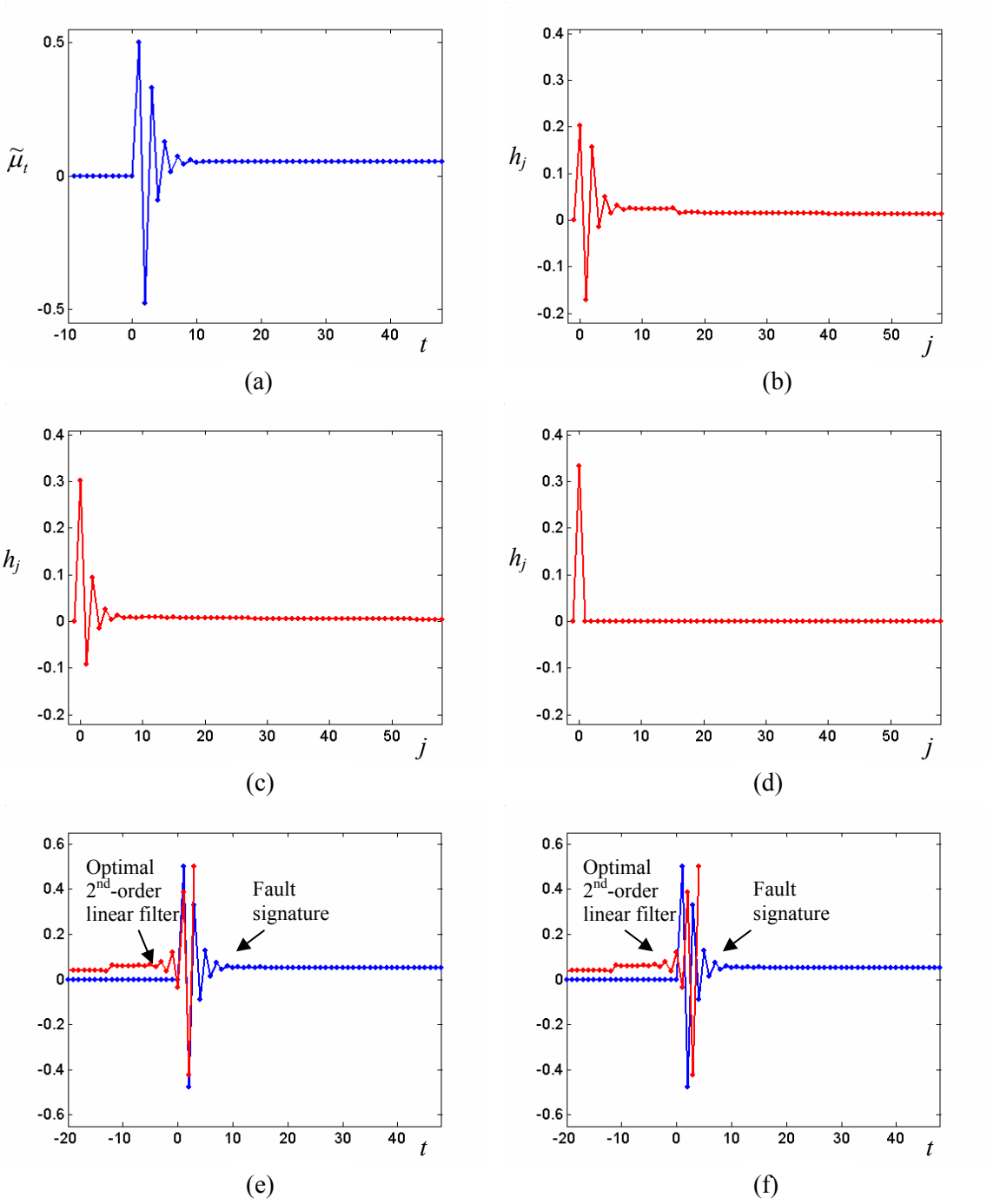


Figure 2.3. Spring-mass-dashpot System: (a) Fault Signature for Δ = .5; (b) Impulse Response of the OGLF for Δ = .5; (c) Impulse Response of the OGLF for Δ = 1; (d) Impulse Response of the OGLF for Δ = 2 and 3; (e) OGLF Applied to the Fault Signature for Δ = .5 at t = 3; (f) OGLF Applied to the Fault Signature for Δ = .5 at t = 4.

II.4.2 Performance Improvement over the Optimal EWMA

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

$$y_t = (1 - \lambda)y_{t-1} + ke_t, \tag{2.12}$$

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 (2.2). 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 two parameters – λ and k − providing one design degree of freedom. In this sense, the GLF is more flexible in design than the EWMA. For the 28 examples in Table 2.3, the GLF and the EWMA are optimally designed to minimize the out-of-control ARL while constraining the in-control ARL to 500. Table 2.3 shows the ARL values obtained based on a simulation with the 250,000 replications with the simulation standard errors shown in parentheses.

For comparison, 28 examples of various processes (i.i.d., AR(1), ARMA(1,1)) and mean shifts (step, spike, sinusoidal) are considered. Mean shifts are assumed to occur at time $t = 1$. The step mean shift is defined in Section II.4.1 and 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 2.3 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 and a period of 8.

Table 2.3. Comparison of the Optimal General Linear Filter (OGLF) and the Optimal EWMA

	Time Series Model		*Shift*		*OGLF*	*Reduction of ARL_1 by OGLF*	*Optimal EWMA*		
No	θ_1	ϕ_1	*Type*	*Size* (μ/σ_a)	ARL_1	(%)	(1-λ)	*h*	ARL_1
1	*0*	*0*	*Step*	.5	28.82 (.03)	0	.953	.11672	28.82 (.03)
2				1.5	5.45 (.01)	0	.758	.21791	5.45 (.01)
3				3	1.86 (.00)	0	.324	.30670	1.86 (.00)
4				4	1.21 (.00)	0	.113	.32161	1.21 (.00)
5	0	.9	Step	.5	355.31 (.57)	0	.998	.05271	355.31 (.57)
6				1.5	130.64 (.18)	0	.993	.06540	130.64 (.18)
7				3	46.91 (.10)	5	.979	.08866	49.43 (.07)
8				4	13.72 (.06)	54	.962	.10802	29.78 (.05)
9	0	.9	Spike	.5	495.39 (.98)	8	0	.32360	497.12 (1.00)
10				1.5	422.01 (.98)	7	0	.32360	454.46 (.99)
11				3	82.72 (.54)	53	0	.32360	177.83 (.76)
12				4	6.72 (.14)	77	0	.32360	28.70 (.32)
13	0	0	Sinusoid	S_1	15.79 (.02)	87	0	32360	124.20 (.42)
14				S_2	30.69 (.04)	86	0	.32363	226.61 (.68)
15				S_3	32.90 (.04)	82	.392	.29861	178.47 (.57)
16				S_4	10.61 (.01)	59	.384	.29965	26.31 (.05)
17	-.9	.9	Step	.5	447.66 (.75)	0	.998	.05271	447.66 (.75)
18				1.5	139.26 (.54)	46	.997	.05565	255.72 (.39)
19				2	41.54 (.36)	79	.996	.05838	194.09 (.28)
20				3	3.12 (.03)	96	0	.32360	76.23 (.49)
21	.5	.9	Step	.5	205.04 (.30)	0	.996	.05839	205.58 (.30)
22				1.5	50.28 (.07)	0	.979	.08874	50.28 (.07)
23				3	10.77 (.03)	0	.88	.16616	10.77 (.03)
24				4	2.74 (.01)	5	.696	.23735	2.88 (.01)
25	.5	.9	Spike	.5	497.47 (.99)	0	0	.32363	497.61 (.99)
26				1.5	461.86 (.99)	2	0	.32360	469.74 (.99)
27				3	208.77 (.80)	20	0	.32360	259.67 (.87)
28				4	50.75 (.41)	41	0	.32360	86.10 (.56)

In Table 2.3, the 7 combinations of the processes and the mean shifts generate 7 different fault signatures, according to which the examples are divided into 7 groups. Each group consists of 4 examples with different mean shift sizes.

The numerical results for all of the 28 examples in Table 2.3 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 $\lambda = 0$ in Equation (2.12) since the Shewhart chart is the most effective form of EWMA for detecting large mean shifts. However, the ARL performance of the Shewhart chart is also significantly improved by the OGLF. This is because the design of the Shewhart chart is determined by the initial magnitude of the fault signature only, 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%. These huge reductions are discussed in detail along with other interesting characteristics of the OGLF in the following section. The OGLFs for all of the examples are graphically shown in Appendix A.

II.4.3 Optimal Filter Characteristics

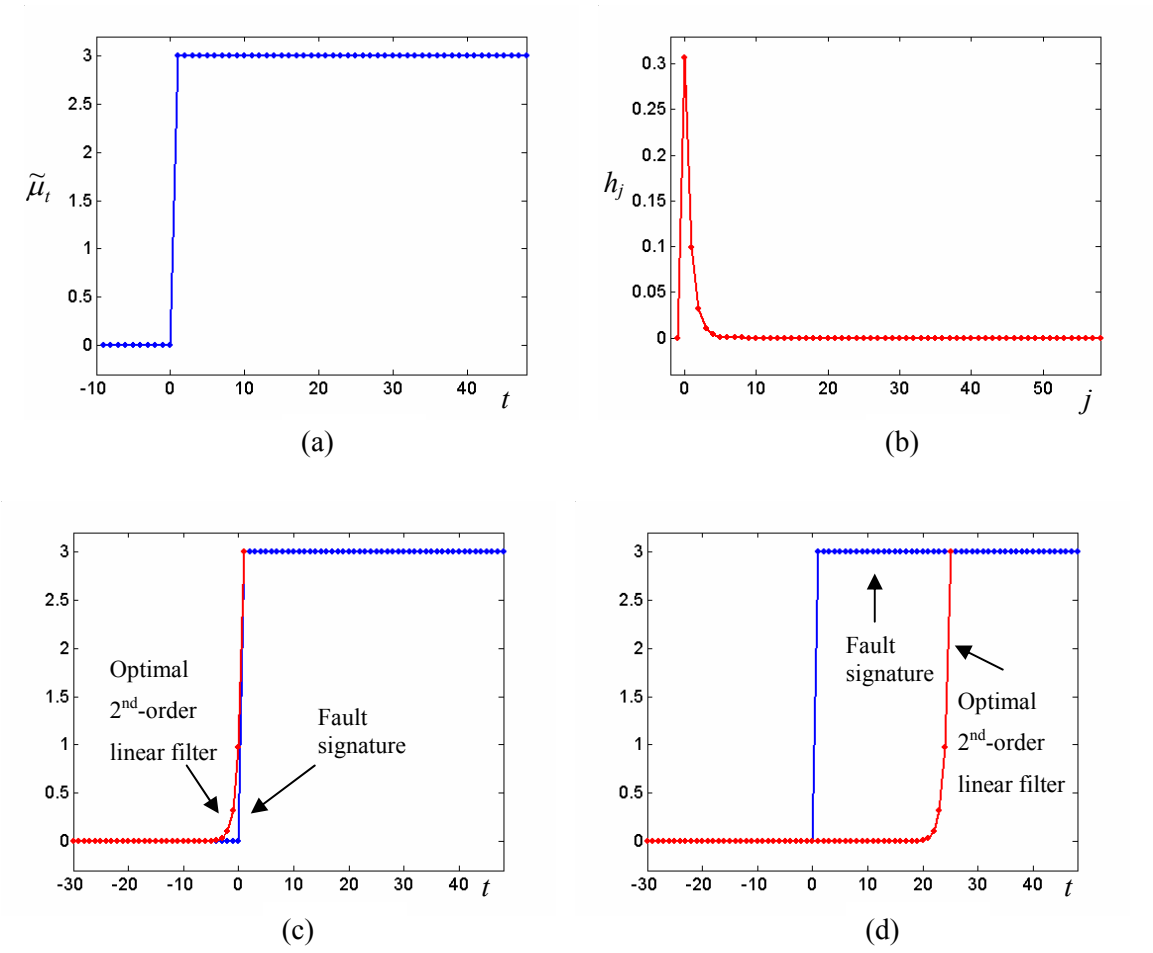


Figure 2.4. Example 3: (a) Fault Signature; (b) Impulse Response of the OGLF; (c) OGLF Applied to the Fault Signature at t = 1; (d) OGLF Applied to the Fault Signature at t = 25.

For the i.i.d. processes with a step mean shift (Examples 1 to 4 in Table 2.3), each OGLF looks very similar to an EWMA. Thus, we tries to estimate the parameters λ and k of the EWMA − a first-order filter representing the OGLF with the ARX function

of MATLAB 6.5. The estimated values of λ are consistent with the optimal values in Table 4 of Lucas and Saccucci (1990). For mean shift $\mu_t = 3$, the estimated λ is .676 with scaling constant k = .30670 for control limits ±1. Note that we define the EWMA as Equation (2.12). In the notation of Lucas and Saccucci (1990), the EWMA is shown to be equivalent to an EWMA with λ = .676 and L = 3.085, which is the optimal combination for a mean shift of 3 standard deviations in Table 4 of Lucas and Saccucci (1990). Figure 2.4(a) and (b) show the fault signature and impulse response of the OGLF for mean shift $\mu_t = 3$. The design parameter λ of the EWMA is determined according to the magnitude of the mean shift. The larger the magnitude of the mean shift is, the larger the λ of the EWMA that is selected with the fixed control limits. Figures 2.4(c) and (d) show how the OGLF is applied to the fault signature, where the OGLF is scaled for illustration purpose so that the largest impulse response coefficient is equal to the largest mean of the residuals.

For the AR(1) processes with a step mean shift (Examples 5 to 8 in Table 2.3), the mean of the residuals settles down to a small steady state value after an initial single spike. The OGLF converges to the optimal EWMA with a small λ for mean shifts $\mu_t = .5\sigma_a$ and $1.5\sigma_a$, since an EWMA is adequate for detecting the small initial spike and small steady state shift. However, the large mean shifts, such as $\mu_t = 3\sigma_a$ and $4\sigma_a$, cause the fault signature to have a large initial spike and a small steady state shift as shown in Figure 2.5(a). Hence, the OGLF is optimally designed to be sensitive for detecting both large and small shifts and, therefore, outperforms the optimal EWMA.

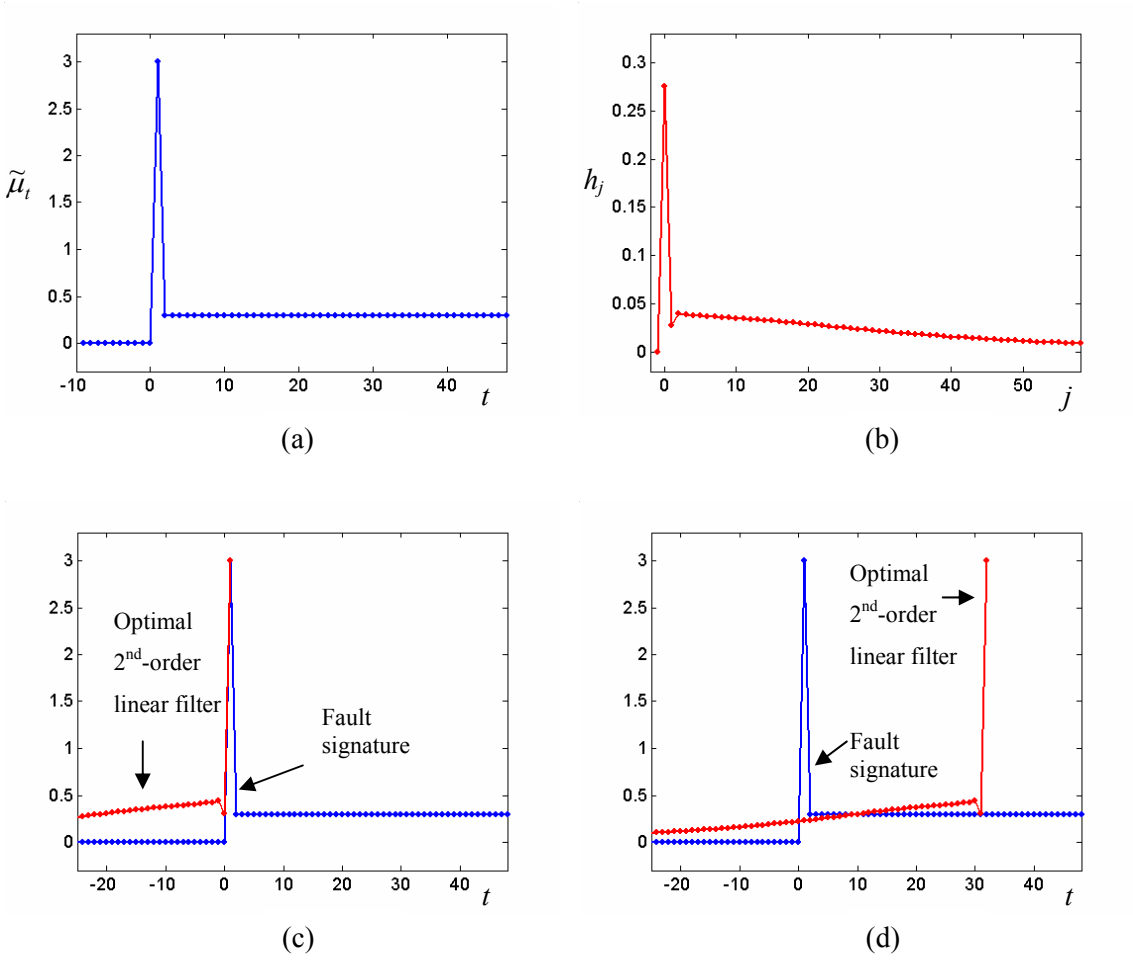


Figure 2.5. Example 7: (a) Fault Signature; (b) Impulse Response of the OGLF; (c) OGLF Applied to the Fault Signature at t = 1; (d) OGLF Applied to the Fault Signature at t = 32.

Figure 2.5(b) shows the OGLF for $\mu_t = 3\sigma_a$, which can be viewed as the weighted combination of a Shewhart chart and an EWMA. In other words, the h_0 and h_j for $j > 1$ in Figure 2.5(b) can be approximated as the impulse responses of a Shewhart chart and an EWMA with a small λ, respectively. The ARX function is used to estimate the parameters of an EWMA taking h_j for $j > 1$ in Figure 2.5(b) as its impulse response, so

that the estimated λ = .0227 and k = .04214. Hence, the OGLF for μ_t = 3 can be approximated as

$$y_t = .23306e_t + \frac{.04214}{1-.9773B}e_t . \qquad (2.13)$$

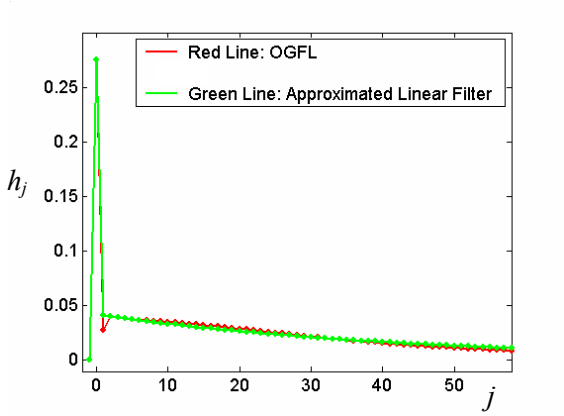


Figure 2.6. Example 7: Impulse Responses of the OGLF and Its Approximated Linear Filter in Equation (2.13).

Figure 2.6 shows the impulse responses of the approximated filter and the OGLF on the same plot. Except for the value of h_1, they are almost identical. This OGLF performs similarly to the combined Shewhart−EWMA scheme proposed by Lucas and Saccucci (1990). The Shewhart chart filter component of the OGLF is effective in detecting the large initial single spike at start-up and its EWMA filter component increases the probability of detection by providing an additional chance to detect the small steady state value of the mean shift following the spike with its long tail. The

properties of the Shewhart chart and the EWMA are optimally combined into one statistic of the OGLF, whereas the combined Shewhart−EWMA scheme considers two statistics with respective control limits at the same time.

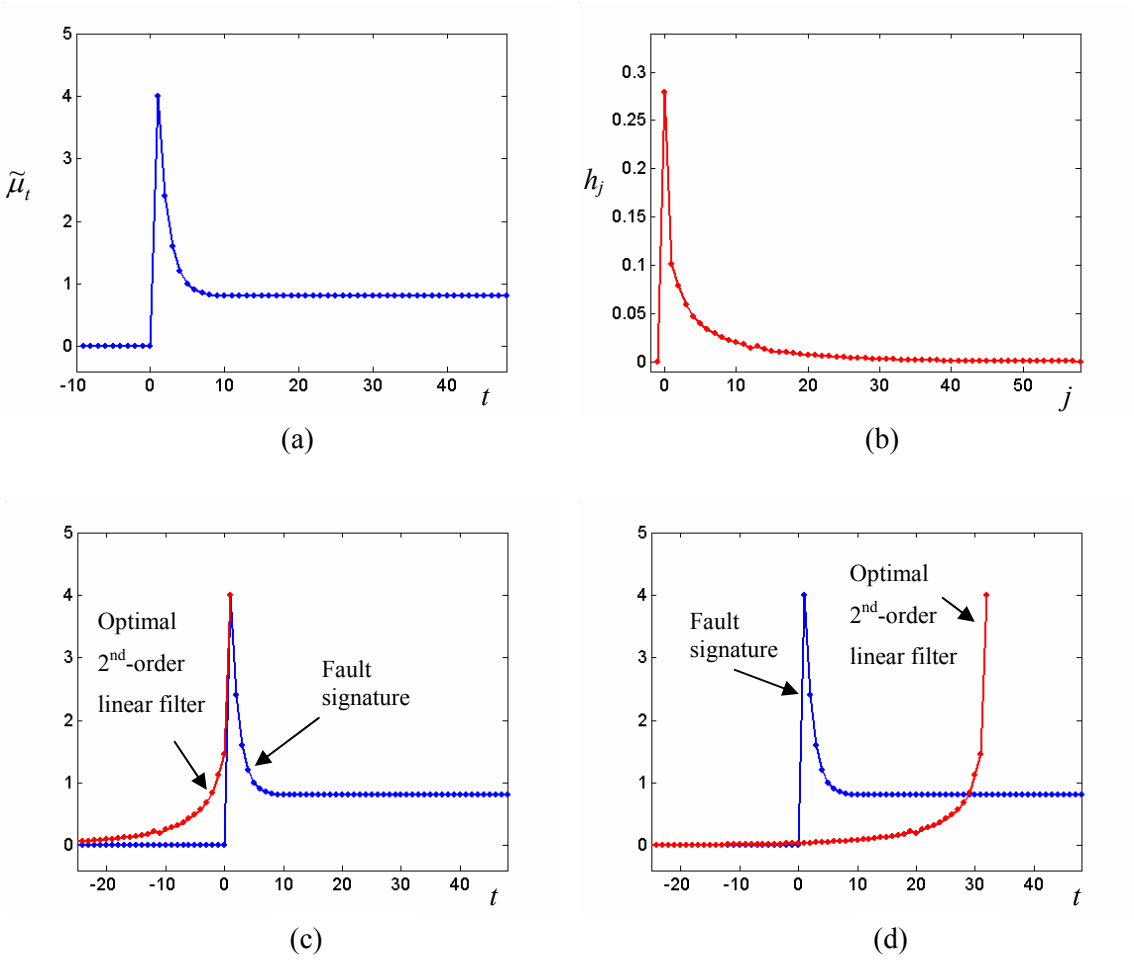


Figure 2.7. Example 24: (a) Fault Signature; (b) Impulse Response of the OGLF; (c) OGLF Applied to the Fault Signature at t = 1; (d) OGLF Applied to the Fault Signature at t = 32.

A similar discussion explains Examples 24 shown in Figure 2.7. The fault signature of the ARMA(1,1) processes with a step mean shift also has an initial single

spike and afterward reaches a moderate steady state value. For other examples with similar fault signatures, see Examples 21 to 23. However, in these cases, the OGLF is designed to be an EWMA because of the relatively small first spike and moderate steady state value of the fault signature.

The OGLF for the AR(1) processes with a spike mean shift (Examples 9 to 12) shows a high correlation with the fault signature, where the initial positive spike is followed by a single negative spike and then settles down to zero. From Equation (2.3) which expresses the charted statistic as a linear combination of the filter coefficients and the residuals, we can easily see that the high correlation between the impulse response coefficient and the residuals contributes to an increase in the magnitude of the charted statistic. Therefore, the OGLF is more effective in detecting this kind of fault signature than the optimal EWMA. As shown in Figure 2.8(a), the fault signature stays non-zero only for the first two timesteps. The optimal EWMA for this kind of fault signature is the Shewhart chart, which considers only the most recent observation. By generating a high correlation with the fault signature, on the other hand, the OGLF shown in Figure 2.8(b) can effectively consider the two non-zero means of the residuals at the same time and, therefore, has higher detection capability. (see Figure 2.8(c) and (d)) As the magnitude of the mean shift increases, the performance improvement over the optimal EWMA becomes more substantial and the reduction in the out-of-control ARL by the OGLF reaches as high as 77%.

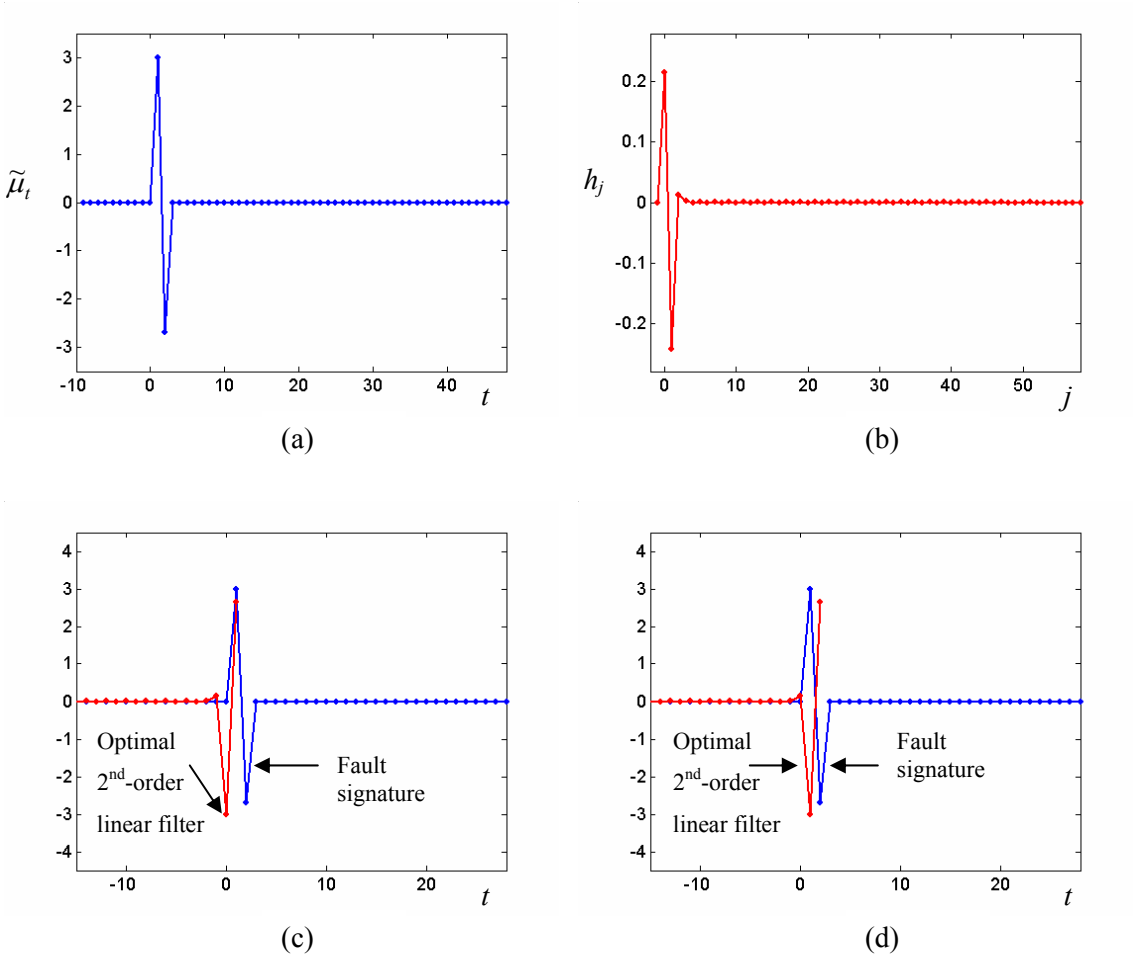


Figure 2.8. Example 11: (a) Fault Signature; (b) Impulse Response of the OGLF; (c) OGLF Applied to the Fault Signature at t = 1; (d) OGLF Applied to the Fault Signature at t = 2.

For the i.i.d. processes with a sinusoidal mean shift, the OGLF significantly outperforms the optimal EWMA. As shown in Figure 2.9, the OGLF for these processes is also designed to utilize the high correlation with the fault signature to increase the detection probability. The shorter the period of the sinusoidal mean shift is, the faster the detection by the OGLF. As shown in Figure 2.9, this is because the magnitude of the

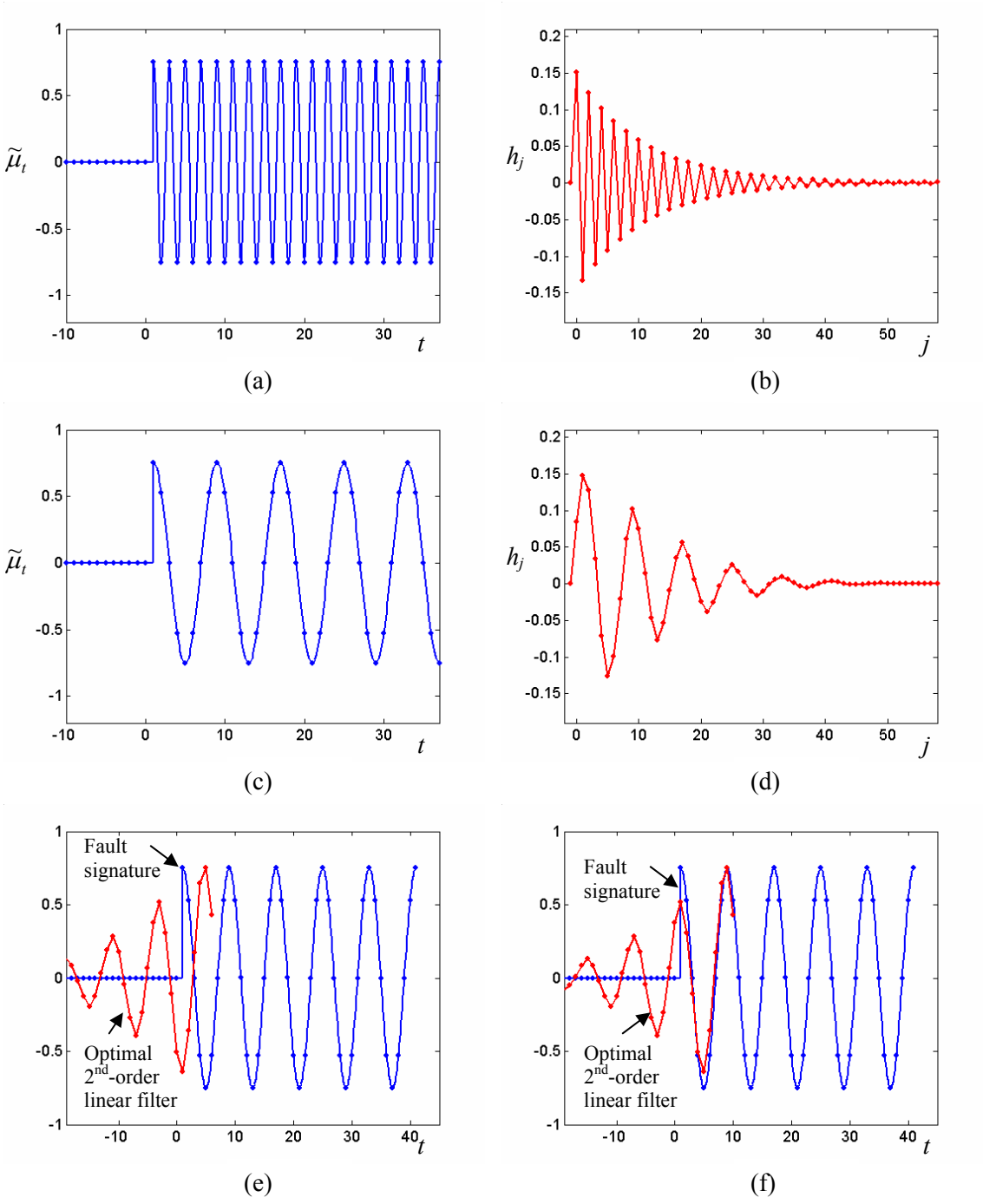


Figure 2.9. Example 13: (a) Fault Signature; (b) Impulse Response of the OGLF; Example 15: (c) Fault Signature; (d) Impulse Response of the OGLF; Example 15: (e) OGLF Applied to the Fault Signature at t = 6; (f) OGLF Applied to the Fault Signature at t = 10.

charted statistic for the OGLF in Example 13 is maximized every 1 timestep after a reasonable amount of time, whereas in Example 15, it is maximized every 4 timesteps. Similarly to the example in Section II.4.1, the OGLF for Example 13 shows a positive correlation and a negative correlation with the fault signature in turn as the timestep moves forward, and the charted statistic comes out to be a large value at each timestep. Therefore, the mean shift in Example 13 is detected more quickly.

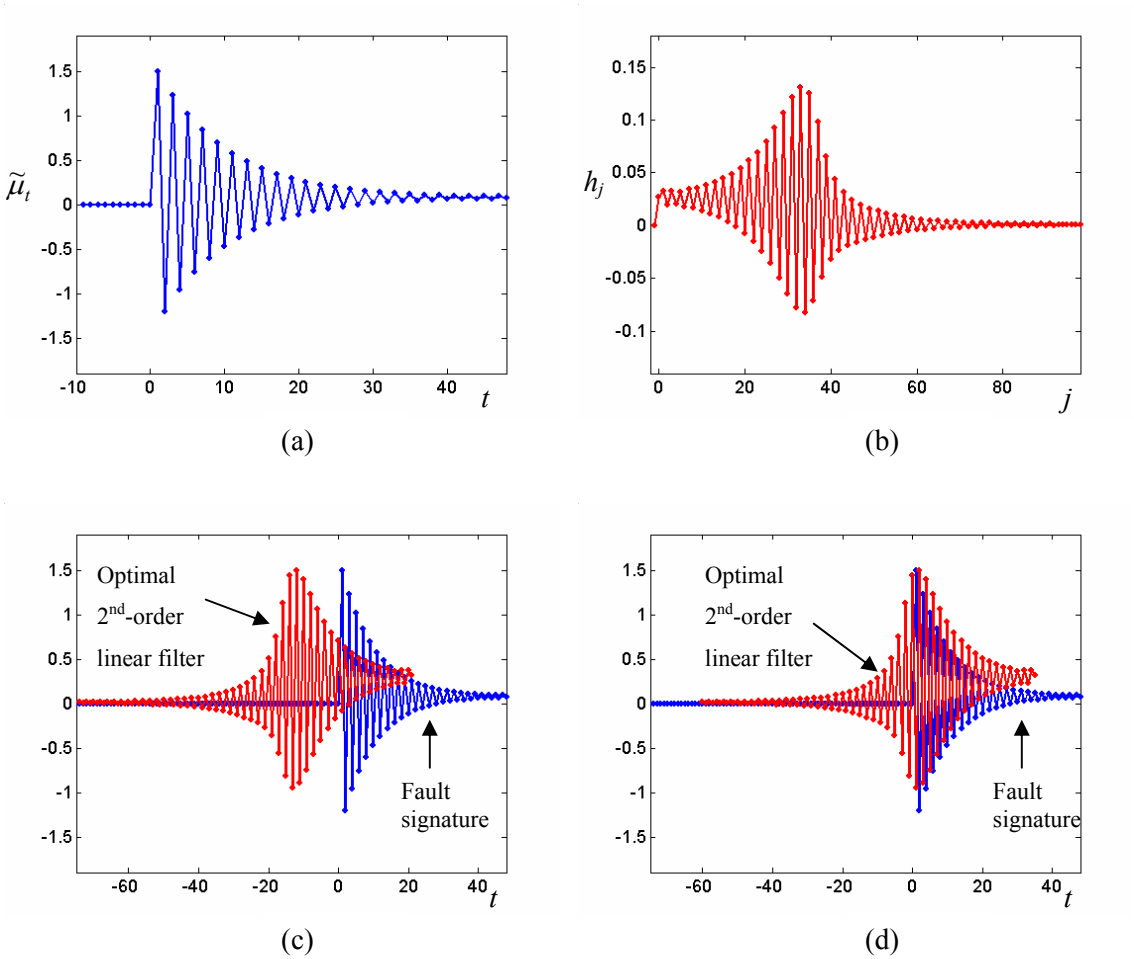


Figure 2.10. Example 18: (a) Fault Signature; (b) Impulse Response of the OGLF; (c) OGLF Applied to the Fault Signature at t = 21; (d) OGLF Applied to the Fault Signature at t = 35.

Figures 2.10 and 2.11 depict the ARMA(1,1) processes with a step mean shift (Examples 17 to 20). The fault signatures shown in Figures 2.10(a) and 2.11(a) are analogous to those of the spring-mass-dashpot system discussed in Section II.4.1. However, the OGLF is designed differently according to the magnitude of the mean shift. For Example 17 with $\mu_t = .5\sigma_a$, the OGLF simply reduces to the optimal EWMA.

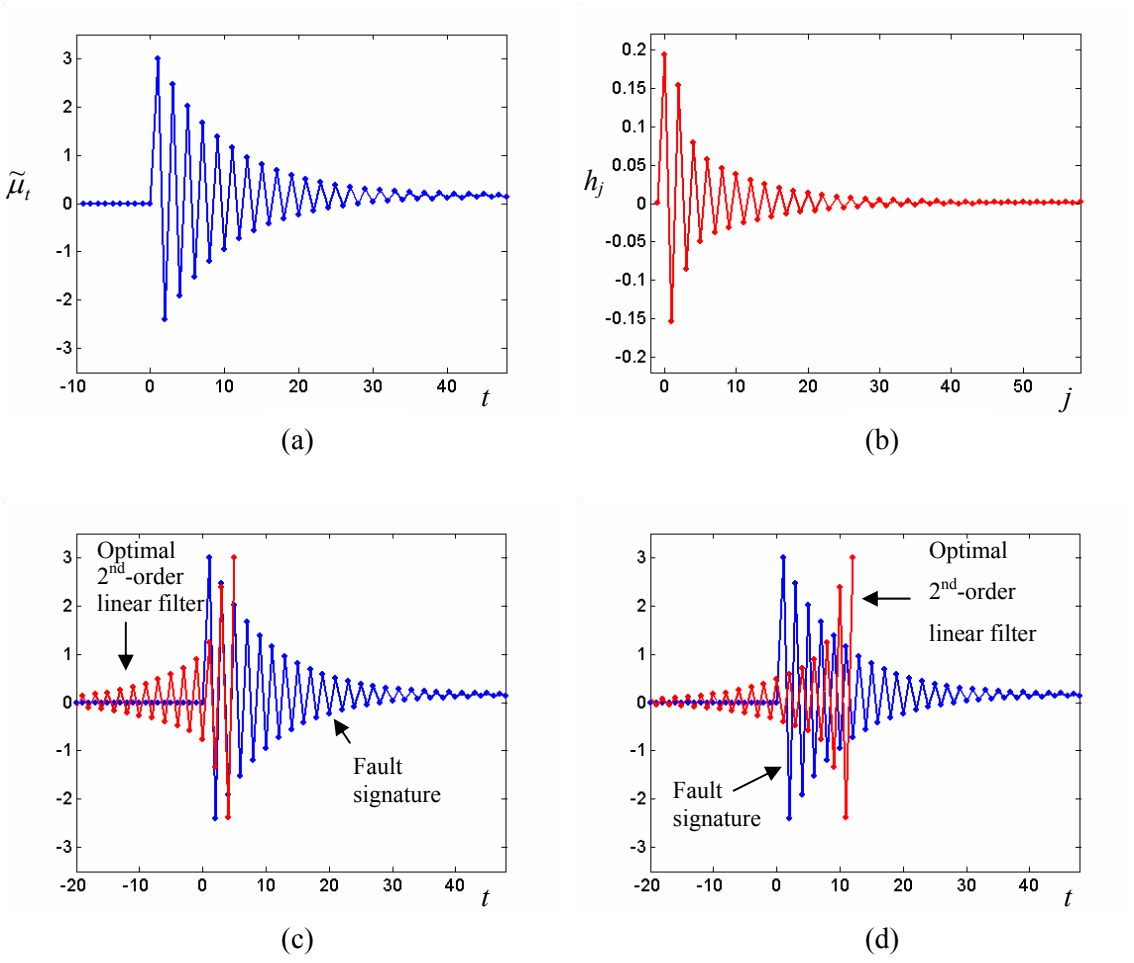


Figure 2.11. Example 20: (a) Fault Signature; (b) Impulse Response of the OGLF; (c) OGLF Applied to the Fault Signature at t = 5; (d) OGLF Applied to the Fault Signature at t = 12.

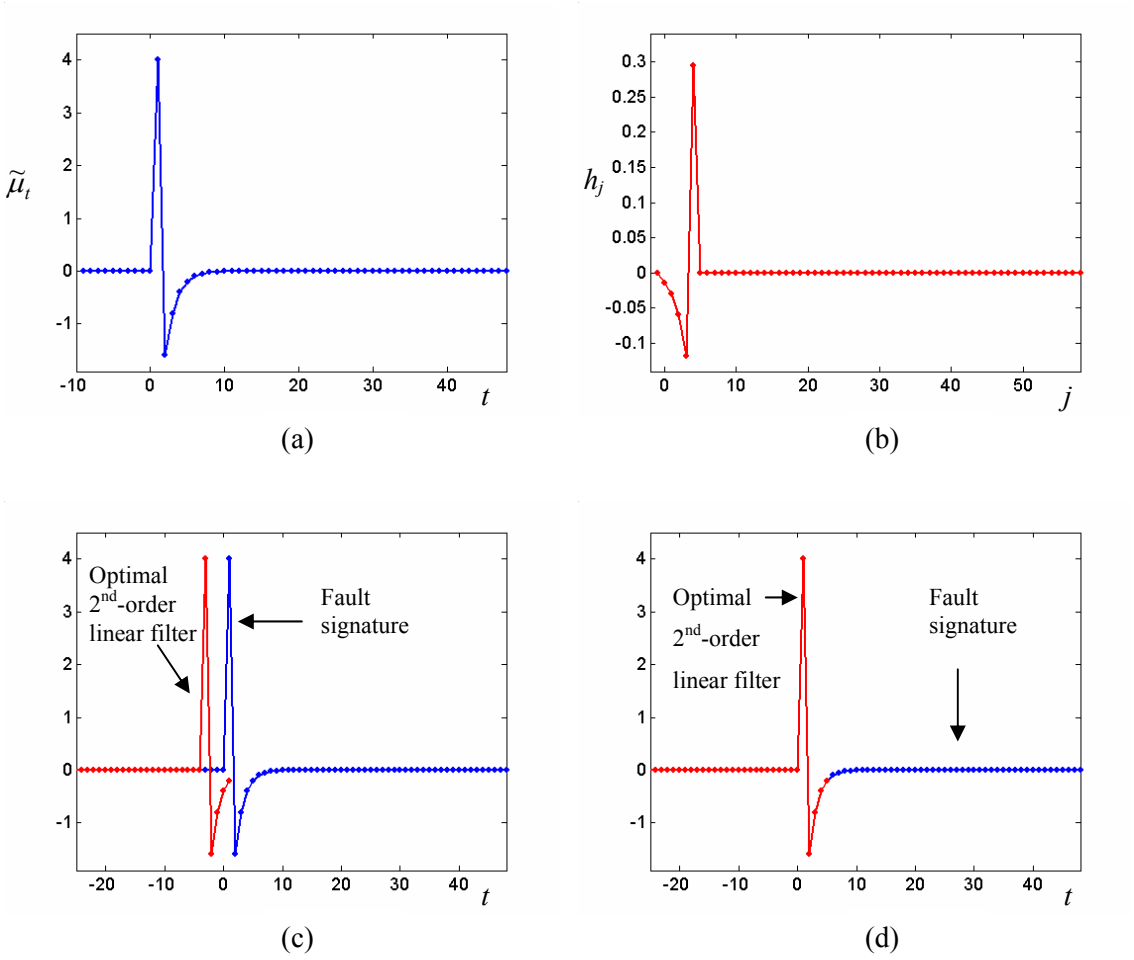


Figure 2.12. Example 28: (a) Fault Signature; (b) Impulse Response of the OGLF; (c) OGLF Applied to the Fault Signature at t = 1; (d) OGLF Applied to the Fault Signature at t = 5.

For Example 19 with $\mu_t = 2\sigma_a$ and Example 20 with $\mu_t = 3\sigma_a$, the OGLF is designed to be highly correlated with the fault signature as in Example 13. (see Figure 2.11(b)) However, note that the magnitude of the charted statistic and the detection probability increases for the first several timesteps and then decreases. This is because, unlike in Example 13, in this case, the fault signature dies out. As shown in Figure

2.10(b), for Example 18 with $\mu_t = 1.5\sigma_a$, the first half of the OGLF is similar to the fault signature and the other half is similar to the OGLF for Example 20. The impulse response of the OGLF oscillates around $\tilde{\mu}_t$ = .026 up to timestep 34, as the amplitude geometrically increases. The OGLF is designed to sacrifice the initial several timesteps by means of placing very small coefficients on the most recent residuals and, instead, to increase the detection probability afterward by generating a high correlation between the coefficients and the fault signature. This mechanism is also found in the OGLF for Examples 25 to 28 of the ARMA(1,1) processes with a spike mean shift. As shown in Figure 2.12, the charted statistic of the OGLF is maximized at timestep 5.

II.5 Chapter Summary

We have established above that many control charting schemes can be described in terms of linear filtering. Based on the generalization of this concept, we have proposed 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 has been compared with other methods and it repeatedly shows remarkable superiority over the others. It always performs at least comparably with the other existing charts, such as the Shewhart chart, the EWMA chart, and the PID chart. Since the OGLF includes the other existing charts as special cases, it sometimes becomes one of them. Therefore, the proposed methodology guarantees the correct choice of control chart as well as the optimization of the chosen chart.

The higher detection capability of the OGLF results from its flexible structure which, in the design procedure, takes into account the transient dynamics and the steady state value of the fault signature. In Examples 7, 8, and 24 with an initial single spike and a non-zero steady state value of the fault signature, the GLF is optimally designed to possess the properties of the Shewhart chart and the EWMA chart. For examples with pronounced transient dynamics in the fault signature, the OGLF shows a high correlation with the fault signature. In some examples, such as Examples 18 and 25 to 28, the OGLF places small coefficients on the most recent residuals to increase the detection probability for the following timesteps. These interesting relationships between the GLF and the fault signature are useful for selecting reasonable starting points for the gradient-based numerical optimization strategy discussed in Section II.3.

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. Chapter III provides an alternative approach to reduce this weakness of the OGLF and enable it to provide comparable charting performance in many cases.

CHAPTER III

OPTIMAL DESIGN OF 2ND-ORDER LINEAR FILTERS FOR STATISTICAL PROCESS CONTROL

III.1 Introduction

Beyond tables, plots, and simple guidelines to assist control chart design (Lin and Adams 1996; VanBrackle and Reynolds 1997; Jiang et al. 2000, Jiang et al. 2002), Chapter II generalizes the concept of linear filters for control charting schemes and develops a design procedure for the GLF that includes existing charts such as the Shewhart chart, the EWMA chart, and the ARMA chart as special cases. It optimally designs GLFs for specific mean shifts in the underlying processes by a gradient-based optimization strategy, where the approximation of the Markov property of the charted statistic results in inaccuracy in the ARL and its derivatives. The Monte Carlo simulation is used to make up for this inaccuracy, but it significantly increases the computational expense required to implement the design procedure. In some cases, moreover, the inaccurate derivative of the ARL debases the optimality of the GLF parameters that were selected by the design procedure. This chapter proposes another control charting scheme, a 2nd-order linear filter, to remove the computational weakness of the OGLF and facilitate the implementation of the linear filter design procedure. The parameters of the proposed linear filter are optimally selected with respect to the statistical criterion about the ARL mentioned in Chapter II.

In Section III.2, our 2nd-order linear filter is introduced with a generalization of the linear filtering concept. Section III.3 discusses how to calculate the ARL for the chart. Calculation of the transition probability is illustrated in Section III.4. Section III.5 develops a computationally efficient method for optimizing the filter parameters. Section III.6 compares the performance of the optimal 2nd-order linear filter with other methods and illustrates some interesting characteristics of the optimal filter for various types of mean shifts (step, spike, sinusoidal) and various ARMA process models. Finally, Section III.7 presents concluding remarks.

III.2 2nd-order Linear Filter

Chapter II generalizes the concept that many common control charts for both i.i.d and autocorrelated data are based on linear filtering, which is illustrated with existing control charts such as the Shewhart chart and the EWMA chart. In order to facilitate the practical implementation of the linear filter design procedure, this chapter, like Jiang et al. (2000), generalizes the simple Shewhart and EWMA linear filtering concepts by utilizing higher-order linear filters for SPC purposes. The focus of this chapter, however, is on optimizing the design of the linear SPC filters. In this chapter, we restrict H(B) to an 2nd-order linear filter and assume that a whitening prefilter is used. Therefore, the control chart statistic can be written as

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + k e_t - k\beta e_{t-1}$$

$$= k\left[\frac{1-\beta B}{1-\alpha_1 B-\alpha_2 B^2}\right] e_t$$

$$= k \frac{\mathrm{M}(B)}{\mathrm{A}(B)} e_t$$

$$= k \frac{\mathrm{M}(B)}{\mathrm{A}(B)} \frac{\Phi(B)}{\Theta(B)} x_t, \qquad (3.1)$$

where k, α_1, α_2 and β are the parameters of our 2nd-order filter, and $\mathrm{A}(B) = [1 - \alpha_1 B - \alpha_2 B^2]$ and $\mathrm{M}(B) = [1 - \beta B]$ are referred to as the AR and MA polynomials for the 2nd-order filter. Equation (3.1) can be expressed similarly to Equation (2.3).

$$y_t = k\left[\frac{1-\beta B}{1-\alpha_1 B - \alpha_2 B^2}\right] e_t = h_0 e_t + h_1 e_{t-1} + h_2 e_{t-2} + \ldots\ldots = \sum_{j=0}^{\infty} h_j e_{t-j}, \qquad (3.2)$$

The 2nd-order filter parameters are selected to directly minimize the out-of-control ARL subject to the in-control ARL equaling, some specified value. This control chart strategy will sound an alarm if the linear filter output y_t falls outside of the specified control limits.

III.3 ARL Calculation

The objective of this chapter is to find the optimal set of filter parameters, which requires that we express the ARL as a function of the filter parameters. To do this, we represent the dynamics of the process as a two-dimensional Markov chain as follows because the y_t in Equation (3.1) does not have the Markov property. Define the vector $V_t = (y_t, z_t)^{\mathrm{T}}$, where $z_t = \alpha_2 y_{t-1} - k\beta e_t$. Thus, the vector V_t can be written as

$$V_t = \begin{bmatrix} y_t \\ z_t \end{bmatrix} = \begin{bmatrix} \alpha_1 & 1 \\ \alpha_2 & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} k \\ -k\beta \end{bmatrix} e_t = DV_{t-1} + We_t, \tag{3.3}$$

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

Note that we can still utilize a two-dimensional state space even if the three ARMA parameters are considered, which is of a lesser order than the suggestion of Jiang (2001). This Markov chain representation releases the optimization routine from the delays due to the use of the Monte Carlo simulation discussed in Chapter II. These two facts result in a significant reduction in computational expense.

Without loss of generality (because of the scalar factor k), we set the control limits for y_t at ±1. The two-dimensional state-space is discretized into a set of rectangles, as shown in Figure 3.1. Although the z-axis technically extends out to ±∞, we may truncate this by defining the upper and lower limits (LL_z, UL_z) such that z_t lies between the limits with very high probability. Let N_z denote the number of discretized subintervals along the z-axis, and let N_y denote the number of discretized subintervals along the y-axis between ±1. The two-dimensional in-control region therefore consists of $N = N_z \times N_y$ nonabsorbing states. Thus, each rectangle between LL_z and UL_z is $\delta_z = (UL_z - LL_z)/N_z$ wide and $\delta_y = 2/N_y$ high. The out-of-control regions (y_t outside the ±1 interval) are treated as a single absorbing state.

The following procedure for calculating the ARL is a two-dimensional version of the Markov chain approach discussed in Brook and Evans (1972) and Lucas and Saccucci (1990) (Runger and Prabhu 1996; Jiang et al. 2000; Jiang 2001). The i^{th} row,

j^{th} column element ($1 \le i,j \le N$) of the transition probability matrix at time t for the nonabsorbing states, denoted Q_t^{ij}, is defined as:

$$Q_t^{ij} = Pr\{V_t \in R_j \mid V_{t\text{-}1} = r_i\}$$

$$= Pr\{r_{j,y} - \delta_y/2 < y_t \le r_{j,y} + \delta_y/2,\ r_{j,z} - \delta_z/2 < z_t \le r_{j,z} + \delta_z/2 \mid y_{t\text{-}1} = r_{i,y}, z_{t\text{-}1} = r_{i,z}\} \quad (3.4)$$

where R_j is the rectangle for state j; r_i is the centroid of R_i; $r_{i,y}$ and $r_{i,z}$ are the YZ coordinates of the centroid of state i. A computational example of Q_t^{ij} is given in Section III.4.

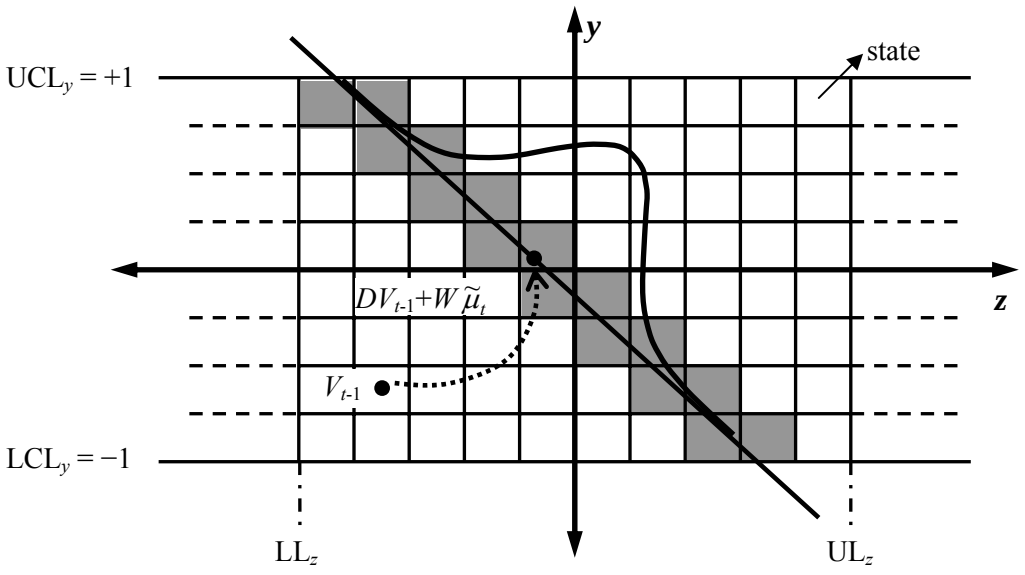


Figure 3.1. Two-dimensional State Space Discretized for the Markov Chain Approach.

Following the analytical expression in Section II.3, in this chapter, the ARL is approximated as

$$\text{ARL} = \sum_{p=1}^{m-1} b_p \underline{1} + b_m [I - Q]^{-1} \underline{1}, \tag{3.5}$$

where m is a sufficiently large integer such that Q_t approaches a steady state value $Q \cong Q_m \cong Q_{m+1} \cong \cdots$, and $b_p = \underline{\pi}_0 \prod_{l=1}^{p-1} Q_l = b_{p-1} Q_{p-1}$ can be calculated recursively for p = 1, 2, . . ., m, respectively with $b_1 = \underline{\pi}_0$.

Equation (3.3) implies that given V_{t-1}, V_t is distributed along a single one-dimensional distribution line in the two-dimensional state space, as illustrated in Figure 3.1. In particular, V_t follows the limiting case of a bivariate normal distribution with mean $DV_{t-1} + W\tilde{\mu}_t$ and rank-1 covariance matrix $WW^{\mathrm{T}}\sigma_a^2$, where $\tilde{\mu}_t$ denotes the mean of e_t. Each Q_t^{ij} can be calculated as the area under the normal density curve (see Figure 3.2) for the segment of the distribution line that falls within rectangle R_j (a more detailed explanation is given in Section III.4). If the distribution line does not pass through a particular rectangle, then the corresponding element of Q_t^{ij} is exactly zero. Although Q_t is an $N \times N$ matrix, each row will contain approximately $2 \times \max\{N_y, N_z\}$ nonzero elements. Thus, Q_t is a sparse matrix, which helps to decrease the computational expense in calculating the ARL.

III.4 Calculation of Q_t^{ij}

An example of calculating an element of the transition probability matrix and its derivative is presented below. The underlying process is assumed to follow the model in Equation (2.1) and then, $e_t \sim NID(\tilde{\mu}_t, \sigma_a^2)$ according to Equation (2.2). As mentioned in Section III.3, the conditional probability $V_t|V_{t-1} \sim N_2(DV_{t-1} + W\tilde{\mu}_t, WW^{\mathrm{T}}\sigma_a^2)$ derived from Equation (3.3) forms a normal distribution line in the two-dimensional state space, where $z_t|V_{t-1} \sim N(\alpha_1 y_{t-1} + z_{t-1} + k\tilde{\mu}_t, k^2\sigma_a^2)$.

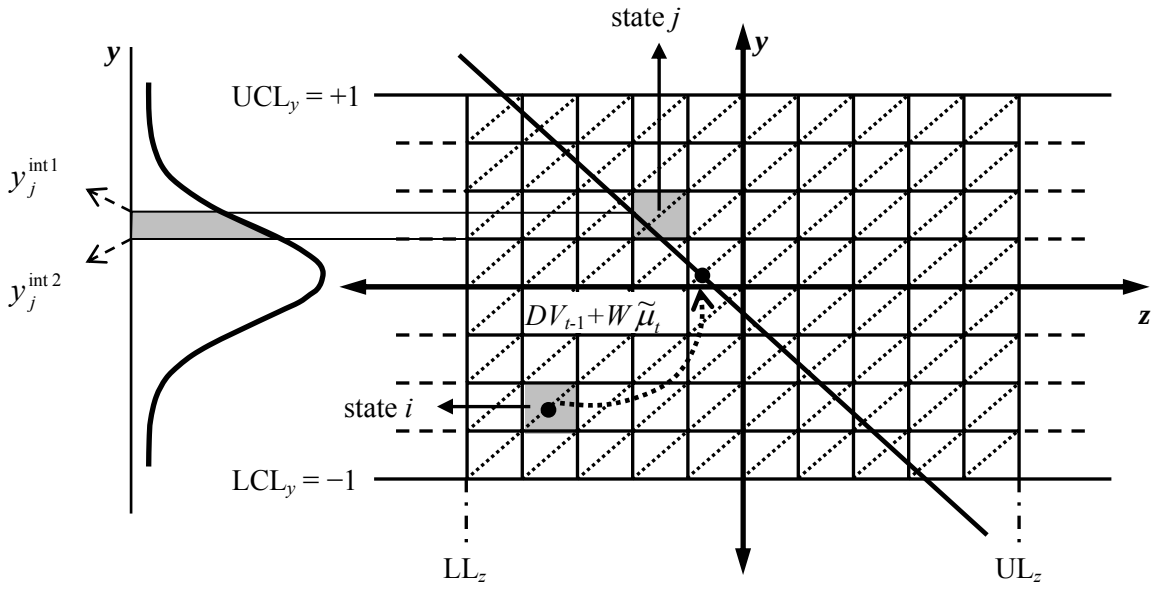


Figure 3.2. Calculation of Q_t^{ij}.

Let $y_j^{\text{int}1}$ and $y_j^{\text{int}2}$ denote the y coordinates of the distribution line intersecting rectangle R_j for state j; the transition probability Q_t^{ij} in Equation (3.4) is calculated as the

area of the normal density curve for the segment of the distribution line falling within the R_j (see Figure 3.2):

$$Q_t^{ij} = F\left(\frac{y_j^{\text{int}1} - (\alpha_1 y_i^c + z_i^c + k\tilde{\mu}_t)}{k\sigma_a}\right) - F\left(\frac{y_j^{\text{int}2} - (\alpha_1 y_i^c + z_i^c + k\tilde{\mu}_t)}{k\sigma_a}\right)$$

$$= F(\tilde{y}_j^{\text{int}1}) - F(\tilde{y}_j^{\text{int}2}), \tag{3.6}$$

where $k > 0$; y_i^c and z_i^c are the coordinates of the centroid of R_i; $\tilde{y}$ is the standardized version of y; and F is the cumulative distribution function of the standard normal distribution.

While the discretization depicted in Figure 3.1 is appropriate to approximate the ARL, it is not an effective way to calculate the derivative of Q_t^{ij} in a case where the distribution line transects only one column of rectangles. For better numerical accuracy, each rectangle in the two-dimensional state space is discretized further into two triangles as shown in Figure 3.2. A triangle represents a state and the transition probabilities are calculated in the same way as in Equation (3.6). The derivative of Q_t^{ij} is obtained as

$$\frac{\partial Q_t^{ij}}{\gamma_q} = f(\tilde{y}_j^{\text{int}1})\frac{\partial \tilde{y}_j^{\text{int}1}}{\partial \gamma_q} - \phi(\tilde{y}_j^{\text{int}2})\frac{\partial \tilde{y}_j^{\text{int}2}}{\partial \gamma_q}, \tag{3.7}$$

where γ_q is the q^{th} element of the filter parameter vector $\gamma = [\alpha_1\ \alpha_2\ \beta\ k]^{\mathrm{T}}$ and f is the probability distribution function of the standard normal distribution. Note that

calculation of the ∂Q_t^{ij} is on the same order of computational complexity as calculating the Q_t^{ij}.

III.5 Optimal Filter Design Strategy

We use a gradient-based numerical optimization strategy to determine the optimal SPC filter parameter vector $\gamma = [\alpha_1 \ \alpha_2 \ \beta \ k]^T$. In this case, only four filter parameters in the optimization procedure of our 2nd-order linear filter need to be tuned, whereas the GLF in Chapter II has the entire impulse response coefficients to design. This method, therefore, contributes to reducing the computational expense and memory use. The user specifies the ARMA process model, the type and magnitude of the mean shift of particular interest, and a desired in-control ARL. The optimization algorithm then finds the filter parameters that minimize the out-of-control ARL for the specified mean shift. The optimization routine is substantially improved if we incorporate gradient information. Although this might seem computationally prohibitive, we propose a method of calculating the gradient that is on the same order of computational complexity as calculating the ARL. This method takes advantage of the sparsity of the transition probability matrix Q_t also.

As in Section II.3, the derivative of the ARL with respect to the q^{th} element of γ denoted by γ_q is approximated as

$$\frac{\partial ARL}{\partial \gamma_q} = \sum_{p=1}^{m-1} b_p \frac{\partial Q_p}{\partial \gamma_q} c_p + b_m [I - Q]^{-1} \frac{\partial Q}{\partial \gamma_q} c_m \,, \tag{3.8}$$

where m and b_p are defined in Section III.3, and $c_p = [\mathrm{I} + Q_{p+1} + Q_{p+1}Q_{p+2} + \cdots]\underline{1} = \underline{1} + Q_{p+1}c_{p+1}$, with initial condition $c_m = [\mathrm{I} + Q + QQ + \cdots]\underline{1} = [I - Q]^{-1}\underline{1}$, can be calculated recursively for $p = 1, 2, \ldots, m$, respectively.

III.6 Discussion and Examples

III.6.1 Comparison with the PID Chart

A example taken from Pandit and Wu (1983) is considered appropriate for comparing our 2nd-order linear filter with existing charts such as the EWMAST chart (= P chart) of Zhang (1998) and the PID chart of Jiang et al. (2002). This experimental example comes from a spring-mass-dashpot system. Jiang et al. (2000) suggested the following ARMA(2,1) process model to fit the first 100 observations:

$$X_t - 1.4385X_{t-1} + .6000X_{t-2} = a_t + .5193a_{t-1}, \tag{3.9}$$

where $\hat{\sigma}_X = 9.130$ and $\hat{\sigma}_a = 2.212$. We assume that Equation (3.9) is the perfect model for the process.

Table 3.1 compares the zero-state ARLs for our optimal 2nd-order linear filter (final ARLs were evaluated based on a Monte Carlo simulation with 250,000 replicates) with those for the residual-based Shewhart chart, P (or EWMAST), PI, and PD charts. For comparison, we use the PID parameters taken from Table 3.1 of Jiang et al. (2002). The OGLF in Table 3.1 is taken from Table 2.2.

Table 3.1. ARLs of the Optimal 2nd-order Linear Filter, the Residual-based Shewhart Chart, and the PID charts

Shift ($\Delta=\mu/\sigma_X$)	*2nd-order Linear Filter (LCL,UCL)=(-1,+1)*					*OGLF (LCL,UCL) =(-1,+1)*	*Residual-based Shewhart chart (L=3.000)*	*P* K_P*=-.8 (L=2.596)*	*PI* (K_P,K_I)*=(-.3, 1.8) (L=2.978)*	*PD* (K_P,K_D)*=(-.8,.5) (L=2.531)*
	α_1	α_2	β	k	*ARL*					
0					370	370	370	370	370	370
					(.68)	(.68)	(.74)	(.73)	(.73)	(.72)
.5	.986	0	0	.08428	76.88	61.26	200	141	351	118
					(.10)	(.15)	(.56)	(.27)	(.72)	(.22)
1	-.529	0	0	.28545	1.59	1.40	3.56	44.9	118	37.3
					(.03)	(.01)	(.06)	(.08)	(.53)	(.06)
2	0	0	0	.33337	1.00	1.00	1.00	11.6	1.00	10.9
					(.00)	(.00)	(.00)	(.02)	(.00)	(.01)
3	0	0	0	.33337	1.00	1.00	1.00	5.44	1.00	5.60
					(.00)	(.00)	(.00)	(.01)	(.00)	(.00)

Note: the simulation standard errors are shown in parentheses.

In Table 3.1, the optimal 2nd-order linear filter reduces to a residual-based EWMA chart for $\Delta = .5$ and a residual-based Shewhart chart for $\Delta = 2$ and $\Delta = 3$. It can be shown from Equation (3.1) that the EWMA filter is a special case of our 2nd-order linear filter. If the parameters of our 2nd-order linear filter are chosen as $A(B) = [1 - (1 - \lambda)B]$, $M(B) = 1$, and $k = \lambda$, the optimal filter is identical to the EWMA filter. In a case where $A(B) = 1$, $M(B) = 1$, and $k = 1$, the 2nd-order linear filter becomes the Shewhart chart. The EWMAST for ARMA(1,1) processes is also a special case of our 2nd-order linear filter when $A(B) = (1 - (1 - \lambda)B)\Phi(B)$, $M(B) = \Theta(B)$, and $k = \lambda$, in Equation (3.1). Therefore, our 2nd-order linear filter may turn out to be an EWMA chart, a Shewhart chart, or an EWMAST chart, as shown in Tables 3.1 and 3.2. We can view the residual-based EWMA as a first-order linear filter with a whitening prefilter and the resudual-based Shewhart chart purely as a whitening prefilter, respectively.

The optimal 2nd-order linear filter performs best in detecting the mean shifts of $\Delta = .5$ and $\Delta = 1$. It produces 35% and 96% reductions in the out-of-control ARL

compared to the best of the PID charts, respectively. A detailed discussion about the huge reduction for a shift $\Delta = 1$ is presented in Section III.6.3. For this example, the optimal 2^{nd}-order linear filter generally outperforms the other charts such as the Shewhart chart, EWMAST chart and PID charts. The following section includes more examples to demonstrate the substantial advantages of the optimal 2^{nd}-order linear filter.

III.6.2 Performance Improvement over the Optimal EWMA

This section illustrates the advantages of the increased complexity that emerges when changing from an EWMA to an 2^{nd}-order linear filter. In this section, the residual-based EWMA chart with control limits ±1 is defined as in Equation (2.12). It is well known that the Shewhart has good detection capability for large mean shifts, that the EWMA chart works better for small mean shifts than the Shewhart chart, and that the Shewhart chart is a special case of the EWMA chart (when $\lambda = 1$ and $k = 1$ in Equation (2.12)). Therefore, the selection of the EWMA chart is a reasonable one for comparing to the performance of our 2^{nd}-order linear filter over a wide range of mean shifts. We consider the examples in Table 3.2 to explain how optimal filters are designed to detect the various types of mean shifts occurring at time $t = 1$ and to compare the ARL performance of the optimal EWMA filter and the optimal 2^{nd}-order linear filter.

Table *3.2. Comparison of the Optimal 2nd-order Linear Filter, the Optimal EWMA, and the OGLF*

	Time Series Model		*Shift*		*OGLF*	*Optimal EWMA Filter*			*Optimal 2nd-order linear Filter*				
No	ϕ_1	θ_1	*Type*	*Size* (μ/σ_a)	ARL_1	(1-λ)	h	ARL_1	α_1	α_2	β	k	ARL_1
1	*0*	*0*	*Step*	.5	28.82 (.03)	.953	.11672	28.82 (.03)	.953	0	0	.11672	28.82 (.03)
2				1.5	5.45 (.01)	.758	.21791	5.45 (.01)	.758	0	0	.21791	5.45 (.01)
3				3	1.86 (.00)	.324	.30670	1.86 (.00)	.324	0	0	.30670	1.86 (.00)
4				4	1.21 (.00)	.113	.32161	1.21 (.00)	.113	0	0	.32161	1.21 (.00)
5	.9	0	Step	.5	355.31 (.57)	.998	.05271	355.31 (.57)	.998	0	0	.05271	355.31 (.57)
6				1.5	130.64 (.18)	.993	.06540	130.64 (.18)	.993	0	0	.06540	130.64 (.18)
7				3	46.91 (.10)	.979	.08866	49.43 (.07)	.86306	.10471	.78365	.27537	47.26 (.1)
8				4	13.72 (.06)	.962	.10802	29.78 (.05)	.86332	.10469	.84730	.29830	13.72 (.06)
9	.9	0	Spike	.5	459.39 (.98)	0	.32360	497.12 (1.00)	-.07002	.04620	.86856	.23679	496.83 (1.00)
10				1.5	422.01 (.98)	0	.32360	454.46 (.99)	-.07075	.03686	.87174	.23643	427.08 (.98)
11				3	82.72 (.54)	0	.32360	177.83 (.76)	-.10326	.00122	.84447	.23596	85.12 (.55)
12				4	6.72 (.14)	0	.32360	28.70 (.32)	-.06867	.03518	.87200	.23669	7.12 (.15)
13	0	0	Sinusoid	S_1	15.79 (.02)	0	32360	124.20 (.42)	-.55750	.32225	.32627	.15058	15.79 (.02)
14				S_2	30.69 (.04)	0	.32363	226.61 (.68)	-.02589	-.90304	-.24291	.14944	30.69 (.04)
15				S_3	32.90 (.04)	.392	.29861	178.47 (.57)	1.1596	-.71570	-1.2077	.08491	43.30 (.08)
16				S_4	10.61 (.01)	.384	.29965	26.31 (.05)	1.0236	-.63649	-1.0699	.10683	11.46 (.01)
17	.9	-.9	Step	.5	447.66 (.75)	.998	.05271	447.66 (.75)	.998	0	0	.05271	447.66 (.75)
18				1.5	139.26 (.54)	.997	.05565	255.72 (.39)	-.92383	.00671	-.03887	.13987	163.10 (.71)
19				2	41.54 (.36)	.996	.05838	194.09 (.28)	-.92383	.00671	-.03887	.13987	43.31 (.37)
20				3	3.12 (.03)	0	.32360	76.23 (.49)	-.86100	-.04540	-.08410	.20510	3.21 (.04)
21	.9	.5	Step	.5	205.04 (.30)	.996	.05839	205.58 (.30)	.996	0	0	.05839	205.58 (.30)
22				1.5	50.28 (.07)	.979	.08874	50.28 (.07)	.979	0	0	.08874	50.28 (.07)
23				3	10.77 (.03)	.88	.16616	10.80 (.03)	.87906	.00020	-.01981	.16390	10.77 (.03)
24				4	2.74 (.01)	.696	.23735	2.88 (.01)	.696	0	0	.23735	2.88 (.01)
25	.9	.5	Spike	.5	497.47 (.99)	0	.32363	497.61 (.99)	-.23811	-.00106	-.03769	.31718	497.47 (1.00)
26				1.5	461.86 (.99)	0	.32360	469.74 (.99)	-.22153	-.00549	-.16337	.32306	469.23 (.99)
27				3	208.77 (.80)	0	.32360	259.67 (.87)	-.21993	-.00610	-.18477	.32339	259.77 (.88)
28				4	50.75 (.41)	0	.32360	86.10 (.56)	-.23006	-.00353	-.15604	.32273	83.72 (.55)

In each example, the parameters of the EWMA filter and our 2nd-order linear filter are optimized to minimize out-of-control ARL with a constraining in-control ARL of 500. The ARL values for control limits ±1 are computed with the Markov chain method introduced in Section III.3. The resulting ARL values are consistent with those based on a Monte-Carlo simulation with 250,000 runs. The simulation standard errors are shown in parentheses. Table 3.2 shows the numerical results of comparisons between optimal EWMA filters and optimal 2nd-order linear filters in terms of ARL performance under ARMA(1,1) processes. The ARLs for the OGLFs considered in Chapter II are also provided in Table 3.2. Mean shifts are assumed to occur at time $t = 1$. The step, spike, and sinusoidal mean shifts are defined as in Chapter II.

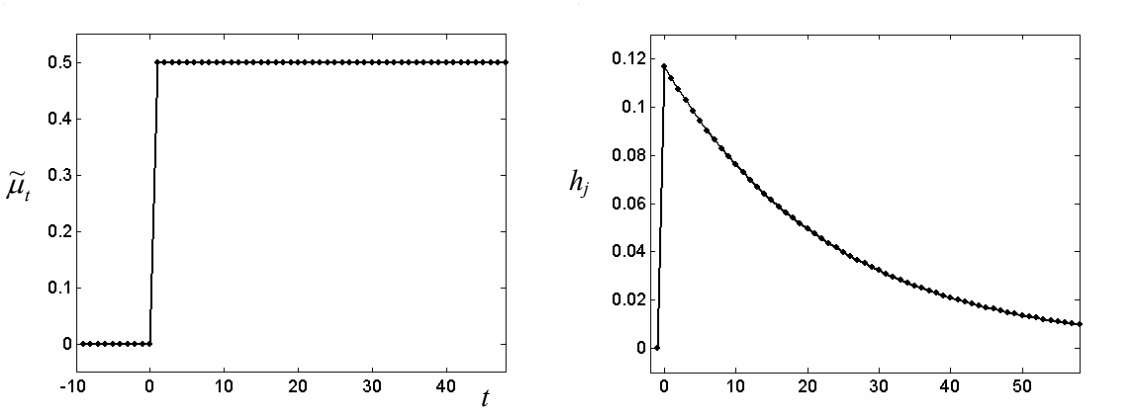
Table 3.2 shows that in the case of a small mean shift of $\mu = .5\sigma_a$, the optimal EWMA filter performs comparably with the optimal 2nd-order linear filter. However, the 2nd-order linear filter provides a more substantial advantage over the lower order EWMA filter as the magnitude of the mean shift increases. Examples 7 through 8, 9 through 12, and 17 through 19 show a gradual increase in the reduction of the out-of-control ARL with the optimal 2nd-order linear filter that is not matched by the optimal EWMA filter. This percentage of reduction reaches 54% for Example 8 with $\mu = 4\sigma_a$, 75% for Example 12 with $\mu = 4\sigma_a$, and 96% for Example 20 with $\mu = 3\sigma_a$.

For Examples 13 to 16 with a sinusoidal mean shift, the performance of both charts is influenced by the amplitude and the period of the mean shift. The EWMA chart performs very poorly with this kind of mean shift which has a small amplitude and a short period, such as in Examples 13, 14, and 15. By contrast, the 2nd-order linear filter

outperforms the EWMA filter by a wide margin and, moreover, shows faster detection with mean shifts that have shorter periods. This superiority of our 2nd-order linear filter over the EWMA filter for the sinusoidal mean shift is further discussed in Section III.6.3. Table 3.2 concisely summarizes how consistently the optimal 2nd-order linear filter outperforms the optimal EWMA filter.

III.6.3 Optimal Filter Characteristics

For the i.i.d. processes with step mean shifts (Examples 1 through 4 in Table 3.2), our search method ended up with the same parameters for the optimal 2nd-order linear filter as those for the optimal EWMA filter. Figure 3.3 shows the fault signature and impulse response coefficients (h_j) of the optimal 2nd-order linear filter for Example 1. h_j indicates how the past and present residuals e_{t-j} affect the present statistic y_t as shown in Equation (3.2).



(a) (b)

Figure 3.3. Example 1: (a) Fault Signature; (b) Impulse Response of the Optimal 2nd-order Linear Filter.

For the i.i.d. data, the optimal 2nd-order linear filter is found to perform best in the form of an EWMA filter with different λ's, which are determined based on the magnitude of the step mean shifts. For Examples 1 and 2, small λ's are more effective in detecting small mean shifts, since the impulse response coefficients of EWMA filters with small λs die out slowly. Conversely, large mean shifts are more rapidly detected by EWMA filters with larger λs as in Examples 3 and 4. A similar discussion is presented on the ARMA(1,1) process with $\phi_1 = .9$ and $\theta_1 = .5$ under a step mean shift (Examples 21 through 24 in Table 3.2).

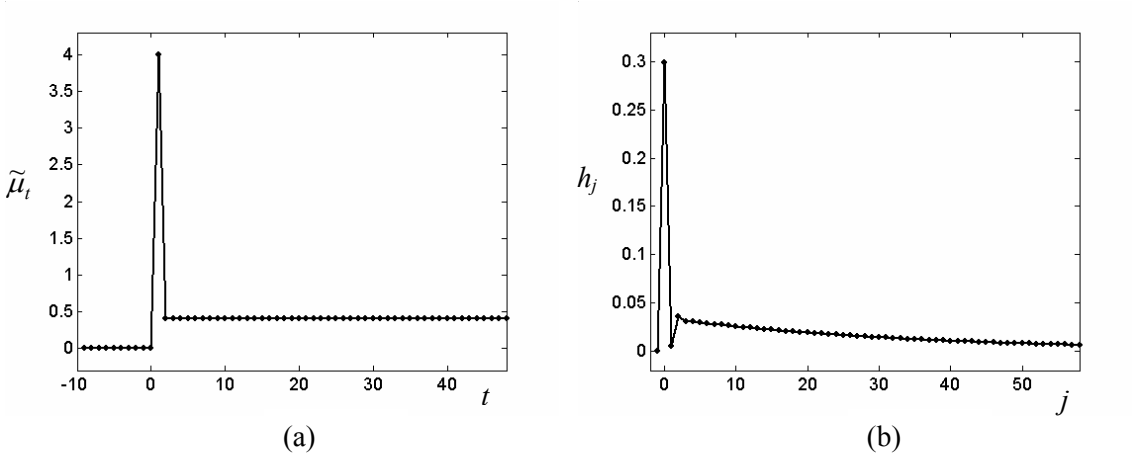


Figure 3.4. Example 8: (a) Fault Signature; (b) Impulse Response of the Optimal 2nd-order Linear Filter.

The AR(1) processes with a step mean shift (Examples 5 through 8 in Table 3.2) show a fault signature with a small steady-state magnitude after an initial single spike as seen in Figure 3.4(a). For step mean shifts of $.5\sigma_a$ and $1.5\sigma_a$, the optimal 2nd-order linear

filters are identical to the optimal EWMA filters – first-order linear filters. For larger mean shifts, the optimal 2nd-order linear filters detect shifts faster than the optimal EWMA filters. This difference can be explained in terms of the impulse response coefficient. Figure 3.4(b) shows the impulse response coefficients of the optimal 2nd-order linear filter for Example 8 – AR(1) process with a step mean shift of $4\sigma_a$. The optimal 2nd-order linear filter is closely related to a combined EWMA-Shewhart scheme. In this case, the optimal 2nd-order linear filter can be decomposed as

$$\begin{aligned} y_t &= k\left[\frac{1-\beta B}{1-\alpha_1 B-\alpha_2 B^2}\right]e_t \\ &= .29830\left[\frac{1-.84730B}{1-.86332B-.10469B^2}\right]e_t \\ &= \left[\frac{.26406}{1+.10780B}+\frac{.03423}{1-.97112B}\right]e_t \\ &\cong .26406e_t+\frac{.03423}{1-.97112B}e_t, \end{aligned} \tag{3.10}$$

which is a weighted combination of a Shewhart individual chart and a EWMA chart with λ = .02888 and k = .03423. This decomposition is graphically illustrated in Figure 3.5, where Figures 3.5(a) and (b) show the impulse response coefficients of the first term and the second term, respectively. For this case, the optimal filter takes advantage of the properties of both the EWMA filter and the Shewhart chart filter. As is well known, the Shewhart chart detects a large mean shift faster. Even if the optimal filter fails to detect the first large spike with the Shewhart chart filter component, it still has a chance to

detect the mean shift by the EWMA filter component that is covering the past observations.

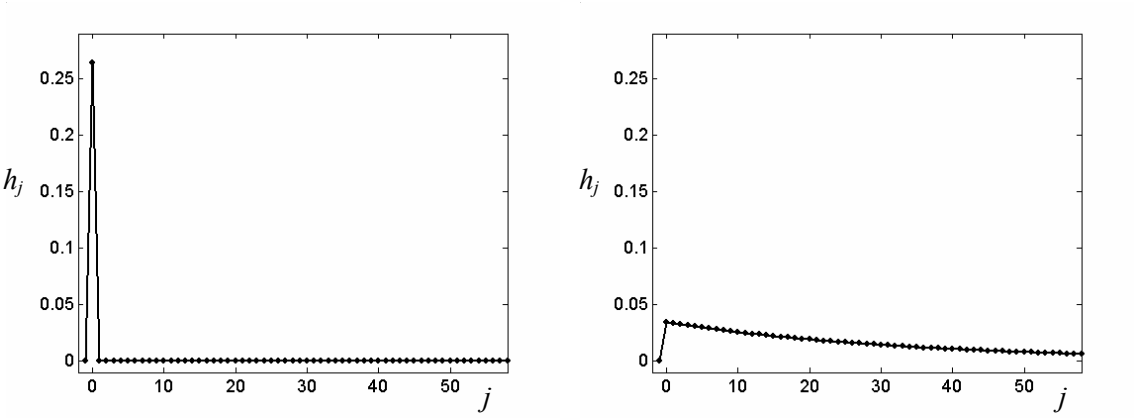


Figure 3.5. Decomposition of the Optimal Filter for Example 8: (a) Shewhart Chart Filter Component; (b) EWMA Filter Component.

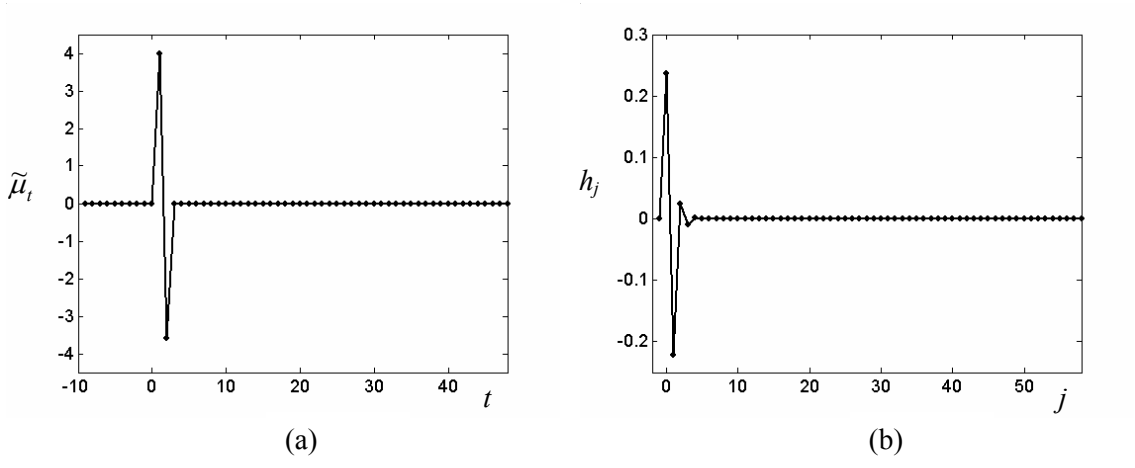
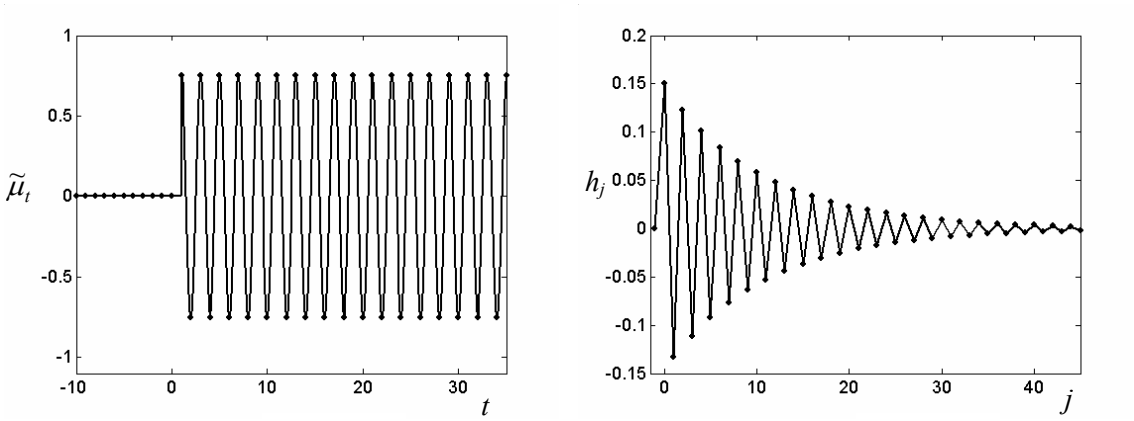


Figure 3.6. Example 12: (a) Fault Signature; (b) Impulse Response of the Optimal 2nd-order Linear Filter.

Similarly Lucas and Saccucci (1990) proposed a combined Shewhart-EWMA control scheme to achieve good performance for both small and large mean shifts simultaneously. The advantage of the optimal 2nd-order linear filter over their scheme is that it optimally combines the properties of both charts into one statistic, whereas the combined Shewhart-EWMA scheme is based on plotting both charts and the control limits of each.

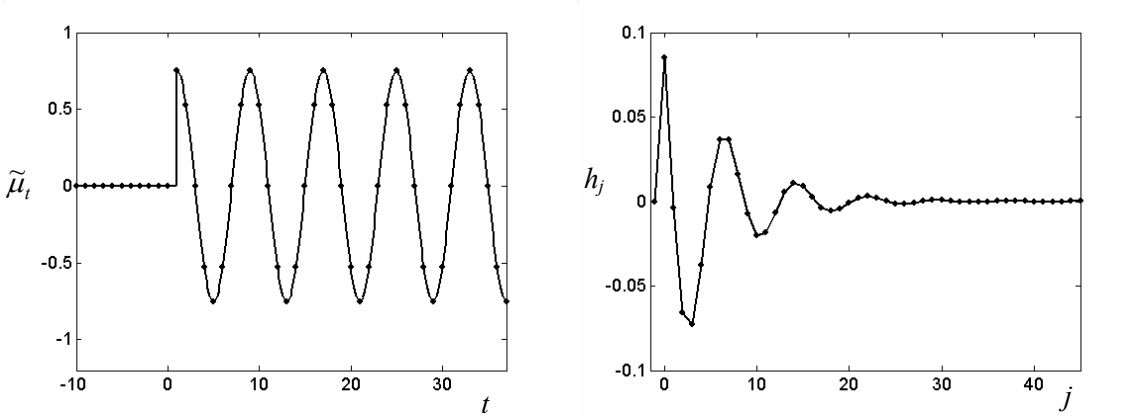


(a) (b)

Figure 3.7. Example 13: (a) Fault Signature; (b) Impulse Response of the Optimal 2nd-order Linear Filter.

With the AR(1) processes that have a spike mean shift (Examples 9 to 12 in Table 3.2), the advantage of our 2nd-order linear filter becomes more prominent as the magnitude of the mean shift increases. For mean shifts of $\mu = 3\sigma_a$ and $4\sigma_a$, optimal 2nd-order linear filters substantially surpass the corresponding optimal EWMA filters. From Equation (3.2), and Figure 3.6, we can see that the impulse response coefficients of the

optimal 2nd-order linear filter are much more effective in detecting residual mean shifts with this kind of spike mean shift than are those of the optimal EWMA filter. The more highly the fault signature is correlated with the impulse response coefficients, the larger the charted statistic y_t is that is generated by Equation (3.2). This advantage of a higher order filter is more substantial in Examples 13 through 16 in Table 3.2.



(a) (b)

Figure 3.8. Example 15: (a Fault Signature; (b) Impulse Response of the Optimal 2nd-order Linear Filter.

As shown in Figures 3.6, 3.7, and 3.8, our 2nd-order linear filter is optimized so that its impulse response coefficients are highly correlated with the residual means of the underlying process when the transient dynamics of the fault signature is pronounced. This high correlation between the impulse response coefficient and the residual mean contributes to an increase in the magnitude of the charted statistic y_t in Equation (3.2). This relationship can be easily explained in Example 13 with S_1. See Figures 3.7. As the

timestep moves forward, the mean residuals and the impulse response coefficients of the optimal filter show a positive correlation and a negative correlation by turns. Thus, the charted statistic y_t in Equation (3.2) comes out to be a large value each time even if the sign changes in turn. With this advantage, the optimal 2nd-order linear filter considerably improves the out-of-control ARL. Better performance results from the higher order structure of the 2nd-order linear filter compared to the EWMA filter. In other words, a process with a sinusoid mean shift oscillating around zero is not a situation where lower order filters, such as the Shewhart chart and the EWMA chart, can use their advantages to the full. There is an 86% reduction in the out-of-control ARL for Example 14. S_2 and S_4 have patterns of impulse response coefficients similar to those of S_1 and S_3, respectively.

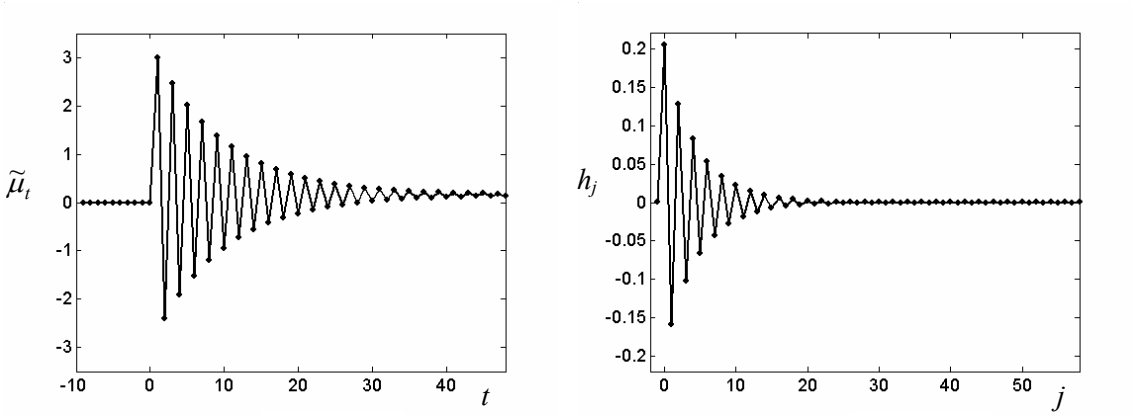
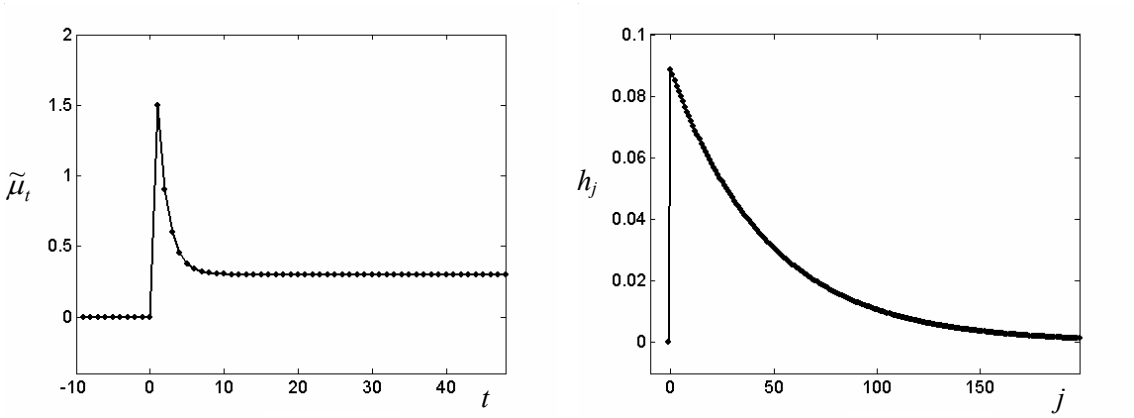


Figure 3.9. Example 20: (a) Fault Signature; (b) Impulse Response of the Optimal 2nd-order Linear Filter.

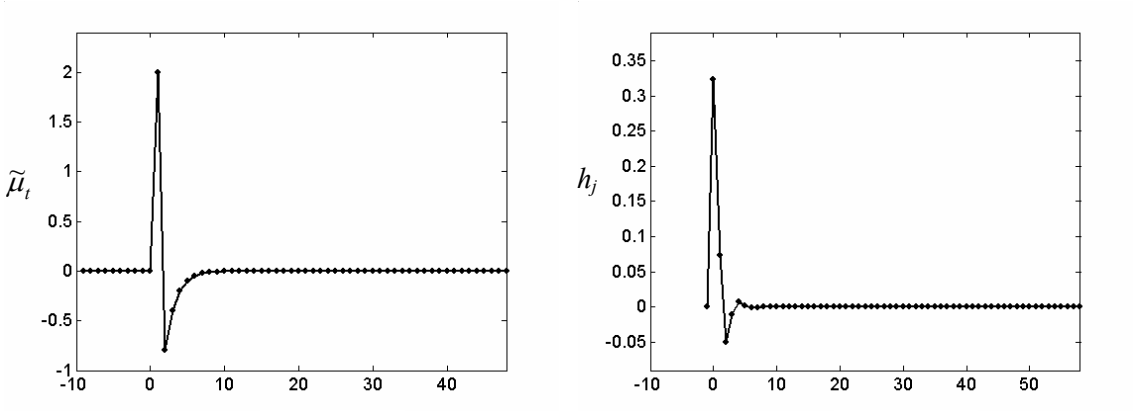
The advantage resulting from the high correlation is also apparent in Examples 17 to 20 – ARMA(1,1) processes with the transient dynamics of residual means oscillating around zero as shown in Figure 3.9(a). The example from Pandit and Wu (1983) in Section III.6.1 has an analogous fault signature. In the examples with mean shifts, $\mu/\sigma_X = \Delta = .5$ and 1 in Table 3.1, the advantage of the high correlation explains the superiority of the optimal 2nd-order linear filter over the other charts. The relationship can be used to select reasonable starting points in our gradient-based search. From Examples 9 to 16, the optimal filters for Examples 17 to 20 are expected to have impulse responses that are highly correlated with the mean of the residuals in order to increase the value of the charted statistic y_t in Equation (3.2) . Thus, we take A(B) = [1 + .9B] and M(B) = [1 + 0B] as the starting point for the search in our optimization strategy, and, in fact, it results in a satisfactory reduction in computational expense.



(m) (n)

Figure 3.10. Example 22: (a) Fault Signature; (b) Impulse Response of the Optimal 2nd-order Linear Filter.

For the ARMA(1,1) processes with $\phi_1 = .9$ and $\theta_1 = .5$ (Examples 21 through 24), the optimal 2nd-order linear filter converges to the optimal EWMA filter when step mean shifts occur, as shown in Figure 3.10. The 2nd-order linear filter for Example 23 has almost the same impulse response as the optimal EWMA chart, regardless of the order. In Examples 25 through 28 with spike mean shifts, the optimal EWMA reduces to the Shewhart chart. The optimal 2nd-order linear filters are also similar to the Shewhart chart, but show the higher correlation with the fault signature than the optimal EWMA. In these examples, the optimal 2nd-order linear filter outperforms only in the case with the largest mean shift since the transient dynamics of the fault signature are not pronounced. See Figure 3.11.



(a) (b)

Figure 3.11. Example 28: (a) Fault Signature; (b) Impulse Response of the Optimal 2nd-order Linear Filter.

III.7 Chapter Summary

The design procedure proposed in this chapter fine-tunes our 2nd-order linear filter to fulfill the specified optimization criterion for detecting a certain type of mean shift in an ARMA process. The result is a more systematic and automatic design procedure than existing heuristic algorithms for seeking the maximum performance of control charts. This optimization strategy, which is based on a derivative of ARL, searches optimal filters effectively and efficiently in terms of computational time and optimal convergence. With a Markov chain representation and a small number of filter parameters, the 2nd-order linear filter attains superiority over the GLF in terms of implementation, and it also performs almost as well as the OGLF in many situations. Moreover, the optimal 2nd-order linear filter provides a good starting point for the OGLF in cases requiring further fine-tuning of the SPC filter.

As shown in the examples, our 2nd-order linear filter is designed to fully utilize the design flexibility that originates from its higher order structure to provide maximum performance. More specifically, the parameters of the 2nd-order linear filter are optimally designed to provide advantageous properties for a given process. In some examples, the 2nd-order linear filter can be tuned to possess the beneficial properties of two existing charts such as the Shewhart chart and the EWMA chart. In a case where the fault signature has pronounced transient dynamics, it can utilize the high correlation of the filter with the fault signature of the process.

CHAPTER IV

MARKOV CHAIN METHOD BASED ON THE PARALLELOGRAM DISCRETIZATION

IV.1 Introduction

The integral equation method and the Markov chain method are generally used to approximate the ARL of control charts. The former is more accurate than the latter (Lucas and Croiser 1982). However, the integral equation method can not be used with certain kinds of control problems (Champ and Ridgon 1991). In fact, only the Markov chain method is applicable for the optimization procedures proposed in this dissertation. Although the Markov chain method is more versatile, it has an important limitation in implementation − a memory space problem due to the large state space. Although finer discretization results in better approximations, finer discretization also increases the dimensions of the transition probability matrix and thereby may cause out-of-memory errors as well as high computational expense. Prabhu and Runger (1996) provided some useful results to simplify the analysis of a two-dimensional Markov chain. The asymptotic formula provided by Brook and Evans (1972) has been used to extrapolate to a continuous scheme. An approximation using this formula requires ARL calculations for several discrete schemes. Hence, it is not desirable for our optimization procedures because the procedures require iterative calculation of ARL.

IV.2 Conventional Markov Chain Method

The conventional discretization approach uses vertical and horizontal lines to partition the two-dimensional state space as shown in Figure 4.1. The control limits of y and the region limits of z constitute the two-dimensional in-control region, which is partitioned into rectangles of $\delta_z = (UL_z - LL_z)/ N_z$ wide and $\delta_y = 2/N_y$ high. Therefore, the transition probability matrix is of dimensions $N \times N$, where $N = (N_z \times N_y)$. Jiang (2001) illustrates the conventional discretization approach for the ARMA(1,1) chart on i.i.d. data.

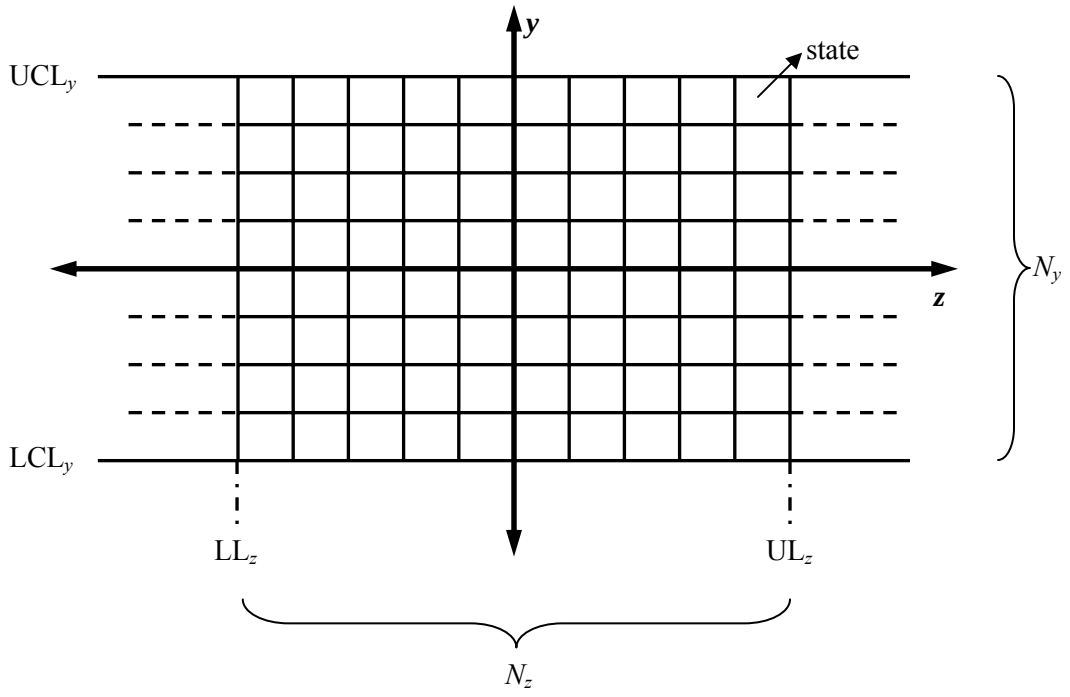


Figure 4.1. Conventional Discretization for Two-dimensional State Space.

The shape of the in-control region may be different depending on the charted statistic. The Multivariate EWMA (Runger and Prabhu 1996) has a circular in-control region, since the charted statistic T^2 is in quadratic form. On the other hand, the ARMA(1,1) chart of Jiang (2001) and the 2nd-order linear filter in Section III have a rectangular in-control region. As noted by Jiang (2001), a rectangular region needs more memory space for implementing the Markov chain method than does a circular one. If the asymptotic value is not used, moreover, larger N_z and N_z are required in order to obtain an approximation accuracy that is competitive with the integral equation method. Thus, the huge size of the transition probability matrix often causes out-of-memory error.

IV.3 Motivation Examples

Table 4.1. Comparison of the ARL Calculations

No.	*Time series model*		*2nd-order linear filter*				N_{mc}	ARL_0		
	ϕ_1	θ_1	α_1	α_2	β	k		*Markov chain method*	*Asymptotic*	*Simulated*
1	0	0	.85	0	0	.18115	41	500.3		499.8 (.99)
2	0	0	.85	0	.2	.21269	61	499.1		501.7 (.99)
3	0	0	.85	0	.9	.32215	91	493.0	501.4	500.6 (1.00)
4	.9	0	.78365	.86306	.10471	.27537	89	483.6	508.8	499.80 (.97)

Table 4.1 shows the in-control ARL calculations of the 2nd-order linear filters for 4 examples, where the one-dimensional Markov chain method is used for the first 2 examples and the two-dimensional Markov chain method is used for the other 2. Example 4 in Table 4.1 is identical to Example 7 in Table 3.2. Table 4.1 includes the in-

control ARL values obtained by the Markov chain method, their asymptotic values, and the ARL values based on a Monte Carlo simulation with 250,000 runs. The simulation standard errors are shown in parentheses. N_{mc} is the largest number of subintervals that can be used to implement the Markov chain method without causing out-of-memory error, where $N_{mc} = N_y = N_z$.

In Examples 1 and 2, the ARL values obtained by the Markov chain method are so reliable that the asymptotic value is not needed. The N_{mc} values for Examples 3 and 4 are 91 and 89, which are larger than those in Examples 1 and 2. However, they are not large enough to provide a good approximation. Using Equation (1.6), thus, the asymptotic values are calculated based on the ARL values for the discrete schemes with N_{mc} = 41, 51, 61, 71, and 81. The asymptotic value for Example 3 is reliable but the value for Example 4 shows some discrepancy, which indicates that even the asymptotic value can be unreliable in some cases. Runger and Prabhu (1996) and Jiang (2001) also discussed the appearance of this discrepancy. The asymptotic value is found to critically depend on the N_{mc} values of the discrete schemes for the approximation in Equation (1.6). Instead of N_{mc} = 41, 51, 61, 71, and 81, the asymptotic value based on N_{mc} = 21, 25, 29, 33, and 37 comes out to be 516.0 for Example 3. Therefore, the discrepancy in the asymptotic value for Example 4 can be explained in terms of insufficient discretizations for the discrete schemes.

As noted by Runger and Prabhu (1996), sometimes a quicker analysis is required in practical applications even though accuracy may be sacrificed as a result. In optimization procedures, however, accuracy and computation expense are both critical.

For this reason, the Markov chain method is not adequate for some optimization problems because obtaining accurate results may significantly increase the computational expense. For Examples 3 and 4, it takes 21.98 minutes and 12.81 minutes, respectively, to complete the calculation of the ARL for one discrete scheme with a N_{mc}, respectively. In addition, the results are not as reliable as the simulated ones. Despite relatively good accuracy, the asymptotic ARL using the Markov chain method might not be appropriate to a particular situation since it requires implementing the Markov chain method for several discrete schemes, thereby increasing the computational expense. In the following section, therefore, we propose a new discretization approach in order to substantially reduce the memory use and computational expense of the two-dimensional Markov chain method.

IV.4 Parallelogram Discretization

The number of partitioned subintervals required to provide a good approximation is based on the area of the in-control region and the charted statistic. In the case of the 2nd-order linear filter, the control limits for y_t are fixed at ± 1 and the limits for z_t are reasonably selected as discussed in Section III.3. Thus, the in-control region is subject to the limits for z_t. Interestingly, the V_t given V_{t-1} forms a single one-dimensional line with the slope of $-1/\beta$. As shown in Figure 3.1, the limits for z_t are influenced by the slope. In other words, the more gradual the slope is, the wider the area of z_t that is under the distribution line. As the limit interval for z_t increases from Example 1 to 4, the desirable number of partitioned subintervals also increases.

Considering the relationship between the desirable number of subintervals and the slope of the distribution line, the new discretization approach, called the Parallelogram Discretization (PD), is proposed. We discretize the state space horizontally as in the conventional discretiztion in Figure 4.1, but the vertical line has a slope equal to that of the distribution line as shown in Figure 4.2.

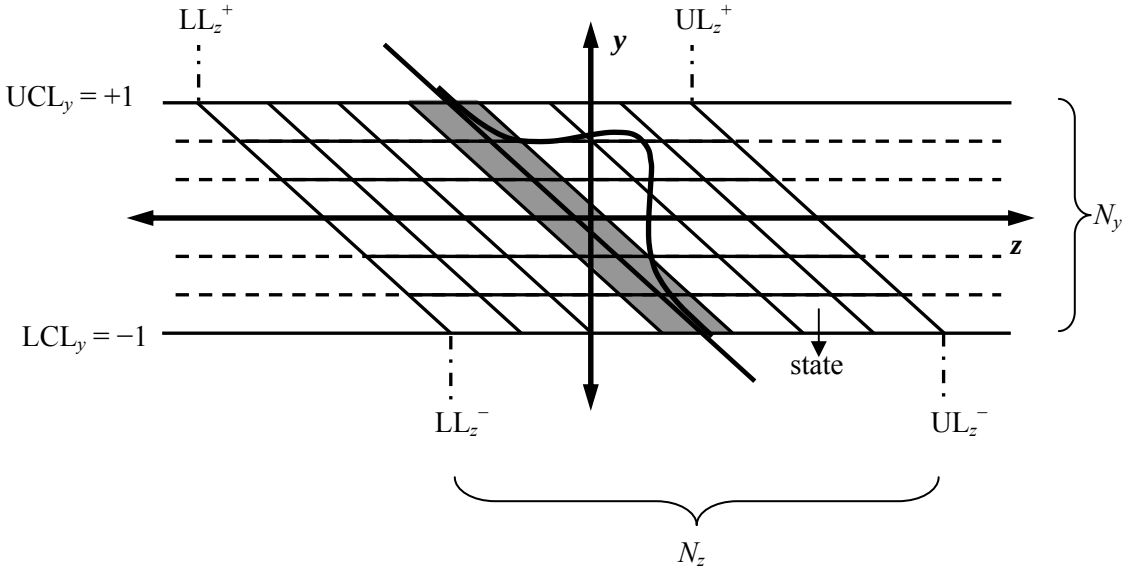


Figure 4.2. Parallelogram Discretization.

The PD surpasses the conventional discretization approach in computational expense, memory use, and accuracy. As discussed in Section III.3, the one-dimensional distribution line of the 2nd-order linear filter over the two-dimensional state space results in a sparse transition probability matrix. This sparse property is used to reduce the memory use with the SPARSE function of MATLAB 6.5, which converts a full matrix to sparse form by squeezing out any zero elements. Thus, the computational expense and

memory use are proportionate to the number of nonzero elements of Q. The PD generates a more sparse transition probability matrix Q. Each row of Q based on the PD has only N_y nonzero elements, whereas those based on the conventional discretization in Figure 3.1 have approximately $2\times\max\{N_y,N_z\}$ nonzero elements. Furthermore, the discretization along the z-axis can be more refined in the PD without increasing the number of non-zero elements. This property of the PD remarkably improves the accuracy over the conventional discretization approach. In the PD for the 2^{nd}-order linear filter, the number of partitioned subintervals along the y-axis is fixed at 41 and the number of subintervals along the z-axis is determined to make the width of each subinterval δ_z equal to .01.

IV.5 Performance Improvement over the Conventional Discretization Approach

This section compares the PD and the conventional discretization approaches applied to the examples in Section IV.2. The Markov chain method is implemented in MATLAB 6.5 on a computer with P4 3.2GHz and 512MB RAM. The N_y and N_z for the conventional discretization approach are determined to be the largest number of subintervals allowed within a limited memory space. The N_y and N_z for the PD are chosen as mentioned in Section IV.4. The numerical results are shown in Table 4.2. For Examples 3 and 4, the PD improves the accuracy of the Markov chain method over the conventional discretization approach. The ARL values are almost the same as the simulated ones. The reduction in computational time with the PD is very significant, especially for Examples 3 and 4.

The reduction in computational time is even more significant than the reduction of the nonzero elements of Q. The reason for this difference is that the conventional discretization approach requires additional computational time to identify those rectangles corresponding to nonzero transition probabilities and their intersections with the distribution line over the state space. Note that Example 4 has a longer computation time than Example 3 with a larger size Q, because its setup time is longer due to the more complex structure of the filter.

Table 4.2. Comparison of the PD and the Conventional Discretization

No.	Parallelogram Discretization Approach					Conventional Discretization Approach			
	N_y	N_z	Size of Q	Time* (min.)	ARL	$N_y = N_z$	Size of Q	Time* (min.)	ARL
1	41	23	943 × 943	.01	500.31	41	1681 × 1681	.31	500.3
2	41	43	1763 × 1763	.03	499.71	61	3721 × 3721	1.36	499.1
3	41	183	7503 × 7503	.10	499.93	91	8281 × 8281	21.98	493.0
4	41	181	7421 × 7421	.12	500.32	89	7921 × 7921	12.81	483.6

*: Computational Time.

VI.6 Chapter Summary

This chapter presents a new discretization approach to calculating ARL with the Markov chain method. The new approach allows us to refine the state space by fixing the number of nonzero transition probabilities and facilitates the computation of transition probability, which increases the accuracy of the ARL and significantly reduces the computational time and memory use. For all of the examples in Tables 3.2 and 4.2, the ARL values calculated by the Markov chain method using this approach are almost as accurate as the simulated values, whereas even the asymptotic values are not reliable

for some of them. This approach can be extended to higher state spaces where the advantage of this method is expected to be even more prominent.

CHAPTER V

OPTIMIZATION STRATEGY

V.1 Overall Strategy

The flowchart in Figure 5.1 represents the optimization strategy used to design OGLFs. We use a gradient-based method in order to optimally design the linear filters to provide a minimum out-of-control ARL under an in-control ARL constraint. The method keeps moving in the direction of the gradient to reduce the out-of-control ARL until it reaches an optimal solution. The parameters of the optimal filters are chosen to be optimal for detecting a specified mean shift. All of the procedures for designing the optimal filters were programmed in MATLAB which searches for the optimal filter for a specific mean shift using the initial values of several parameters representing the real process, the starting point of the search, the magnitude and type of the mean shift, and the number of partitioned subintervals along the axes.

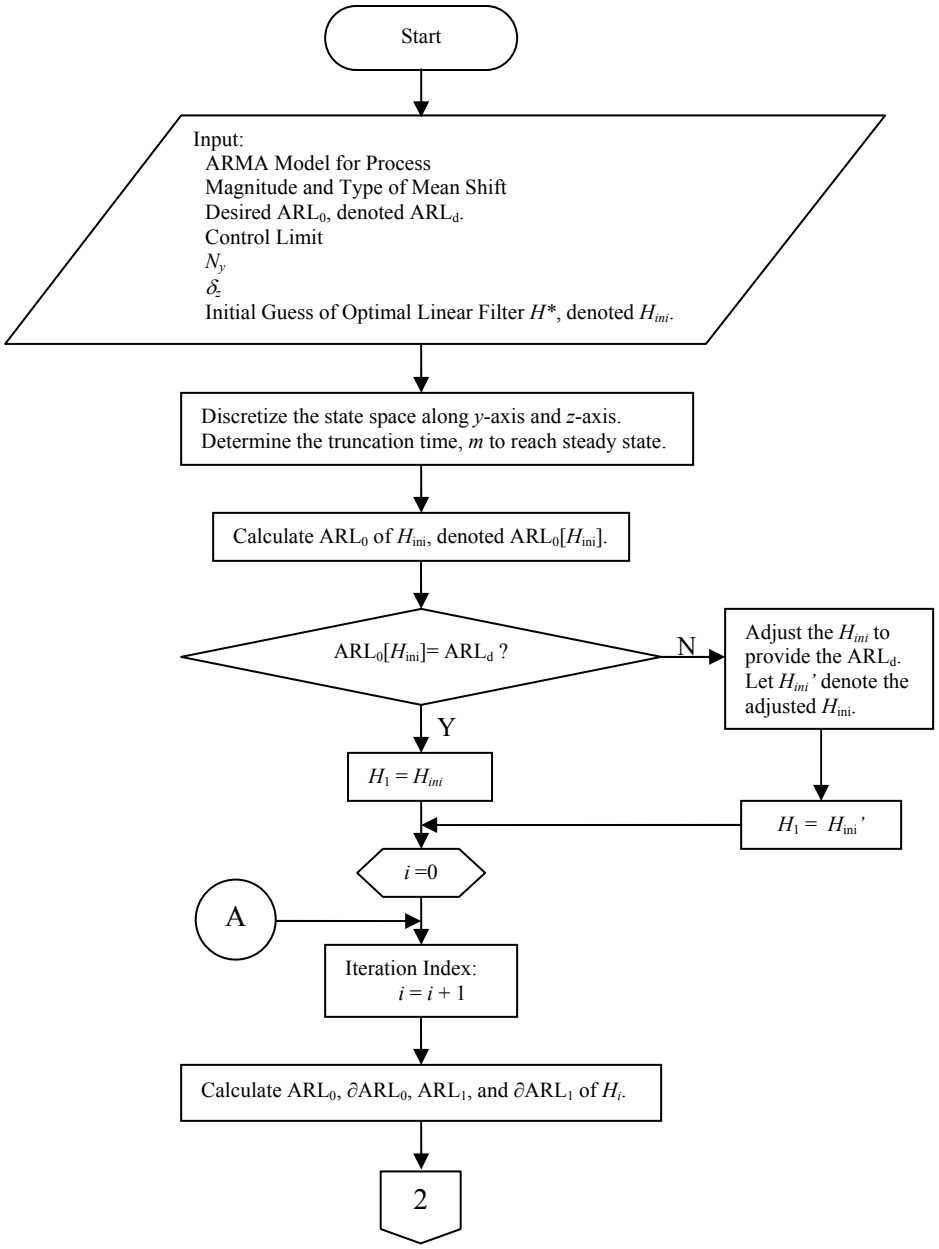


Figure 5.1. Flowchart of the Optimization Strategy.

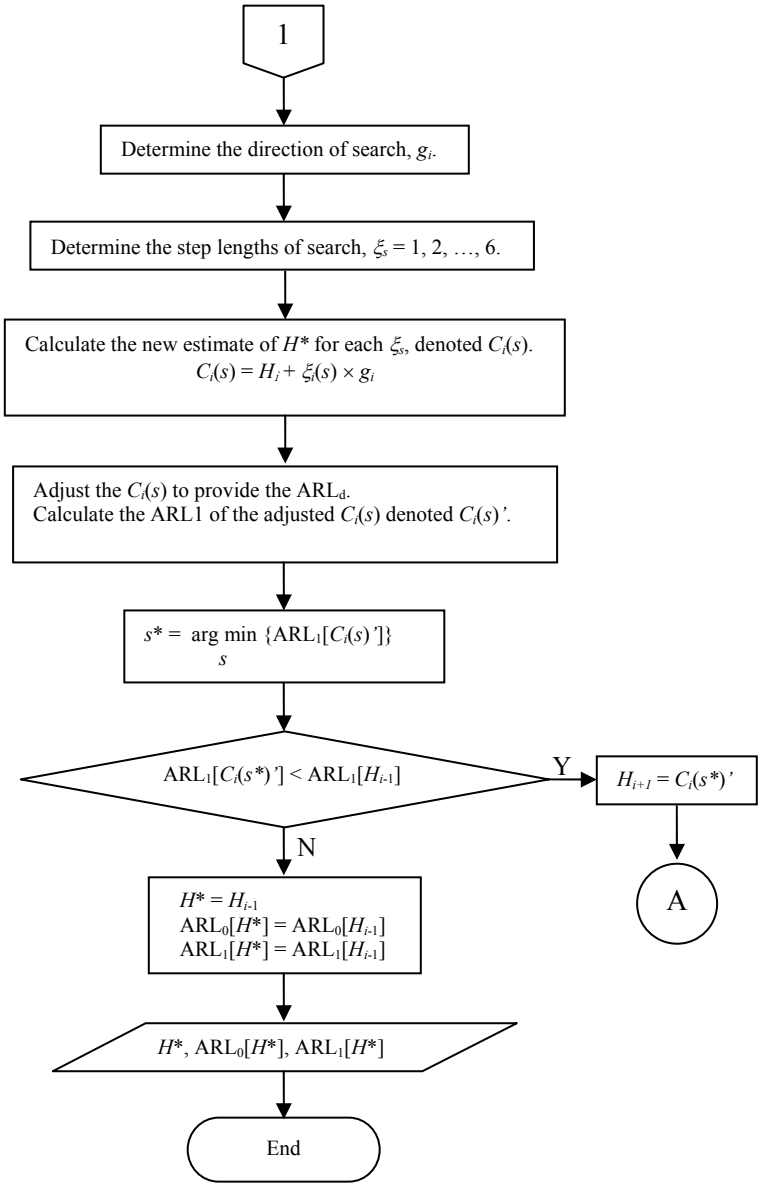


Figure 5.1. Continued.

V.2 Gradient-based Search

The optimization criterion for the GLFs and the 2nd-order linear filters is to minimize the ARL_1 while constraining the ARL_0 to some specific value (i.e., 500). To describe the gradient-based search, let H_i be the current estimate of a minimizer of ARL_1 at Iteration i, and let g_i be the search direction at point H_i. As shown in Equation 5.1, at Iteration i, 6 step lengths denoted $\xi_i(s)$ are used to calculate new estimates denoted $C_i(s)$, where $s = 1, 2, \ldots, 6$.

$$C_i(s) = H_i + \xi_i(s) \times g_i \tag{5.1}$$

For $s = 1, 2, \ldots 6$, the $C_i(s)$ is adjusted in the direction of $\partial\mathrm{ARL}_0$ to provide the desired ARL (see Figure 5.2). Let $C_i(s)'$ denote the adjusted $C_i(s)$. The ARL_1 of the $C_i(s)'$ is calculated next. Then, the $C_i(s)'$ with the minimum ARL_1 is assigned to the new estimate of the optimal linear filter H^*, denoted by H_{i+1} at Iteration i.

$$H_{i+1} = C_i(s^*)', \tag{5.2}$$

where $\mathrm{ARL}_1[C_i(s)']$ denotes the ARL_1 of the $C_i(s)'$ and $s^* = \arg\min_s \{\mathrm{ARL}_1[C_i(s)']\}$.

As shown in Figure 5.2, the search direction g_i is defined as the orthogonal projection of $\partial\mathrm{ARL}_1$ onto the space perpendicular to the $\partial\mathrm{ARL}_0$. Thus, the g_i is obtained as

$$g_i = -\partial ARL_1 + \left(\frac{\partial ARL_1 \bullet \partial ARL_0}{\left| \partial ARL_0 \right|^2} \right) \partial ARL_0 . \tag{5.3}$$

The gradient information substantially improves the optimization routine, especially for GLFs which must design all of the impulse response coefficients.

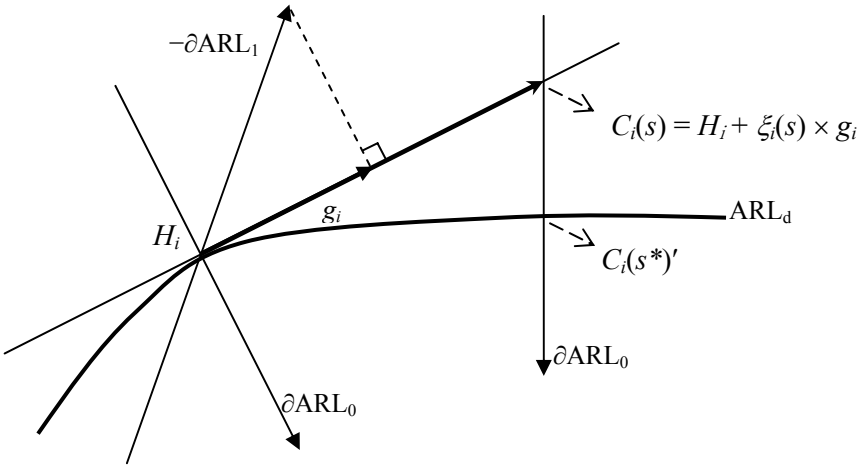


Figure 5.2. Gradient-based Search.

V.3 Selecting Starting Points of Search

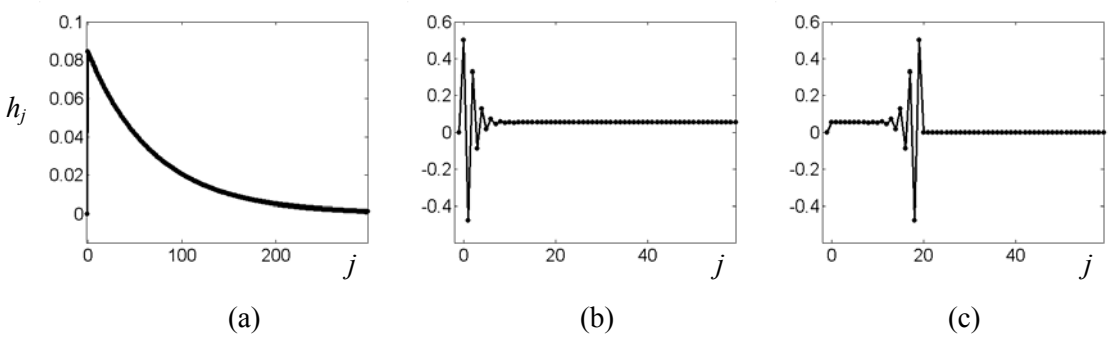
Since finding the global optimum is essential when using a search strategy in practice, avoiding being trapped in a local optimum is critical. Local search is based only on the information within a small area, and, therefore, when using local search methods, we cannot be convinced, without additional assumptions, that a chosen best solution is the global optimum. Random search is a good way out of the local optimum dilemma, but it sometimes leads to infeasibility due to the large search space required. Ideal search methods should have the merits of both types of searches in order to facilitate solid optimization. The issue of a local optimum is also very important in our gradient method

where our results vary according to the starting point of the search. Because the power of random search results from using random starting points, we begin our search from several reasonable points in each example.

As the starting point of a search, we make reasonable guesses as to the optimal linear filter. These guesses are intuitively selected based on the transient dynamics and the steady state value of the mean shift that are of interest. For the GLF, it could be one of the following: 1) the Shewhart chart filter, 2) the optimal EWMA filter, 3) the optimal 2nd-order linear filter, 4) the fault signature, and 5) the flipped fault signature. As a starting point, the flipped fault signature is defined as

$$h_j^{ini} = \begin{cases} \tilde{\mu}_{q-j} & 0 \le j < q \\ 0 & q \le j \le r \end{cases}, \tag{5.4}$$

where $q = \min\{20, t^*\}$; t^* is the largest t for $\tilde{\mu}_t > \varepsilon$; ε is a fixed positive constant sufficiently small; r is the window length of the GLF; $0 < t \le 20$; and h_j^{ini} is the $(j+1)^{th}$ impulse response coefficient of the initial guess of the optimal linear filter H(B) for the underlying process of the mean shift that is of interest. As the starting points of the GLF in Example 1 with $\Delta = .5$ in Table 3.1, the optimal 2nd-order linear filter (= the optimal EWMA filter), the fault signature, and the flipped fault signature shown in Figure 5.3 were used. The search starting from only the fault signature converges to the best GLF. Table 5.1 lists the starting point converging to the OGLF for each example considered in Sections II.4.1 and II.4.2. The starting points for the 2nd-order linear filters are similarly selected.



(a) (b) (c)

Figure 5.3. Starting Points of the Optimization Search for Example 1 in Table 3.1: (a) Optimal 2nd-order Linear Filter; (b) Fault Signature; (c) Flipped Fault Signature.

Table5.1. Starting Points Converging to the OGLF

No.	Best Starting Point	No.	Best Starting Point
1	Optimal EWMA	17	EWMA
2	Optimal EWMA	18	Fault Signature
3	Optimal EWMA	19	Fault Signature
4	Optimal EWMA	20	Fault Signature
5	Optimal EWMA	21	2nd-order linear filter
6	Optimal EWMA	22	EWMA
7	Optimal 2nd-order linear filter	23	2nd-order linear filter
8	Optimal 2nd-order linear filter	24	Fault Signature
9	Fault Signature	25	Flipped Fault Signature
10	Fault Signature	26	Flipped Fault Signature
11	Optimal 2nd-order linear filter	27	Flipped Fault Signature
12	Optimal 2nd-order linear filter	28	Flipped Fault Signature
13	Optimal 2nd-order linear filter	1*	Fault Signature
14	Optimal 2nd-order linear filter	2*	Fault Signature
15	Shewhart	3*	Shewhart
16	Fault Signature	4*	Shewhart

*: Examples of Table 2.2 in Section II.4.1.

CHAPTER VI

CONCLUSIONS AND FUTURE WORK

VI.1 Conclusions

In this dissertation, control charting schemes are generalized in terms of linear filtering and two linear filters are proposed as new control charting schemes. The optimal design methodologies for these filters are developed based on the Markov chain method. A new discretization approach for the Markov chain method and a gradient-based optimization strategy enable the implementation of optimal design methodologies in two-dimensional space by reducing the computational expense and memory use. This research forms a general basis for more powerful and broad control charting methods.

The ARL performance of the optimized linear filters − the OGLF and the optimal 2^{nd}-order linear filter − is compared with that of the residual-based Shewhart chart, the PID chart, and the optimal EWMA. The optimal linear filters significantly 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 optimal linear filters 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 that is mentioned in Section II.4.3 and III.6.3, the flexibility of the filter structure plays a key

role in determining its performance. Additional flexibility from a higher order filter guarantees better detection capability for more kinds of mean shifts. In other words, in this capacity, the optimal linear filters are superior to the Shewhart chart, EWMA chart, and the ARMA(1,1) chart for ARMA(1,1) processes of Jiang et al. (2000). Our optimal filter design procedures are programmed in MATLAB. The programs search for an optimal linear filter beginning with the initial guesses of the optimal linear filter. Neither a heuristic algorithm nor a reference table is needed to find the optimal filter.

VI.2 Future Work

- This research has considered various types of mean shifts (step, spike, sinusoidal) and various ARMA processes. Optimal linear filters perform well for several combinations of mean shifts and processes. However, this research is restricted to processes with one type of mean shift at a time. In reality, several types of mean shifts may happen within one process at the same time. Optimal filter designs for such processes should be investigated in the future.
- This research is restricted to detecting deterministic mean shifts. It can be extended to optimization over a distribution of mean shift magnitude.
- The search of the gradient-based optimization strategy starts from an intuitively selected initial guess of the optimal linear filter. The starting point significantly influences the computational time and the optimality of the resulting linear filters. In many examples, the fault signature turns out to be a good starting point, showing a high correlation with the optimal linear filter. However, selecting

reasonable initial guesses for the optimal linear filters based on preliminary information about the underlying process should be an area for future study. More work is also needed on methods for avoiding being trapped in local optima.

- In this research, the ARMA model for the underlying process is assumed to exist and be known. In practice, the parameters of the ARMA model are unknown and are estimated. Sensitivity analysis, the relaxation of the assumption, and robust design methodologies are all areas that should be studied further.
- In all of the examples considered in this research, the PD provides results that are almost as accurate as those from Monte Carlo simulation. However, the accuracy of the Markov chain method based on the PD still depends on the number of partitioned subintervals. Thus, there is still a memory limitation with this method. In order to resolve the memory space problem that is due to the fine discretization, a design methodology using the integral equation method is under investigation. In addition, if the mathematical expression for the relationship between the discrepancy of the ARL and the parameters of the control chart can be identified, the optimization routine will be remarkably improved and also free of out-of-memory error.
- The OGLF performs best, but it is less practical because of its high computational expense. Thus, the 2^{nd}-order linear filter is developed to simplify the optimization procedure. However, there is still much room for improvement in this method in terms of computation and performance. Extensions to higher-

order linear filters and the consequent computational expense problem need further investigation.

REFERENCES

Alwan, L. C., and Roberts, H. V. (1988), "Time-Series Modeling for Statistical Process Control," *Journal of Business & Economic Statistics*, 6, 87-95.

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

Artiles-León, N., David, H. T., and Meeks, H. D. (1996), "Statistical Optimal Design of Control Charts with Supplementary Stopping Rules," *IIE Transactions*, 28, 225-236.

Bagshaw, M., and Johnson, R. A. (1975), "The Effect of Serial Correlation on the Performance of CUSUM Tests II," *Technometrics*, 17, 73-80

Baker, C. T. H. (1977), *The Numerical Treatment of Integral Equations*, Oxford, England: Clarendon Press.

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

Champ, C. W. and Rigdon, S. E. (1991), "Comparison of the Markov Chain and the Integral Equation Approaches for Evaluating the Run Length Distribution of Control Charts," *Communications in Statistics: Simulation and Computation*, 20, 191-204.

Crosier, R. B. (1986), "A New Two-Sided Cumulative Sum Quality Control Scheme," *Technometrics*, 28, 187-194.

Crowder, S. V. (1987), "A Simple Method for Studying Run-Length Distributions of Exponentially Weighted Moving Average Charts," *Technometrics*, 29, 401-407.

Crowder, S. V. (1989), "Design of Exponentially Weighted Moving Average Schemes," *Journal of Quality Technology*, 21, 155-162.

Harris, T. J., and Ross, W. H. (1991), "Statistical Process Control Procedures for Correlated Observations," *Canadian Journal of Chemical Engineering*, 69, 48-57.

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

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

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

Johnson, R. A., and Bagshaw, M. (1974), "The Effect of Serial Correlation on the Performance of CUSUM Tests," *Technometrics,* 16, 103-112.

Johnson, R. A., and Wichern, D. W. (1998), *Applied Multivariate Statistical Analysis* (4th ed.), Upper Saddle River, NJ: Prentice-Hall.

Jones, L. A. (2002), "The Statistical Design of EWMA Control Charts with Estimated Paramters," *Journal of Quality Technology*, 34, 277-288.

Kantorovich, L. V., and Krylov, V. I. (1964), *Approximate Methods of Higher Analysis*, New York: John Wiley.

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

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

Lucas, J. M., and Crosier, R. B. (1982), "Fast Initial Response for CUSUM Quality-Control Schemes: Give Your CUSUM A Head Start," *Technometrics*, 24, 199-205.

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

Page, E. S. (1954), "Continuous Inspection Schemes," *Biometrika*, 41, 100-114.

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

Parkhideh, S., and Parkhideh, B. (1998), "Design of a Flexible Zone Individuals Control Chart," *International Journal of Production Research*, 36, 2259-2267.

Prabhu, S. S., and Runger, G. C. (1996), "Analysis of a Two-Dimensional Markov Chain," *Communications in Statistics - Simulation and Computation*, 25, 75-80.

Renolds, M. R., Jr. (1995), "Evaluating Properties of Variable Sampling Interval Control Charts," *Sequential Analysis*, 14, 59-97.

Reynolds, M. R., Jr., Amin, R. W., and Arnold, J. C. (1990), "CUSUM Charts With Variable Sampling Intervals," *Technometrics*, 32, 371-384.

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

Runger, G. C., Willemain, T. R., and Prabhu, S. (1995), "Average Run Lengths for CUSUM Control Charts Applied to Residuals," *Communications in Statistics - Theory and Methods*, 24, 273-282.

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

Vasilopoulos, A. V., and Stamboulis, A. P. (1978), "Modification of Control Chart Limits in the Presence of Data Correlation," *Journal of Quality Technology*, 1, 20-30.

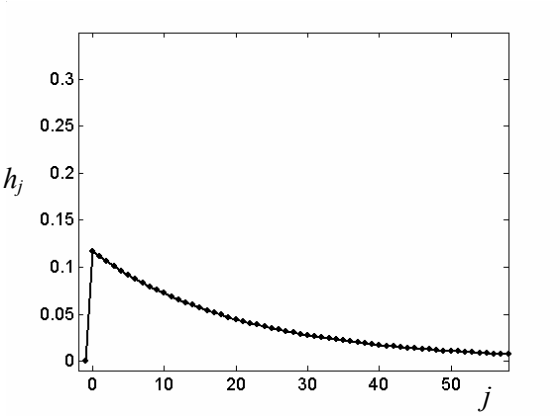
Yang J. and Makis V. (1997), "On the Performance of Classical Control Charts Applied to Process Residuals," *Computers and Industrial Engineering*, 33, 121-124.

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

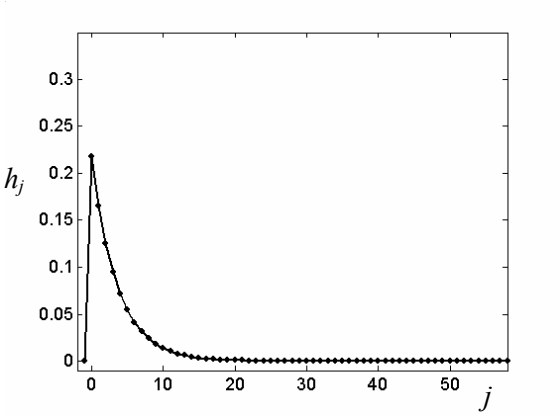
APPENDIX A

OPTIMAL GENERAL LINEAR FILTERS IN TABLE 2.3

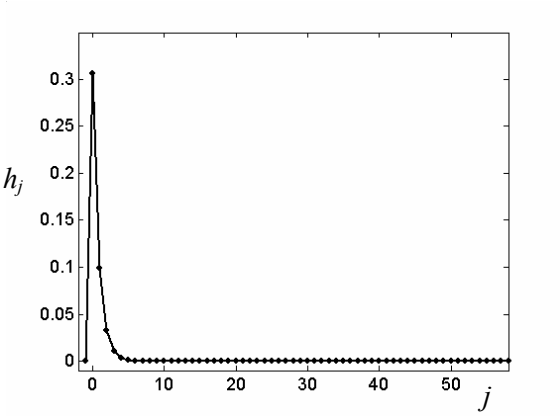
This appendix graphically shows the OGLFs for the 28 examples in TABLE 2.3.



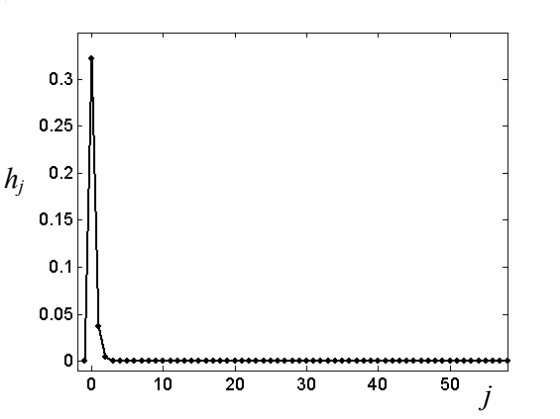
Ex 1



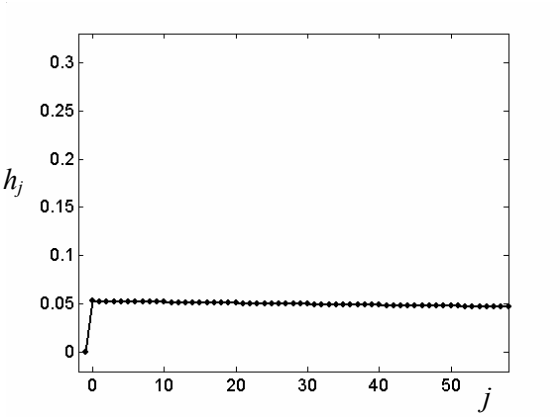
Ex 2



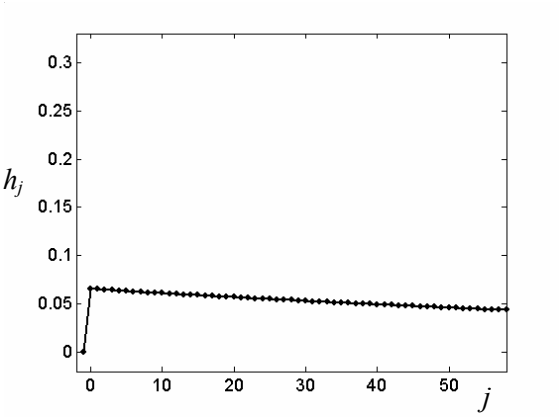
Ex 3



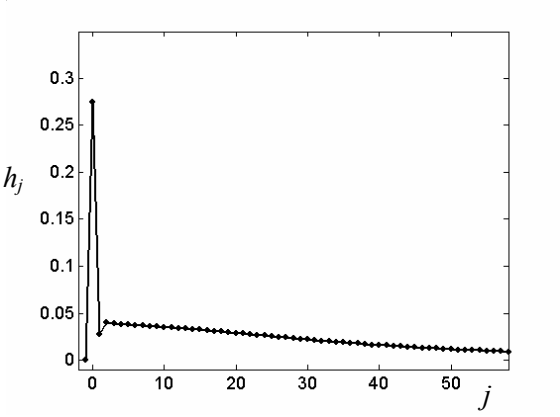
Ex 4



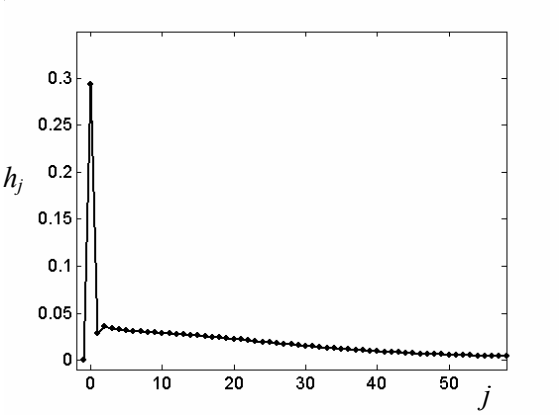
Ex 5



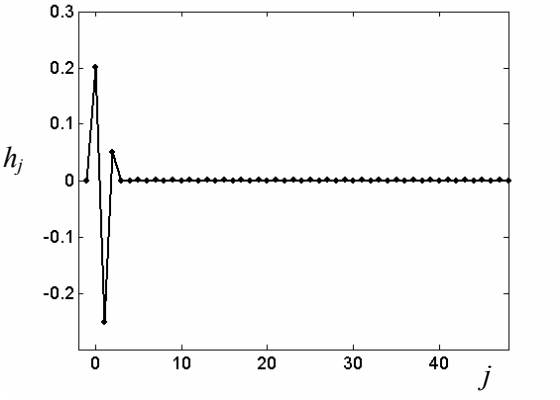
Ex 6



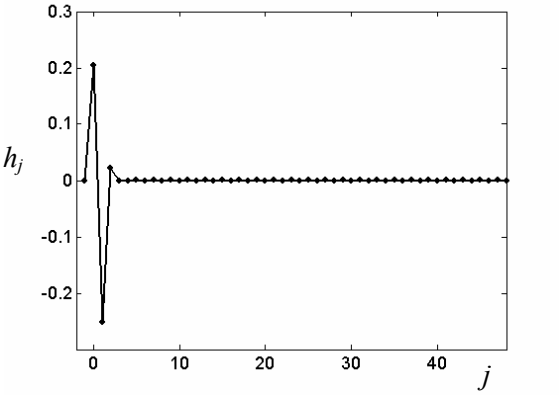
Ex 7



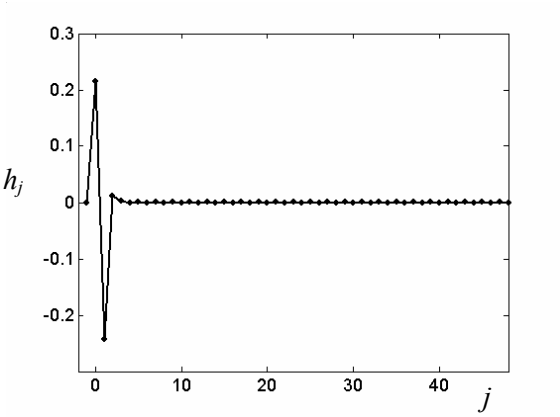
Ex 8



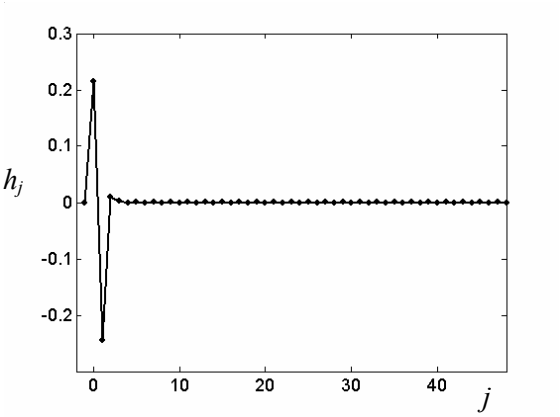
Ex 9



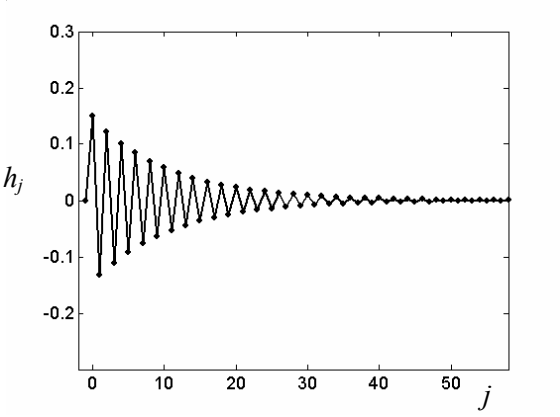
Ex 10



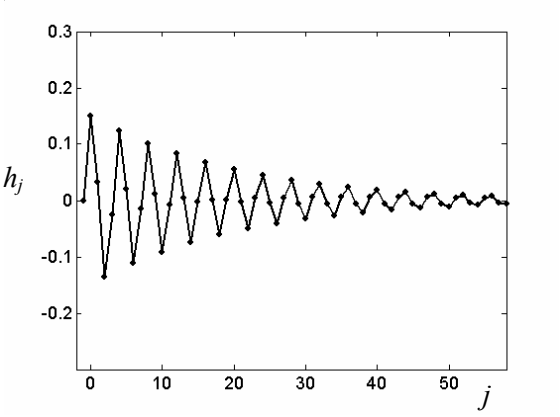
Ex 11



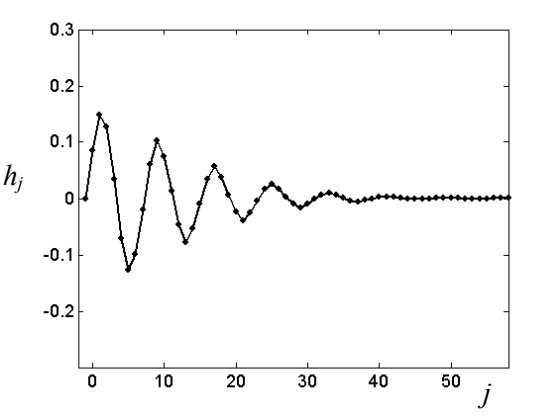
Ex 12



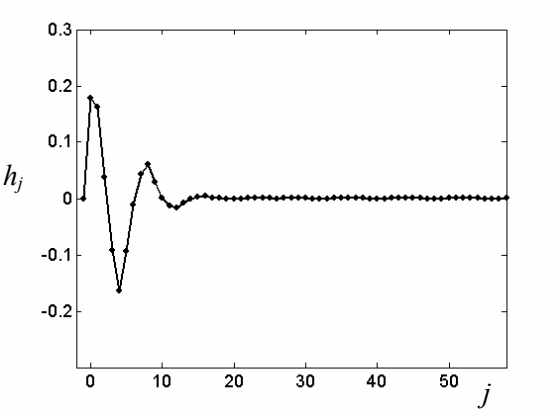
Ex 13



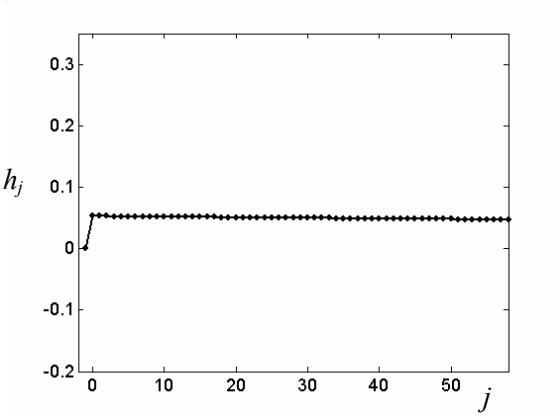
Ex 14



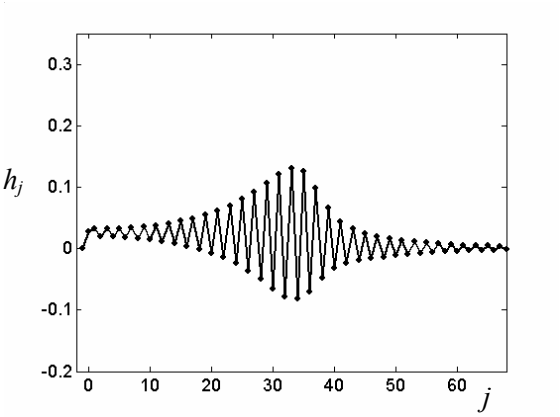
Ex 15



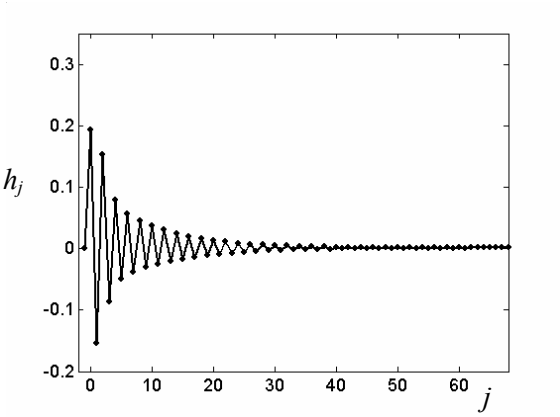
Ex 16



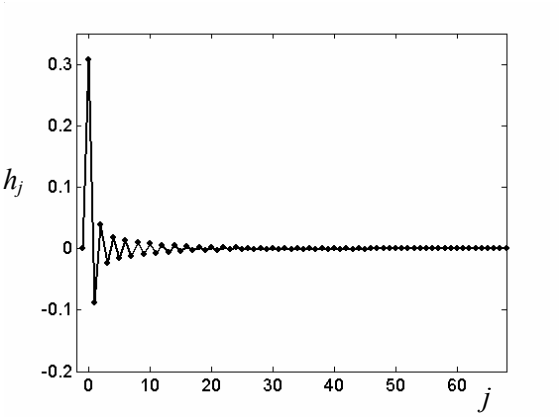
Ex 17



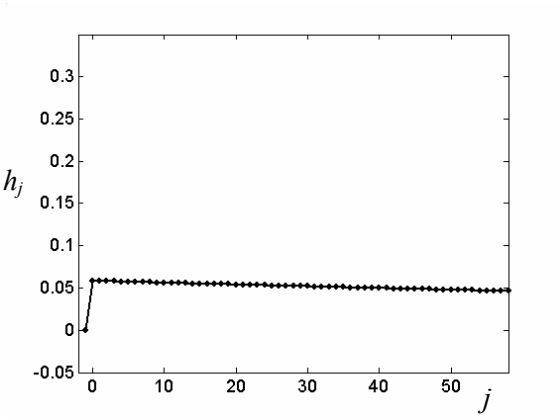
Ex 18



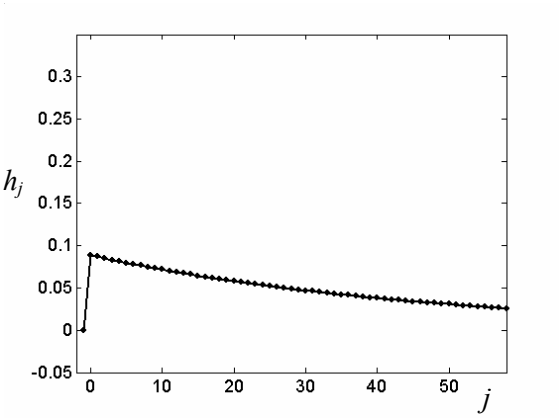
Ex 19



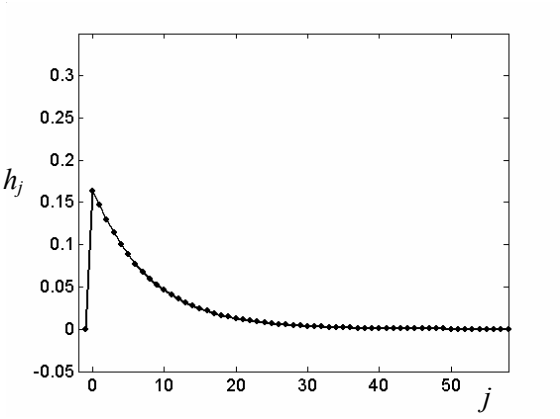
Ex 20



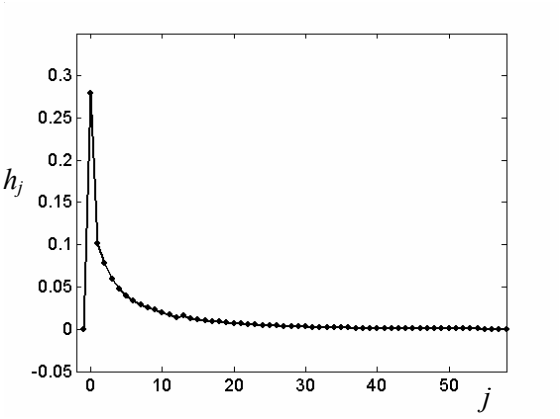
Ex 21



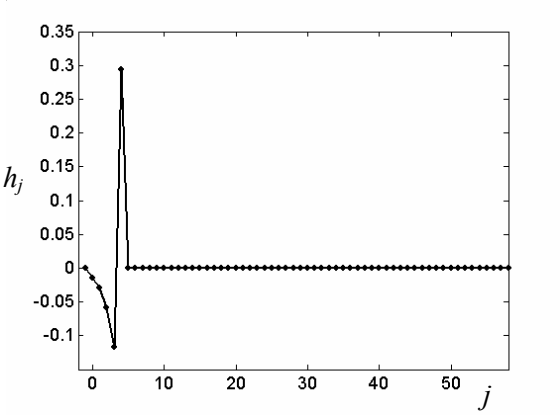
Ex 22



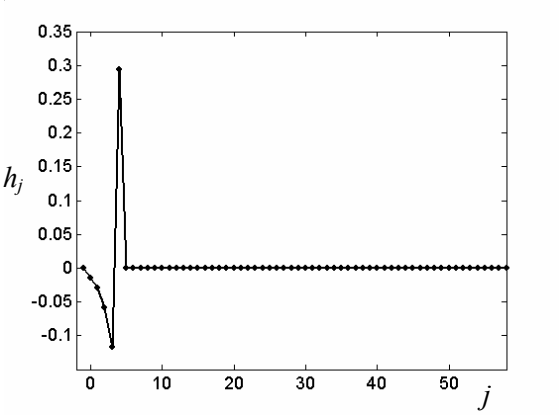
Ex 23



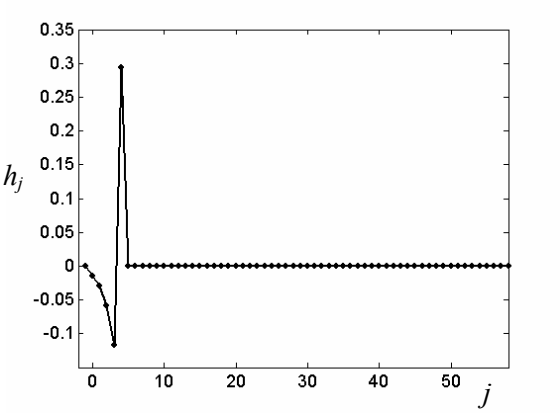
Ex 24



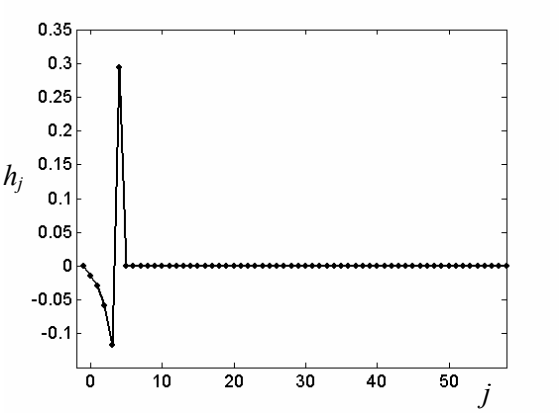
Ex 25



Ex 26



Ex 27



Ex 28

VITA

CHANG-HO CHIN

Permanent Address

Sangdae Hanbo APT 105-1105,
Sangdae 2 Dong, Jinju, Kyung-Nam, Rep. of Korea
660-764

Education

- Ph.D. in Industrial Engineering, Texas A&M University, College Station, Texas, August 2004
- M.S. in Industrial Engineering, Texas A&M University, College Station, Texas, December 1999
- B.S. in Industrial Engineering, Korea University, Seoul, Rep. of Korea, August 1996

Professional Experience

Research Assistant

- Team leader of the project "*Characterizing and Diagnosing Manufacturing Variation with In-process Measurement Data*", sponsored by Higher Education Coordinating Board, TX. (September 2002 − August 2004)

- Team leader of the project "*Defect Detection and Prevention in Printed Circuit Board Assembly Via Information Integration*", sponsored by Solectron, Austin, TX. (January 2000 – August 2002)

- Team leader of the project "*Statistical Process Control for Low-Volume Composite Manufacturing*", sponsored by Bell Helicopter Textron Inc., Dallas, TX. (June 1999 – February 2000)