CHANGE DETECTION
IN AUTOCORRELATED PROCESSES

by

Jiangbin Yang

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Mechanical and Industrial Engineering
University of Toronto

©Copyright by Jiangbin Yang 1999
The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author’s permission.

L’auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L’auteur conserve la propriété du droit d’auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.
Change Detection in Autocorrelated Processes

Jiangbin Yang, Ph.D.

Department of Mechanical and Industrial Engineering

University of Toronto, 1999

Abstract

The problem of change detection is about quick detection of a change in a dynamic system or process at a low rate of false alarm by sequentially observing the system or process. It has important applications in quality control, signal processing and other areas. This thesis studies the problem in the context of autocorrelated processes.

First, we systematically investigate the properties of process residuals (one-step ahead forecast errors) for change detection. We show that process residuals are statistically sufficient for the problem of change detection, and change detection can be done by using process residuals. We show that process residuals are mutually uncorrelated with zero means when there is no change to the process, that is, when the process is in-control. We develop a general procedure for specifically deriving the forms of process residuals. Using the procedure, we derive the forms of residuals of general autoregressive integrated moving average (ARIMA) processes and state space models, and obtain some specific properties of the residuals under several situations. Under the Gaussian assumption, for an ARIMA process or a steady-state state space model subject to a change in process mean level, the residuals are i.i.d. with zero means before the occurrence of change. After the occurrence of change, the residuals are still mutually independent with the same variance as before, but with time-varying and generally nonzero means. For an autocorrelated process subject to a change in process mean level, we find that the properties of residuals are independent of the feedback control
applied to the process.

We then concentrate on detection of a change in process mean level in an autocorrelated process by using the process residuals. Cumulative sum (CUSUM), exponentially weighted moving average (EWMA) and Shewhart control chart procedures are applied to the residuals. For computation of the average run lengths (ARLs) of the control chart procedures applied to the residuals whose means are time-varying after change, we derive an explicit formula for Shewhart, establish integral equations for CUSUM and EWMA, and develop efficient numerical procedures for solving the integral equations. Under the ARL criterion, we numerically study the performance of the control chart procedures applied to the residuals of autocorrelated processes under several situations.

We study the likelihood ratio (LR) testing procedure applied to the process residuals. We extend the classical LR testing procedure by replacing its constant threshold value with a time-varying threshold sequence. We propose a combined CUSUM and Shewhart control chart procedure to approximate the extended LR testing procedure. We develop numerical procedures based on integral equations for computation of the ARLs of these change detection procedures, and numerically study their performance.

We study the problem of optimal sequential testing on process mean levels. The problem is a special situation of change detection, an extension of the Wald's sequential probability ratio testing (SPRT) problem, and has wide application in signal processing and other areas. We formulate the problem as an optimal stopping problem, derive the optimal stopping rule, obtain some important properties of the optimal stopping boundaries, develop a numerical procedure for computation of the optimal stopping boundaries, and present a numerical example.
To my father, Jin Yang

who is so warm-hearted and diligent

and

to whom I feel deeply indebted
Acknowledgments

I thank my supervisor, Prof. Viliam Makis, who has offered me valuable guidance and assistance during the course of my Ph.D. study.

I thank Mr. Youhua Chen for his friendly suggestion made to me in Spring, 1994, which inspired me to pursue my Ph.D. in the engineering field.

I thank the members of my departmental oral examination committee, Prof. Viliam Makis (Supervisor and Chair), Prof. Andrew K. S. Jardine, Prof. Subbarayan Pasupathy (Department of Electrical and Computer Engineering), Prof. Morton J. M. Posner, and Prof. Jim G. C. Templeton, for their valuable comments and feedback on the manuscript of my thesis.

I thank the external appraiser, Prof. Hoang Pham of the Department of Industrial Engineering, Rutgers University, for his valuable appraisal on my thesis.

I acknowledge the Connaught Scholarships offered to me in 1994, 1995 and 1996 by the University of Toronto. I acknowledge the Ontario Graduate Scholarships offered to me in 1996, 1997 and 1998. I acknowledge the continuous financial support since January, 1995 from the Condition Based Maintenance (CBM) Consortium at the University of Toronto led by the principal investigators Prof. Andrew K. S. Jardine and Prof. Viliam Makis. I acknowledge the financial support from Prof. Viliam Makis’ NSERC fund.

I thank Ms. Jayne Beardsmore, an efficient administrator of the CBM Consortium, who has been very helpful and kind. I thank Mr. Oscar del Rio, a knowledgeable Computing Administrator of our department, who has offered me much prompt and effective help. I thank Ms. Brenda Fung and the graduate office who have efficiently helped me organize the final stage of my Ph.D. study.

I thank my wife, Hongmei Cao, who has accompanied me for many days and nights of hard work, who has understood and appreciated me, who has shared many things with me, and who has supported and encouraged me.
Contents

Abstract ii

Acknowledgments iv

1 Introduction 1

1.1 The Problem ................................................. 2

1.2 Two Production Processes .................................... 5

1.2.1 A Machining Process ...................................... 5

1.2.2 A Cutting Process ........................................ 6

1.3 Problem Formulation of Change Detection ..................... 8

2 Residuals for Model Transformation 10

2.1 Introduction .................................................. 11

2.2 Sufficiency of Residuals for Change Detection ................ 13

2.3 Second-Order Properties of Residuals ......................... 15

2.4 Procedure for Deriving Forms of Residuals ..................... 17

2.5 Properties of Residuals of ARIMA Processes ................... 18

2.5.1 Processes Subject to Additive Changes ..................... 18
2.5.2 Processes Subject to Nonadditive Changes .................................................. 20
2.6 Properties of Residuals of State Space Models .............................................. 23
  2.6.1 Residuals of a General Model ................................................................. 23
  2.6.2 Residuals of the Machining Process .......................................................... 25
  2.6.3 Residuals of the Steady-State Machining Process ..................................... 28
2.7 Independence of Residuals of Feedback Control ........................................... 31
2.8 Summary and the Model Transformation ....................................................... 33

3 Performance of Control Charts Applied to Residuals .................................... 35
  3.1 Introduction .................................................................................................. 36
  3.2 Integral Equations for Computation of the ARLs .......................................... 40
    3.2.1 Shewhart Control Chart Procedure ....................................................... 40
    3.2.2 CUSUM Control Chart Procedure ......................................................... 41
    3.2.3 EWMA Control Chart Procedure ............................................................. 43
  3.3 Numerical Analysis ....................................................................................... 45
    3.3.1 Results on the Machining Process Subject to a Step Shift ....................... 46
    3.3.2 Results on the Machining Process Subject to a Drift Rate Change ............ 49
    3.3.3 Results on the Cutting Process ................................................................. 52
    3.3.4 Summary of the Performance of Control Charts ...................................... 55

4 Sequential Testing on Process Mean Levels ..................................................... 56
  4.1 Introduction .................................................................................................. 57
  4.2 Formulation as an Optimal Stopping Problem ............................................... 60
  4.3 Derivation of the Optimal Stopping Rule ..................................................... 63
4.4 Properties of the Optimal Stopping Boundaries ............................................. 71
4.5 Computation of the Optimal Stopping Boundaries ............................................. 76
4.6 Numerical Results ............................................................................................... 78

5  Likelihood Ratio Testing Procedures ................................................................. 80

  5.1 Introduction ........................................................................................................... 81
  5.2 The Structure of LR Testing Procedures ............................................................. 82
  5.3 Extended LR Testing Procedures ......................................................................... 84
  5.4 A Combined CUSUM and Shewhart Control Chart Procedure ......................... 86
  5.5 Performance Analysis ........................................................................................... 87
     5.5.1 The LR Testing Procedure Applied to AR(1) Processes ............................... 88
     5.5.2 The Extended LR Procedure Applied to IMA(1,1) Processes ..................... 92
     5.5.3 The CCS Procedure ..................................................................................... 96
  5.6 Summary of the Performance Analyses .................................................................. 99

6  Conclusion and Future Research ............................................................................ 100

  6.1 Contributions ....................................................................................................... 101
  6.2 Future Research ................................................................................................... 103

Bibliography ............................................................................................................... 104
List of Tables

3.1 Optimum ARLs and parameters of the CUSUM, EWMA and Shewhart control chart procedures applied to the residuals of the machining process with $w = 0.553$ subject to an additive step shift of size $\Delta$ in its state equation when their in-control ARLs are 100. .................................................. 46

3.2 Optimum ARLs and parameters of the CUSUM, EWMA and Shewhart control chart procedures applied to the residuals of the machining process with $w = 0.553$ subject to a drift rate change of $c'$ in its state equation when their in-control ARLs are 300. .................................................. 49

3.3 Optimum ARLs and parameters of the CUSUM, EWMA and Shewhart control chart procedures applied to the residuals of the AR(2) cutting process subject to an observational additive step shift of size $\Delta$ when their in-control ARLs are 100. .................................................. 53

4.1 The optimal stopping boundaries when $\delta(i) = 0.9^i$ for $i \geq 0$. ................... 78
5.1 ARLs of the LR testing procedure and the CUSUM and EWMA control chart procedures applied to the residuals of an AR(1) process with autoregressive coefficient \( \alpha_1 = 0.45 \) subject to an observational additive step shift of size \( \Delta \) when their in-control ARLs are 100. ARLs are optimized at \( \Delta = 1 \) in the upper panel, and at \( \Delta = 3 \) in the lower panel.

5.2 ARLs of the extended LR (XLR) testing procedure in comparison with the classical LR testing procedure and the CUSUM and EWMA control chart procedures applied to the residuals of an IMA(1,1) process with moving average coefficient \( b_1 = 0.1 \) subject to an observational additive step shift of size \( \Delta \) when their in-control ARLs are 100. ARLs are optimized at \( \Delta = 1 \) in the upper panel, and at \( \Delta = 3 \) in the lower panel.

5.3 ARLs of a CCS procedure in comparison with the LR testing procedure and the CUSUM control chart procedure applied to the residuals of an AR(1) process with autoregressive coefficient \( \alpha_1 = 0.45 \) subject to an observational additive step shift of size \( \Delta \) when their in-control ARLs are 100.
List of Figures

3.1 Optimum ARLs versus sizes of change (upper panel) and the ARL curves with the chart parameters optimizing the ARLs at $\Delta = 2$ (lower panel) of the CUSUM, EWMA and Shewhart control chart procedures applied to the residuals of the machining process with $w = 0.553$ subject to an additive step shift of size $\Delta$ in its state equation, when the fixed in-control ARL is 100. . . . 48

3.2 Optimum ARLs versus sizes of change (upper panel) and the ARL curves with the chart parameters optimizing the ARLs at $c' = 2$ (lower panel) of the CUSUM, EWMA and Shewhart control chart procedures applied to the residuals of the machining process with $w = 0.553$ subject to a drift rate change of $c'$ in its state equation, when the fixed in-control ARL is 300. . . . . . . . 51

3.3 Optimum ARLs versus sizes of change (upper panel) and the ARL curves with the chart parameters optimizing the ARLs at $\Delta = 2$ (lower panel) of the CUSUM, EWMA and Shewhart control chart procedures applied to the residuals of the $AR(2)$ cutting process subject to an observational additive step shift of size $\Delta$, when the fixed in-control ARL is 100. . . . . . . . . . 54

4.1 A typical illustration of $g_n(\lambda)$ and the related functions and quantities. . . . 70

4.2 The optimal stopping boundaries of the example. . . . . . . . . . . . . . . . . . 78
5.1 ARLs of the LR testing procedure and the CUSUM and EWMA control chart procedures applied to the residuals of an $AR(1)$ process with autoregressive coefficient $a_1 = 0.45$ subject to an observational additive step shift of size $\Delta$. For the fixed in-control ARL, the chosen chart parameters optimize the ARL at $\Delta = 1$ in the upper panel and at $\Delta = 3$ in the lower panel.

5.2 ARLs of the extended LR (XLR) testing procedure in comparison with the classical LR testing procedure and the CUSUM and EWMA control chart procedures applied to the residuals of an $IMA(1,1)$ process with moving average coefficient $b_1 = 0.1$ subject to an observational additive step shift of size $\Delta$. For the fixed in-control ARL, the chosen chart parameters optimize the ARL at $\Delta = 1$ in the upper panel and at $\Delta = 3$ in the lower panel.

5.3 ARLs of a CCS detection procedure in comparison with the classical LR testing procedure and the CUSUM and EWMA control chart procedures applied to the residuals of an $AR(1)$ process with autoregressive coefficient $a_1 = 0.45$ subject to an observational additive step shift of size $\Delta$. 

91

95

98
Chapter 1

Introduction
1.1 The Problem

The problem of change detection is about quick detection of a change in a dynamic system or process at a low rate of false alarm by sequentially observing the system or process. The instant of change is unknown, and randomness is usually involved in the system. The problem arises naturally in the areas of quality control, signal processing, automatic control and others. Specific examples of applications, such as statistical process control, navigation system monitoring, seismic data processing, signal segmentation, and vibration monitoring of mechanical systems, can be found in Basseville and Nikiforov (1993).

Autocorrelation may exist in the process or the sequential observations of the system. In quality control, more than 70% of the studied processes subject to change detection are autocorrelated, according to Alwan and Roberts (1995). Because of more complexity, more effort is involved in designing change detection procedures for autocorrelated processes than for independent and identically distributed (i.i.d.) processes, and neglect of the presence of autocorrelation may seriously reduce the effectiveness of the designed detection procedures. The impact of autocorrelation on the performance of control charts directly applied to the output of an autocorrelated process has been reported in the literature, for example, by Alwan (1992). Therefore, care must be taken to deal with autocorrelation properly.

In production and other areas, an autocorrelated process may be controlled in order to compensate for expected deviations of the process from its target value. This will reduce the systematic variability of the process and improve the process capability in quality control. As a result, both the quality and productivity levels will be increased. A controlled process should be monitored in order to detect changes to the process. A special-cause change can disturb the process and invalidate the process model that has been used to develop the control
policy. If it is not detected, the control policy will become ineffective. The complementary roles of process control and process monitoring have been discussed by MacGregor (1988) and Box and Kramer (1992). Integration of process control and process monitoring has been proposed by Tucker, Faltin and Vander Wiel (1993) and Montgomery, Keats, Runger and Messina (1994). Usually, for a controlled autocorrelated process, autocorrelation still exists in the closed-loop process output.

Since Wald (1947) founded sequential analysis and Page (1954) proposed the CUSUM procedure, substantial research has been done for detection of step shifts in process mean levels in i.i.d. processes and their continuous analogue – drift rate changes in Wiener processes. Works including Shiryaev (1961), Shiryaev (1963), Shiryaev (1978), Lorden (1971), Lai (1995), and Siegmund and Venkatraman (1995) have made contributions to the related problems. The optimality of the likelihood ratio CUSUM procedure under a well-defined minimax criterion for detection of a step shift in process mean level in an i.i.d. process was first proven by Moustakides (1986) by using optimal stopping theory, and later by Ritov (1990) by using saddle point decision theory. Similar optimality results have also been obtained by Beibel (1996) and Yakir (1997). The situation where an i.i.d. process is subject to a step shift in process mean level has been referred to as the i.i.d. case in the literature, because under this situation, the process is i.i.d. both before and after change. Moustakides (1998) extended the optimality condition, that is, the process is i.i.d. both before and after change, of the likelihood ratio CUSUM procedure to a more general condition, where only the likelihood ratio sequence is required to be i.i.d. both before and after change. However, as stated by Moustakides, the generalized condition is very restrictive, and the extension for autocorrelated processes is very limited.
Since 1970s, change detection in autocorrelated processes has received growing attention. Kligene and Telkis (1983) surveyed and analyzed the works published before 1980. Basseville (1988) surveyed and described the then state-of-the-art statistical methods for detecting changes in signals and systems. Basseville and Nikiforov (1993) presented the existing methodology comprehensively, and included the relevant references. Lai (1995), with discussions, surveyed a large variety of sequential change detection procedures widely scattered in quality control, fault detection and signal processing literature. Since change detection in autocorrelated processes is more complicated than in the i.i.d. case, there are more challenging problems. And less attention has been drawn to change detection in controlled processes.

This thesis studies some problems on change detection in autocorrelated processes. Specifically, this thesis systematically investigates the properties of process residuals and by using residuals, achieve a model transformation of change detection in autocorrelated processes either controlled or not, establish a set of procedures for accurately and precisely computing the ARLs of control charts and studying their performance, study a situation of change detection in autocorrelated processes – optimal sequential testing on process mean levels, analyze and extend a widely used change detection procedure – the LR testing procedure, propose a combined CUSUM and Shewhart scheme for change detection in autocorrelated processes, and study the performance of these change detection procedures.

In the following sections, we will introduce two production processes that have motivated this work, and formulate the problem of change detection in a general framework.
1.2 Two Production Processes

1.2.1 A Machining Process

An application that motivated this work was a machining process introduced in Crowder (1992), which is described as follows. A computer-controlled bar lathe forms cylindrical metal pieces from a metal bar in a turning operation by cutting away excess metal. The process is subject to tool wear and stochastic shocks resulting from small differences in the bars, environmental condition changes, etc.. The outside diameters of metal cylinders are measured by an automatic gauging system which contributes moderate measurement errors.

The process can be represented by a controlled random walk model with linear drift and measurement errors as follows. For \( t = 1, 2, \ldots \),

\[
Y_t = \Theta_t + \epsilon_t, \\
\Theta_t = \Theta_{t-1} + u_{t-1} + c + \nu_t,
\]

where \( \Theta_t \) is the outside diameter value of the part at time \( t \), \( Y_t \) is its measurement, \( \epsilon_t \sim N(0, \sigma^2_\epsilon) \) is the measurement error, \( \nu_t \sim N(0, \sigma^2_\nu) \) is the stochastic shock, \( c \) is the linear drift rate, which may well approximate the effect of tool wear during a time period, and \( u_{t-1} \) is the amount of feedback control determined by \( \{ Y_i, i \leq t - 1 \} \). We assume that \( \Theta_0 \sim N(\theta_0, \sigma^2_0) \) is the initial setting, \( \{ \epsilon_t \} \) and \( \{ \nu_t \} \) are two independent sequences of i.i.d. random variables. Vander Wiel (1996 b) studied such a machining process with target outside diameter value of 15010 units (1 unit = 0.0001 inch), and obtained the following maximum likelihood estimates of the process parameters:

\[
\hat{\sigma}_\epsilon = 3.75, \\
\hat{\sigma}_\nu = 3.10, \\
\hat{c} = 0.84.
\]
The machining process modeled by (1.1) is representative of nonstationary production processes subject to tool wear and measurement errors. Its optimal control has been studied by several authors. Crowder (1992) studied optimal control of the process when no drift was present, Jensen and Vardeman (1993) developed an optimal control policy for the process with additional adjustment errors, and Vander Wiel (1996b) found an optimal control policy under the assumption that the lathe can automatically compensate for the deterministic tool wear with negligible cost. The process model belongs to the class of state-space models, which have been extensively examined in the process control literature such as Astrom and Wittenmark (1997).

1.2.2 A Cutting Process

A second application that further stimulated this work was a metal cutting process studied by Allen and Huang (1994). During rough turning operations of a tool, the vertical cutting force experienced by the tool was measured by two strain gauges mounted on the top and bottom faces of the tool. The force measurements were digitized by an analogue-to-digital converter, and input to a computer numerical controller. The controller then manipulated the feed rate in order to maintain the cutting force at a desired level.

The process was modeled by an autoregressive process of order 2 (AR(2) process) under feedback control as follows.

\[ Y_t = a_1 Y_{t-1} + a_2 Y_{t-2} + b_1 u_{t-1} + b_2 u_{t-2} + \epsilon_t, \]

where \( Y_t \) represents the deviation of the vertical cutting force from its target level at time \( t \), \( u_{t-1} \) is the control input (feed rate) that is determined by \( \{Y_i, i \leq t - 1\} \), and \( \epsilon_t \) is the noise input. It's assumed that \( \{\epsilon_t\} \) is a sequence of i.i.d. normal random variables with mean 0.
and variance $\sigma^2_e$. Allen and Huang (1994) obtained the following least square estimates of the process parameters:

$$
\begin{align*}
    a_1 &= 0.45, \quad a_2 = 0.2, \\
    b_1 &= 0.1, \quad b_2 = 0.04,
\end{align*}
$$

and derived a so-called general minimum variance control policy.

The model (1.3) is a special case of ARIMA processes under feedback control. With their powerful capability of modeling, ARIMA processes have been widely applied in the areas of automatic control, quality control and others, which can be seen in the literature such as Box, Jenkins and Reinsel (1994).

In the turning operations, the cutting conditions such as depth of cutting and cutting speed are usually changeable. These changes may cause the cutting force to largely deviate from its target. Special-cause changes occur to the machining process in § 1.2.1 as well, which may result from abnormal status of the hydraulic system or from changes in the cutting tool conditions. Two approaches can be employed to deal with the changes, which are adaptive control and change detection. By employing change detection, we can detect and eliminate the root cause of change if possible. Or, we can detect the change, and then adjust the control policy when it is necessary. Thus, we can leave the system more stable.
1.3 Problem Formulation of Change Detection

The change detection problem can be formulated in a general statistical inference framework.

Let \{Y_t\} be a stochastic process having a parametric model \( \Psi \) that is subject to a change at an unknown time point \( r \). The probability distribution functions of \{Y_t\} belong to \( \{F_{\Psi(r)}, r \in \Omega\} \), where \( \Omega = \{1, 2, \cdots, \infty\} \). \( r \) is called the change point, and we want to have it detected.

The parameter value of \( \Psi \) may be a vector, and may be a function of time \( t \). \( \Psi(r) \) denotes how the process model parameter changes. If \( r = \infty \), then no change occurs to the process, and \( \Psi = \psi_0 \). If \( r < \infty \), then the change occurs to the process,

\[
\Psi = \begin{cases} 
\psi_0 & \text{for } t < r, \\
\psi_1 & \text{for } t \geq r.
\end{cases}
\]

That is, the process model parameter jumps to \( \psi_1 \) from \( \psi_0 \) at \( t = r \), and then remains \( \psi_1 \).

When \( \Psi = \psi_0 \), the process is said to be in-control, and when \( \Psi = \psi_1 \), the process is said to be out-of-control.

We assume that we start observing the process at time \( t = 1 \). Let \( y_t \) be the observation of \( Y_t \), and \( y^*_t = (y_1, \cdots, y_t) \) be the observations obtained up to \( t \geq 1 \). Let \((\cdots, y_{-1}, y_0)\) denote the initial condition of the process prior to \( t = 1 \), assumed to be given. Let \( y^t = (\cdots, y_{t-1}, y_t) \).

Based on \( y^*_t \) (or equivalently \( y^t \)), detection of the change point \( r \) is statistical inference on \( r \), and the detection or inference problem is the following sequential testing problem. At time \( t = 1, 2, \cdots \), test the hypotheses

\[
H_0: r > t \quad \text{versus} \quad H_1: r \leq t
\]

based on the observations \( y^*_t \) (or equivalently \( y^t \)). We call these two hypotheses the change point hypotheses. Observation is continued and the test is carried out until a decision in favor
of $H_1$ is made for the first time. When $H_1$ is decided at stage $t$, occurrence of change is alarmed, and the observation and test are stopped.

For $t = 1, 2, \ldots$, let $\mathcal{F}_t$ be the minimum $\sigma$-algebra induced by $\{Y_i, i = 1, 2, \ldots, t\}$. Then, the change detection procedure or equivalently, the sequential testing procedure is a stopping rule (or a stopping time) $N \geq 1$ with respect to $\{\mathcal{F}_t, t = 1, 2, \ldots\}$. 
Chapter 2

Residuals for Model Transformation
2.1 Introduction

When autocorrelation became prominent in the 1970s, concern was raised about the performance of control charts directly applied to autocorrelated processes. As observed and studied by Johnson and Bagshaw (1974), Bagshaw and Johnson (1975), Vasilopoulos and Stamboulis (1978), Alwan (1992) and Wardell, Moskowitz and Plante (1992), control charts directly applied to autocorrelated processes are seriously misleading. They either result in too many false alarms or can fail to alarm when a change really occurs.

To circumvent autocorrelation, residuals have been used for change detection in autocorrelated processes, since residuals are mutually independent under certain circumstances. An early work on monitoring residuals of autocorrelated processes was Berthouex, Hunter and Pallesen (1978). Since it was formally proposed by Alwan and Roberts (1988), applying control charts to residuals for change detection in autocorrelated processes has been a popular practice in the quality control area. In some special cases, some properties of residuals have appeared or been studied in the literature. Harris and Ross (1991) studied the properties of residuals of an integrated moving average process of order (1, 1) ($IMA(1,1)$ process) and an autoregressive process of order 1 ($AR(1)$ process) subject to step shifts in the process mean levels. Wardell, Moskowitz and Plante (1992) studied how residuals of an autoregressive moving average process of order $(1,1)$ respond to a step shift in process mean level by assuming the process is deterministic. Wardell, Moskowitz and Plante (1994) derived the forms of residuals of an autoregressive moving average process of order $(p,q)$ process subject to a change in process mean level by using Z-transform. Basseville and Nikiforov (1993) studied some properties of residuals, and utilized the convenience of residuals in designing change detection procedures when the residuals were mutually independent.
Although residuals have been used for change detection in practice and some properties of residuals have been around in the literature, the usage of residuals has been *ad hoc*, properties of residuals have not been systematically investigated and especially, the sufficiency of residuals for change detection has been overlooked. In this chapter, we are going to systematically investigate the properties of residuals for change detection, show the sufficiency of residuals for change detection, develop a general procedure for derivation of forms of residuals of autocorrelated processes, and show the independence of feedback control of residuals. The sufficiency of residuals will provide us a theoretical background for the use of residuals, knowledge of the properties of residuals will enable us to design better change detection procedures and monitor residuals more effectively, and the independence of residuals of feedback control will turn out to be a convenient bridge for the combination of process control and change detection.

For the stochastic process \( \{Y_t\} \) defined in § 1.3, let

\[
g(y^{t-1}) = E_{r=\infty}(Y_t | y^{t-1}),
\]

for \( t = 1, 2, \cdots \). We call \( g(y^{t-1}) \) the *forecast function*. Obviously, \( g(y^{t-1}) \) is independent of the unknown value of \( r \). The *process residual* at \( t = 1, 2, \cdots \) is defined as

\[
e_t = Y_t - g(y^{t-1}).
\]  

\( \{e_1, e_2, \cdots\} \) forms a stochastic process. We will also use \( e_t \) denote its realization. That is,

\[
e_t = y_t - g(y^{t-1}).
\]

Let \( e_1^t = (e_1, \cdots, e_t) \). For convenience, we also let \( e^t \) denote \( e_1^t \).
2.2 SUFFICIENCY OF RESIDUALS FOR CHANGE DETECTION

We first study the sufficiency of process residuals for change detection. We have noted that the initial condition $y^0$ is always given. For $t = 1, 2, \cdots$,

\[
\begin{align*}
e_1 &= y_1 - g(y^0) \\
e_2 &= y_2 - g(y_1, y^0) \\
& \vdots \\
e_t &= y_t - g(y_{t-1}, \cdots, y_1, y^0)
\end{align*}
\]

define a one-to-one transformation $G$ from the observations $y^t_i$ to the residuals $e^t_i$, that is, $e^t_i = G(y^t_i)$. $e^t_i$ is a statistic. For any given $e^t_i$, there is one and only one $y^t_i = G^{-1}(e^t_i)$. That is, the conditional distribution of $y^t_i$ given $e^t_i$ is

\[
P(y^t_i|e^t_i) = \begin{cases} 
1, & \text{if } y^t_i = G^{-1}(e^t_i), \\
0, & \text{otherwise,}
\end{cases}
\]

which is independent of the unknown parameter $r$. Therefore, by the definition of sufficiency, at any time $t \geq 1$, $(e^t_1, \cdots, e^t_t)$ is a sufficient statistic on $\{F_{\theta(r)}, r \in \Omega\}$, or equivalently a sufficient statistic on the unknown parameter $r$. That is, all the information about the unknown parameter $r$ contained in $y^t_i$ is also fully preserved in $e^t_i$. Hence, the residuals are sufficient for the change detection problem.

One can design change detection procedures based on the observations directly. However, due to the one-to-one transformation $G^{-1}$, any change detection procedure based on the observations can be transformed to a change detection procedure based on the residuals, and vice versa. And a detection procedure based on the residuals will have exactly the same detection properties as its counterpart based on the observations.

Because residuals are sufficient, and equivalent to observations for a change detection prob-
lem, when residuals have nicer properties than observations, we can concentrate on residuals in designing and searching for effective change detection procedures. As we will see, residuals can have much nicer properties, such as mutual independence, than observations, which will provide us great convenience in designing and analyzing change detection procedures.

We note that the sufficiency of residuals always holds for the change detection problem, no matter what the process model is, if a change occurs, or how the change occurs. More on the sufficiency of residuals for change detection will be discussed in § 2.3 and § 2.5.2.
2.3 Second-Order Properties of Residuals

We now study the properties of residuals further. Because of the definition of residuals (2.1),

\[ E_{r=\infty}(e_t|y^{t-1}) = E_{r=\infty}(Y_t - g(y^{t-1})|y^{t-1}) = E_{r=\infty}(Y_t|y^{t-1}) - g(y^{t-1}) = 0. \]

Then, by using total probability,

\[ E_{r=\infty}(e_t) = 0, \]

for any \( t \geq 1 \). By a similar approach, for any \( i \geq 1 \), since

\[ E_{r=\infty}(e_t Y_{t-i}|y^{t-i}) = E_{r=\infty}(e_t|y^{t-i}) \cdot y_{t-i} = 0, \]

we have

\[ E_{r=\infty}(e_t Y_{t-i}) = 0, \]

and furthermore

\[ E_{r=\infty}(e_t e_{t-i}) = 0, \]

for any \( t \geq 1 \). Therefore, when there is no change, the residuals are mutually uncorrelated, and their means are zero. For a Gaussian process, the mutual uncorrelatedness implies that the residuals are mutually independent when there is no change.

One may now be interested in looking at the distributions of residuals. We consider this in the case of continuous variables. Let \( f_{\psi(r)} \) be the probability density function corresponding to \( F_{\psi(r)} \). That is, \( Y_t^r \) has a joint density function \( f_{\psi(r)}(y_t^r) \). The Jacobian of \( G \),

\[ |G| = 1. \]
Therefore, the joint density function of \( e_i^t = G(y_i^t) \) is

\[
f_{\psi(t)}^s(e_i^t) = f_{\psi(t)}(G^{-1}(e_i^t)),
\]

and

\[
f_{\psi(t)}^s(e_i^t) = f_{\psi(t)}(y_i^t).
\]

That is, the likelihood of the residuals \( e_i^t \) and the likelihood of the corresponding observations \( y_i^t \) are equal. This from another side confirms the sufficiency of residuals for change detection.

If the residuals are mutually independent, then

\[
f_{\psi(t)}(y_i^t | y^0)
= \prod_{i=1}^{t} f_{\psi(t)}(y_i | y^{i-1})
= \prod_{i=1}^{t} f_{\psi(t)}^s(e_i),
\]

and therefore, a LR testing procedure based on residuals will have much simpler form than its counterpart based on observations.
2.4 Procedure for Deriving Forms of Residuals

In order to analyze the properties of the residuals \( \{e_t\} \) in detail, and especially to learn how they behave when a change \( \Psi(r) \) occurs to the process \( \{Y_t\} \), we need to derive the forms of \( \{e_t\} \). Two things must be kept in mind when we are deriving the forms. The first thing is that the forecast function \( g(y^{t-1}) \) used to compute the residuals is always defined under the assumption that \( \{Y_t\} \) is in control, that is, \( r = \infty \), because we don't know the occurrence of the change until it's detected, and before the detection, we make forecasts by assuming that there is no change. The second thing is that \( \{Y_t\} \) is subject to the change \( \Psi(r) \).

The forms of residuals may have to be specifically derived for a particular type of process \( \{Y_t\} \). However, there is a general procedure to follow, which is the following.

(i) Derive the forecast function \( g(y^{t-1}) \), that is, the one step ahead forecast by assuming \( r = \infty \).

(ii) Find the conditional distribution of the process output \( Y_t \) for given \( y^{t-1} \) by taking into account the effect of \( \Psi(r) \).

(iii) Find the relationship between the conditional expectation of \( Y_t \) in (ii) and \( g(y^{t-1}) \) in (i).

(iv) Subtract \( g(y^{t-1}) \) from \( Y_t \) having the conditional distribution in (ii) by using the relationship of (iii). Then, the forms of residuals \( e_t = Y_t - g(y^{t-1}) \) are ready.
2.5 Properties of Residuals of ARIMA Processes

Now, we can derive the forms of residuals of an ARIMA process.

Before going on, we need to distinguish between two kinds of changes in an autocorrelated process, namely, additive changes and nonadditive changes. An additive change occurs in the process mean level, and has an additive effect on the mean values of process observations. Nonadditive changes are model parameter changes, including changes in spectral characteristic or system dynamics.

2.5.1 Processes Subject to Additive Changes

An ARIMA process of order \((p, l, q)\) (\(ARIMA(p, l, q)\) process) under feedback control subject to an additive change can be represented as

\[
Y_t = u_{t-1} + \frac{\beta(B)}{\alpha(B)(1 - B)^l} \varepsilon_t + \Delta(t - r),
\]

where by using time series notations as those in Box, Jenkins and Reinsel (1994), \(B\) is the backward shift operator applied to the time index \(t\), \(\alpha(B) = 1 + a_1 B + \cdots + a_p B^p\) is the stationary autoregressive operator, \(\beta(B) = 1 + b_1 B + \cdots + b_q B^q\) is the invertible moving average operator, \(l\) is a nonnegative integer. The process is stationary when \(l = 0\), and nonstationary when \(l \geq 1\). \(\{\varepsilon_t\}\) is the innovation sequence that is a sequence of white noise input with mean zero and variance \(\sigma^2\). \(u_{t-1}\) is the net effect of the feedback control that is determined by \(y^{t-1}\), and \(u_{t-1} \equiv 0\) represents that there is no control applied. \(r\) is the unknown change point. \(\Delta(\cdot)\) is a deterministic function, \(\Delta(i) \equiv 0\) for \(i < 0\) and \(\Delta(i)\) may be nonzero for \(i \geq 0\). (2.2) can also be equivalently represented as

\[
Y_t = \frac{\beta(B)}{\alpha(B)(1 - B)^l} [\varepsilon_t + \delta(t - r)],
\]
where
\[ \delta(t - r) = \frac{\alpha(B)(1 - B)^l}{\beta(B)} \Delta(t - r). \]

We call \( \Delta(t - r) \) an *observational* additive change, and \( \delta(t - r) \) an *innovational* additive change.

In real world, these two types of additive changes may come from different backgrounds. However, they are mathematically equivalent.

For the ARIMA\((p, l, q)\) process subject to the additive change \( \Delta(t - r) \), we can derive the forms of residuals by following the general procedure in § 2.4.

(i) 
\[ g(y_t) = [1 - \frac{\alpha(B)(1 - B)^l}{\beta(B)}]y_t + \frac{\alpha(B)(1 - B)^l}{\beta(B)}u_{t-1}. \]

(ii) Given \( y_t \),
\[ Y_t = [1 - \frac{\alpha(B)(1 - B)^l}{\beta(B)}]y_t + \frac{\alpha(B)(1 - B)^l}{\beta(B)}u_{t-1} + \frac{\alpha(B)(1 - B)^l}{\beta(B)}\Delta(t - r) + \epsilon_t. \]

(iii) The relationship between \( g(y_t) \) and \( E(Y_t | y_t) \) is obvious.

(iv)
\[ e_t = \begin{cases} 
\epsilon_t, & \text{for } t < r, \\
\frac{\alpha(B)(1 - B)^l}{\beta(B)}\Delta(t - r) + \epsilon_t, & \text{for } t \geq r.
\end{cases} \tag{2.3} \]

Two results can be observed for the ARIMA process subject to the additive change. The first is that the residuals \( \{e_t\} \) are the *inverse filtering* of the observations \( \{Y_t\} \) by removing the term \( u_t \),
\[ e_t = \frac{\alpha(B)(1 - B)^l}{\beta(B)}(Y_t - u_{t-1}). \]

The second is that the residuals form an i.i.d. process subject to an additive change,
\[ e_t = \epsilon_t + \delta(t - r), \]
where $\delta(t - r)$ is the inverse filtering of $\Delta(t - r)$ and $\delta(i) = 0$ for $i < 0$. Therefore, detection of an additive change in an ARIMA process is equivalent to detection of an additive change in an i.i.d. process.

### 2.5.2 Processes Subject to Nonadditive Changes

For an ARIMA process, a model parameter change from $(\alpha, \beta, l, \sigma_i)$ to $(\alpha', \beta', l', \sigma'_i)$ at change point $r$, with the process mean value unchanged, is a nonadditive change to the ARIMA process. Under such a change, we have the following development for the residuals.

(i) 

$$g(y^{i-1}) = [1 - \frac{\alpha(B)(1-B)^i}{\beta(B)}]y_t + \frac{\alpha(B)(1-B)^i}{\beta(B)}u_{t-1}.$$  

(ii) Given $y^{i-1}$, 

$$Y_t = \begin{cases} 
[1 - \frac{\alpha(B)(1-B)^i}{\beta(B)}]y_t + \frac{\alpha(B)(1-B)^i}{\beta(B)}u_{t-1} + \epsilon_t, & \text{for } t < r, \\
[1 - \frac{\alpha'(B)(1-B')^{i'}}{\beta'(B')}y_t + \frac{\alpha'(B)(1-B')^{i'}}{\beta'(B')}u_{t-1} + \epsilon'_t, & \text{for } t \geq r,
\end{cases}$$

where $\epsilon'_t = \epsilon_t$ for $t < r$, and $\{\epsilon'_t, t \geq r\}$ is a sequence of white noise input with mean zero and standard deviation $\sigma'_i$.

(iii) The relationship between $g(y^{i-1})$ and $E(Y_t|y^{i-1})$ is obvious as well.

(iv) 

$$\epsilon_t = \begin{cases} 
\epsilon_t, & \text{for } t < r, \\
\left(\frac{\alpha(B)(1-B)^i}{\beta(B)} - \frac{\alpha'(B)(1-B')^{i'}}{\beta'(B')}\right)(y_t - u_{t-1}) + \epsilon'_t, & \text{for } t \geq r.
\end{cases}$$

Therefore, before the change, the residuals are an i.i.d. process. After the change, the residuals either exhibit autocorrelation or have a changed variance. The possible autocorrelation among residuals will not provide us the convenience for constructing change detection procedures as
the mutual independence does in the situation of additive changes. However, being autocorrelated or not does not affect the sufficiency, and equivalence to observations, of residuals, because the sufficiency holds for change detection problems in general.

Detecting an additive change in a normal i.i.d. process was defined as a basic problem in Basseville and Nikiforov (1993). They established the sufficiency of likelihood ratios for change detection. For change detection in ARIMA processes subject to innovational additive changes, they recognized the sufficiency of residuals because the likelihood ratios were functions of the residuals. However, they did not recognize the prevalent sufficiency of residuals for change detection in general. For detection of a nonadditive change in an ARIMA process, they concluded wrongly that residuals were no longer sufficient for change detection just because they could no longer be mutually independent.

One probably could be puzzled by the sufficiency of residuals for detection of an additive change in a nonstationary ARIMA process, for example, an observational additive step change in an $IMA(1,1)$ process. Because of the differencing operation $(1 - B)^l$, $l \geq 1$ in (2.3) for nonstationary ARIMA processes, the residual mean will go to zero exponentially fast after an observational additive step change has occurred to the ARIMA process. This means that the effect of the observational step change on the residuals will vanish. So, how could the residuals be sufficient? The puzzle will not exist if one can realize that for a nonstationary ARIMA process, the process observations wander around anywhere because of the integrating operation $(1 - B)^{-l}$ in (2.2). Therefore, the effect of the observational step change on the process observations themselves will be buried into the wandering, that is, will vanish as well. The puzzle can also be resolved by the fact that any change detection procedure based on observations (or feedback control actions) can be transformed to a procedure based on
residuals with the same detection properties preserved, and vice versa.
2.6 Properties of Residuals of State Space Models

2.6.1 Residuals of a General Model

Consider a general state space model subject to an additive change $\Delta(t-r)$ in its measurement equation, which is described as follows. For $t = 1, 2, \cdots$,

\[
Y_t = H_t \Theta_t + \epsilon_t + \Delta(t-r),
\]

\[
\Theta_t = \Phi_t \Theta_{t-1} + u_{t-1} + \nu_t,
\]

where at time $t$, $\Theta_t$ is the state of the dynamic system, $\Phi_t$ is the transition matrix, $H_t$ is the state observation transformation matrix, $Y_t$ is the measurement of the transformed $\Theta_t$, $\epsilon_t$ is the measurement error, $\nu_t$ is the noise input, $u_{t-1}$ is the amount of feedback control determined by $y^{t-1}$. $\{\epsilon_t\}$ and $\{\nu_t\}$ are independent normal white noise processes, $\epsilon_t \sim \mathcal{N}(0, \Sigma_{\epsilon})$ and $\nu_t \sim \mathcal{N}(0, \Sigma_{\nu})$. $r \geq 1$ is the change point, $\Delta(\cdot)$ is a deterministic function and $\Delta(i) \equiv 0$ for $i < 0$. $\Theta_0 \sim \mathcal{N}(\theta_0, \Sigma_0)$ is initialized. Although it is not necessary to explicitly specify their dimensions, $\Theta_t$, $Y_t$, $\epsilon_t$, $\nu_t$, $u_{t-1}$ and $\Delta(\cdot)$ all can be vectors.

Following the general procedure in §2.4 and using Kalman filtering, we can derive the forms of residuals of the model (2.4).

(i)

\[
g(y^{t-1}) = H_t[H_t E_{r=\infty}(\Theta_{t-1}|y^{t-1}) + u_{t-1}],
\]

\[
E_{r=\infty}(\Theta_t|y^t) = E_{r=\infty}(\Theta_t|y^{t-1}) + K_t[y_t - H_t E_{r=\infty}(\Theta_t|y^{t-1})]
\]

\[
= K_t y_t + (I - K_t H_t)E_{r=\infty}(\Theta_{t-1}|y^{t-1}) + u_{t-1},
\]

where

\[
K_t = \text{Cov}(\Theta_t|y^{t-1})H_t'[H_t \text{Cov}(\Theta_t|y^{t-1})H_t' + \Sigma_{\epsilon}]^{-1},
\]

\[
\text{Cov}(\Theta_t|y^{t-1}) = \Phi_t \text{Cov}(\Theta_{t-1}|y^{t-1}) \Phi_t' + \Sigma_{\nu},
\]
and

\[ Cov(\Theta_t|y') = [I - K_t H_t] Cov(\Theta_t|y^{t-1}). \]

(ii)

\[ Y_t[y^{t-1}] \sim N(H_t[\Phi_t E(\Theta_{t-1}|y^{t-1}) + u_{t-1}] + \Delta(t - r), \]

\[ H_t[\Phi_t Cov(\Theta_{t-1}|y^{t-1}) \Phi'_t + \Sigma_s] H'_t + \Sigma_s). \]

(iii)

\[ E(\Theta_t|y^t) = K_t [y_t - \Delta(t - r)] + (I - K_t H_t)[\Phi_t E(\Theta_{t-1}|y^{t-1}) + u_{t-1}]. \]

For \( t < r \),

\[ E(\Theta_t|y^t) = E_{r=\infty}(\Theta_t|y^t). \]

By mathematical induction, for \( t \geq r \),

\[ E(\Theta_t|y^t) = E_{r=\infty}(\Theta_t|y^t) - \sum_{i=0}^{t-r-1} \prod_{j=0}^{i-1} [(I - K_{t-i} H_{t-j}) \Phi_{t-i-j}] K_{t-i} \Delta(t - r - i), \]

where \( \prod_{j=0}^{i-1} \equiv 1. \)

(iv) For \( t < r \),

\[ e_t \sim N(0, H_t[\Phi_t Cov(\Theta_{t-1}|y^{t-1}) \Phi'_t + \Sigma_s] H'_t + \Sigma_s). \]

For \( t \geq r \),

\[ e_t \sim N(\Delta(t - r) - H_t \Phi_t \sum_{i=0}^{t-r} \prod_{j=0}^{i-1} [(I - K_{t-i-j} H_{t-i-j}) \Phi_{t-i-j}], \]

\[ K_{t-i} \Delta(t - 1 - r - i), H_t[\Phi_t Cov(\Theta_{t-1}|y^{t-1}) \Phi'_t + \Sigma_s] H'_t + \Sigma_s), \]

where \( \sum_{i=0}^{t-1} \equiv 0. \)
By an approach similar to that in § 2.3, we can prove that \( \{e_t, t = 1, 2, \ldots\} \) are mutually uncorrelated. Therefore, \( \{e_t, t = 1, 2, \ldots\} \) are mutually independent because of the Gaussian property.

Similarly, we can derive the forms of residuals of a state space model subject to an additive change in its state equation. We will illustrate this with the machining process in the next section.

2.6.2 Residuals of the Machining Process

We have noted that the model of the machining process is a special case of state space models. We now consider the machining process subject to an additive change \( \Delta(t - r) \) in the state equation, which is represented as follows. For \( t = 1, 2, \ldots \),

\[
Y_t = \Theta_t + e_t,
\]

\[
[\Theta_t - \Delta(t - r)] = [\Theta_{t-1} - \Delta(t - 1 - r)] + u_{t-1} + c + \nu_t,
\]

or equivalently,

\[
Y_t = \Theta_t + e_t,
\]

\[
\Theta_t = \Theta_{t-1} + u_{t-1} + c + \nu_t + [\Delta(t - r) - \Delta(t - 1 - r)],
\]

where \( r \geq 1 \) is a change point, \( \Delta(\cdot) \) is a deterministic function and \( \Delta(i) \equiv 0 \) for \( i < 0 \).

The change \( \Delta(t - r) \) represents an additive change to \( \Theta_t \) in (2.5). Looking at (2.6), we see the differencing of \( \Delta(t - r) \) with respect to \( t \), \( \Delta(t - r) - \Delta(t - 1 - r) \), can also represent additive changes in the drift rate \( c \) and in the mean level of the stochastic shocks \( \{\nu_t\} \). Therefore, we use \( \Delta(t - r) \) to represent the total effect of additive changes to \( \Theta_t \), which may result from changes in tool conditions, metal bar quality and environmental conditions and others.
The function $\Delta(\cdot)$ can represent various types of additive changes.

$$\Delta(i) \equiv \Delta \text{ for } i \geq 0$$

represents a step shift.

$$\Delta(i) = (i + 1)c' \text{ for } i \geq 0$$

represent a drift rate change from $c$ to $c + c'$ per part or an external ramping shift.

$$\Delta(0) = \iota$$

and

$$\Delta(i) = 0 \text{ for } i \neq 0$$

represents a impulse shift.

We have mentioned that $\Theta_0 \sim N(\theta_0, \sigma_\Theta^2)$ is initialized. For $t = 1, 2, \ldots$, let

$$q_t^2 = \text{Var}(\Theta_t | y^t).$$

By Kalman filtering,

$$q_t^2 = \frac{q_{t-1}^2 + \sigma_\nu^2}{q_{t-1}^2 + \sigma_\nu^2 + \sigma_\varepsilon^2}. \quad (2.7)$$

Let

$$s_t^2 = q_{t-1}^2 + \sigma_\nu^2 + \sigma_\varepsilon^2, \quad (2.8)$$

and

$$w_t = \frac{q_{t-1}^2 + \sigma_\nu^2}{q_{t-1}^2 + \sigma_\nu^2 + \sigma_\varepsilon^2}. \quad (2.9)$$

Then, following the general procedure in § 2.4 and using Kalman filtering, we can derive the forms of residuals of the process represented by (2.6).
(i)

\[ g(y^{t-1}) = E_{r=\infty}(\Theta_{t-1}|y^{t-1}) + u_{t-1} + c, \]

\[ E_{r=\infty}(\Theta_t|y^t) = w_t y_t + (1 - w_t)[E_{r=\infty}(\Theta_{t-1}|y^{t-1}) + u_{t-1} + c]. \]

(ii)

\[ Y_t|y^{t-1} \sim N(E(\Theta_{t-1}|y^{t-1}) + u_{t-1} + c + \Delta(t - r) - \Delta(t - r - 1), s_t^2) \]

(iii)

\[ E(\Theta_t|y^t) = w_t y_t + (1 - w_t)[E(\Theta_{t-1}|y^{t-1}) + u_{t-1} + c] + (1 - w_t)\Delta(t - r) - \Delta(t - r - 1)]. \]

For \( t < r, \)

\[ E(\Theta_t|y^t) = E_{r=\infty}(\Theta_t|y^t). \]

By mathematical induction, for \( t \geq r, \)

\[ E(\Theta_t|y^t) = E_{r=\infty}(\Theta_t|y^t) + \sum_{i=0}^{t-r} \Delta(t - i - r) - \Delta(t - i - 1 - r) \prod_{j=0}^{i} (1 - w_{t-j}). \]

(iv) For \( t < r, \)

\[ e_t \sim N(0, s_t^2). \]

For \( t \geq r, \)

\[ e_t \sim N(\sum_{i=0}^{t-r} \Delta(t - i - r) - \Delta(t - i - 1 - r)) \prod_{j=1}^{t-r} (1 - w_{t-j}), s_t^2), \]

where \( \prod_{j=1}^{t-r} \equiv 1. \)

Because of the one-to-one transformation \( G \) from \( y_t^t \) to \( e_t^t, \) by an approach similar to that in § 2.3, we can prove that \( \{e_t, t = 1, 2, \cdots\} \) is a sequence of mutually uncorrelated random variables. By the Gaussian property, \( \{e_t, t = 1, 2, \cdots\} \) is a sequence of mutually independent normal random variables.
Therefore, for the machining process, its residuals form a sequence of mutually independent normal random variables with variances independence of the change. The mean values of residuals are constant zero before the change point \( r \) and start to change after \( r \).

Since \( s_i^2 \) is independent of the change \( \Delta(t - r) \), the residuals \( \{e_t\} \) depend on \( \Delta(t - r) \) only through their mean values

\[
E(e_t) = \sum_{i=0}^{t-r} [\Delta(t - i - r) - \Delta(t - i - 1 - r)] \prod_{j=1}^{i} (1 - w_{t-j}).
\]  

(2.10)

We now examine how \( E(e_t) \) responds to some specific types of additive changes. For a step shift of size \( \Delta \), \( \Delta(i) \equiv \Delta \) for \( i \geq 0 \),

\[
E(e_t) = \begin{cases} 
0, & \text{for } t < r, \\
\Delta \prod_{j=1}^{t-r} (1 - w_{t-j}), & \text{for } t \geq r.
\end{cases}
\]

For a drift rate change, \( \Delta(i) = (i+1)c' \) for \( i \geq 0 \),

\[
E(e_t) = \begin{cases} 
0, & \text{for } t < r, \\
c' \sum_{i=0}^{t-r} \prod_{j=1}^{i} (1 - w_{t-j}), & \text{for } t \geq r.
\end{cases}
\]

2.6.3 Residuals of the Steady-State Machining Process

One can see from (2.9) and (2.8) that the filtering rate \( w_t \) and the residual variance \( s_i^2 \) depend on \( t \) only through \( q_i^2 \) defined by (2.7), which in turn is fully determined by \( q_0^2 \), the variance of the initial process setting \( \Theta_0 \). For any initial value \( q_0^2 \), \( \{q_t^2\} \) converges to

\[
q^2 = (\sqrt{\sigma_e^4 + 4\sigma_e^2 \sigma_e^2 - \sigma_e^2})/2,
\]

which is called the steady state variance. When \( q_t^2 \) is stabilized at \( q^2 \), we say that the process is in steady state.

It is interesting to investigate the effect of \( q_0^2 \) on the convergence of \( \{q_t^2\} \). It is because of (2.7) that \( \{q_t^2\} \) converges very fast to \( q^2 \) for any value of \( q_0^2 \). Numerical results have shown
the convergence of \( \{q_t^2\} \) for the machining process with the parameters given by (1.2) in two extreme cases: (i) \( q_0^2 = 0 \), that is, the initial setting of the process mean is error-free and (ii) \( q_0^2 \) is large, say 10^2. In both cases, \( \{q_t^2\} \) stabilizes at \( q^2 \) after two iterations. Furthermore, if \( q_0^2 = q^2 \), then \( q_t^2 = q^2 \) for all \( t \). Therefore, we assume that the process is in steady state by assuming \( q_0^2 = q^2 \).

When the process is in steady state,

\[
s_t^2 = s^2 = q^2 + \sigma_\nu^2 + \sigma_\xi^2,
\]

and the filtering rate

\[
w_t \equiv w = (\sqrt{r_{SN}^2 + 4r_{SN} + r_{SN}})/(\sqrt{r_{SN}^2 + 4r_{SN} + r_{SN} + 2}),
\]

where \( r_{SN} = \sigma_\nu^2/\sigma_\xi^2 \) is the signal-to-noise ratio. It is easy to see that \( w(\cdot) \) is an increasing function of \( r_{SN} \). That is, the higher the signal-to-noise ratio, the larger the filtering rate.

From (2.10), for \( t \geq r \),

\[
E(e_t) = \sum_{i=0}^{t-r} (1 - w)^i [\Delta(t - i - r) - \Delta(t - i - 1 - r)].
\]

When the change is a step shift of size \( \Delta \),

\[
E(e_t) = (1 - w)^{t-r} \Delta.
\]

That is, the residual mean jumps to \( \Delta \) at \( r \) and then decreases exponentially back to zero.

When the change is a drift rate change from \( c \) to \( (c + c') \) per part,

\[
E(e_t) = [1 - (1 - w)^{t-r+1}] \frac{c'}{w}.
\]

That is, the residual mean will converge exponentially to a non-zero level of \( c'/w \).
For the machining process with parameters in (1.2),

\[ q^2 = 2.788^2, \quad s^2 = 5.608^2 \quad \text{and} \quad w = 0.5528. \]

In summary, for the machining process in steady state, its residuals form a sequence of mutually independent normal random variables with a constant variance \( s^2 \). Let

\[
\delta(k) = \begin{cases} 
0, & \text{for } k < 0, \\
\sum_{i=0}^{k} (1 - w)^i [\Delta(k - i) - \Delta(k - i - 1)], & \text{for } k \geq 0.
\end{cases}
\]

Then the residuals form an i.i.d. normal process subject to an additive change \( \delta(t - r) \).
2.7 Independence of Residuals of Feedback Control

An objective of quality control is to reduce process variability and improve process capability. For autocorrelated processes, by applying automatic process control, we can reduce systematic variation. By applying change detection, we can possibly detect and remove special causes in order to improve the processes. As we have mentioned in § 1.1, automatic process control and change detection (process monitoring) are complementary rather than competitive tools, and their combination for better quality control has been advocated in the literature.

Examining the results and developments so far, we see that the forms of residuals of the ARIMA processes and state space models subject to additive changes are independent of the applied feedback control. This may at first appear surprising. However, the reason behind this is simple. That is, $u_{t-1}$ is fully determined by $y_{t-1}$, and the conditioning operation in computation of the residuals cancels out the effect of $u_{t-1}$ on $e_t$. Actually, the form of residuals of a process is independent of any deterministic component of the process, for example, the deterministic drift rate $c$ in the machining process.

However, we must note that the forms of residuals of a process may depend on the applied feedback control when the process is subject to a nonadditive change. For example, when the autoregressive or moving average operator of an ARIMA process changes, feedback control will impose some effect on the means of residuals.

The independence of residuals of feedback control under situations of additive changes serves as a bridge for the combination of automatic process control and change detection in quality control. Under different optimization criteria and control situations, different feedback control policies can be obtained. When adjustment cost is minor, simple minimum variance control can be applied. When adjustment is disruptive or time consuming, adjustment cost
has to be taken into account. Under some situations, such as that described in Vander Wiel (1996 b), tool wear can be automatically corrected by the lathe with negligible cost while in others it cannot. However, no matter what specific form of feedback control is used, residuals behave the same in response to additive changes. Therefore, an independent change detection procedure based on residuals can be developed and used regardless of the specific form of feedback control applied.

One may wonder why the feedback control cannot remove the effect of additive changes in the residuals. The reason is that the effect of all feedback control is incorporated into the forecast and the forecast \( g(y^t-1) \) is done under the assumption that the process is in-control. Consider an example. Suppose a step shift of size \( \Delta \) occurs at \( r \) and an additional amount of \((-\Delta)\) of control is applied to the process at \( (r-1) \). Then the effect of the step shift on process output can be compensated because \( \Theta_r = \Theta_{r-1} + (u_{r-1} - \Delta) + c + \nu_r + \Delta = \Theta_{r-1} + u_{r-1} + c + \nu_r \). Hence, \( Y_r | y^{r-1} \sim N(E(\Theta_{r-1} | y^{r-1}) + u_{r-1} + c, s_r^2) \). However, at \( (r-1) \), \( \Theta_r \) and \( Y_r \) is forecast as \( E(\Theta_{r-1} | y^{r-1}) + (u_{r-1} - \Delta) + c \). Therefore, \( e_r = Y_r - [E(\Theta_{r-1} | y^{r-1}) + (u_{r-1} - \Delta) + c] \sim N(\Delta, s_r^2) \). That is, the feedback can not remove the effect of the change in the residuals. This from another side confirms that the properties of residuals of a process subject to an additive change are independent of the applied feedback control.
2.8 Summary and the Model Transformation

Residuals are always sufficient and equivalent to process observations for change detection. Any change detection procedure based on observations can be transformed to a change detection procedure based on residuals with the same detection properties such as optimality preserved, and vice versa. When there is no change, the residuals are mutually uncorrelated with zero means, and furthermore mutually independent if the process is Gaussian. For an ARIMA process or a steady-state space model subject to an additive change, the residuals are mutually independent. Actually, the residual process is an i.i.d. process subject to an additive change that is the inverse filtering of the original additive change.

Because of the properties of residuals, detection of an additive change in an ARIMA process or a steady-state state space model, possibly under feedback control, is transformed to detection of an additive change in an i.i.d. process with detection properties preserved. This model transformation will offer us much convenience for designing and analyzing change detection procedures based on residuals.

In the rest of this work, we will focus on the residual process

\[ \{e_t, t = 1, 2, \cdots\}, \]

which is an i.i.d. process with zero means and variance \( \sigma^2 \) subject to an additive change \( \delta(t - r) \), where \( r = 1, 2, \cdots, \infty \) is the change point. \( E(e_t) = \delta(t - r) \). \( \delta(i) \equiv 0 \) for \( i < 0 \), \( \delta(i) \) for \( i \geq 0 \) is the mean of residual after change. The formulation of the change detection problem is now as follows. At \( t = 1, 2, \cdots \), test the hypotheses

\[ H_0: r > t \quad \text{versus} \quad H_1: r \leq t \]
based on $e_t$. The corresponding stopping time $N$ is now with respect to $\{\mathcal{F}_t, t = 1, 2, \cdots\}$, where $\mathcal{F}_i$ is the minimum $\sigma$-algebra induced by $\{e_i, i = 1, 2, \cdots, t\}$. 
Chapter 3

Performance of Control Charts

Applied to Residuals
3.1 Introduction

From now on, we will study change detection procedures. In this chapter, we study the CUSUM, EWMA and Shewhart control chart procedures applied to the residual process \( \{e_t, t = 1, 2, \ldots \} \) defined in § 2.8.

Although applying control charts to residuals has been a popular practice for change detection, especially, in the quality control area, a systematic investigation on the performance of the control chart procedures applied to residuals of autocorrelated processes has not been available in the literature. The lack of research in this area was due to two reasons.

- Reason one, lack of a systematic investigation on the properties of residuals in the previous literature.
- Reason two, the time-varying feature of the mean of residual after change, which made the computation of performance measures not so easy as in the i.i.d. case.

By using a discrete-state Markov chain approach, Vander Wiel (1996 a) studied the performance of two-sided CUSUM, EWMA and Shewhart control charts applied to the residuals in a special case, that is, for an \( IMA(1,1) \) process subject to an observational additive step shift.

In this chapter, for a general pattern of the mean of residual after change, that is, a general \( \delta(\cdot) \), we will derive an explicit formula for computation of the ARLs of the Shewhart control chart procedure, establish integral equations for the CUSUM and EWMA control chart procedures, and develop numerical procedures of Gaussian quadrature for solving the integral equations. Because of the time-varying feature of the mean of residual after change, the integral equations and the numerical procedures based on them have not been done by
other authors. Under the ARL criterion, we will then study the performance of the control chart procedures applied to the residuals under three typical situations, where the mean of residual after change is decreasing to zero, decreasing at first and then stabilized at a nonzero level, and gradually increasing and stabilized at a nonzero level. By the ARL criterion, we mean that for a change detection procedure $N$, for a fixed value of its in-control ARL that is $E_{r=\infty}(N)$, the smaller its out-of-control ARL that is $E_{r=1}(N)$, the faster the procedure $N$ for detection of the change $\delta(\cdot)$.

For computation of the ARLs, both the integral equation approach and the Markov chain approach can be used. We are using the integral equation approach here, because it is more efficient and can offer us more accurate and precise solution than the Markov chain approach. Before we examine the advantage of the integral equation approach, we look at the difference between the two approaches, which exists in the following aspects.

- First, in the case of a continuous state variable, by using the integral equation approach, we establish an integral equation of ARL function on the base of the continuous state variable directly. By using the Markov chain approach, however, one has to discretize the continuous state variable and then establish transition probability matrices on the base of the discretization intervals.

- Second, when solving the ARL values, by using the Markov chain approach, one approximates the ARL values at each discretization interval by a constant (zero-degree polynomial). That is, a continuous ARL function is approximated by a discontinuous
step function. By using the integral equation approach, however, we fit the integrand (the continuous ARL function) on each quadrature interval by a polynomial of higher degree when solving the integral equation by using quadrature methods. That is, the continuous ARL function is fitted by a smooth spline curve.

• Third, by using Gaussian quadrature to solve an integral equation, we can choose the location of the quadrature abscissas in an optimal way to achieve the closest fitness of the ARL function by a spline curve. However, there is no such an optimal way for discretization in the Markov chain approach.

For an $N$-point quadrature rule, basically we solve an $N$-dimensional system of linear equations in the integral equation approach. In an $N$-discretization interval Markov chain approach, one needs to carry out $N$-dimensional matrix operations. Therefore, when the number of quadrature points is about the same as the number of discretization intervals, the computational complexities of the two approaches are about the same, the integral equation approach, however, can offer much more accurate and precise solutions than the Markov chain approach. Usually, only a small number of quadrature points are needed to achieve desired computational accuracy and precision. However, the same degree of accuracy and precision may not be achieved by a large number of discretization intervals in the Markov chain approach. Merely increasing the number of discretization intervals cannot increase the accuracy and precision of final solutions, because in the Markov chain approach, (i) the continuous ARL function is always approximated by a discontinuous step function, which is a vital shortcoming; (ii) increasing the number of discretization intervals will increase the computational complexity and intermediate computational errors dramatically. More discussions on quadrature methods and solution of integral equations can be found in Press, Teukolsky, Vetterling and Flannery.
When both the integral equation approach and the Markov chain approach can be used, the former is superior to the later in terms of computational efficiency, accuracy and precision. However, in some cases (for example, when the state variable is more than two dimensional), it may be hard to solve the integral equations efficiently, while the discrete-state Markov chain approximation can be still applied. The less efficiency and accuracy makes the Markov chain approach more flexible and versatile. Lucas and Crosier (1982) computed the ARLs of CUSUM charts in the i.i.d. case by using the two approaches. They discussed and their numerical results demonstrated the relative advantages and disadvantages of the two approaches.
3.2 Integral Equations for Computation of the ARLs

Under the ARL criterion, for a change detection procedure $N$, since we are only interested in $E_{r=\infty}(N)$ and $E_{r=1}(N)$, we can simply assume that

- $r = 1$,

and concentrate on computation of $E_{r=1}(N)$. By substituting constant 0 for $\delta(t - 1)$, $t = 1, 2, \cdots$, $E_{r=\infty}(N)$ is just a special case of $E_{r=1}(N)$.

Two other assumptions made in this chapter are

- $e_t$ has a normal distribution;

- $\delta(t - 1)$ converges as $t$ increases, and the limiting value is denoted by $\delta(\infty)$.

Let $f_{N(\mu, \sigma^2)}$ be the normal density function with mean $\mu$ and variance $\sigma^2$, and $\Phi(\cdot)$ be the standard normal distribution function.

3.2.1 Shewhart Control Chart Procedure

The Shewhart control chart procedure with control limit $C$ applied to $\{e_1, e_2, \cdots\}$ is a stopping time,

$$N_S = \inf\{t \geq 1 : |e_t| \geq C\sigma\}.$$

For $i \geq 0$, define $ARL_S(i)$ the ARL of $N_S$ applied to $\{e_{i+1}, e_{i+2}, \cdots\}$. Then,

$$E_{r=1}(N_S) = ARL_S(0).$$

By conditioning,

$$ARL_S(i) = 1 + ARL_S(i + 1)[\Phi(C - \delta(i)/\sigma) - \Phi(-C - \delta(i)/\sigma)],$$
which leads to the following formula.

\[
ARL_S(i) = 1 + \sum_{t=1}^{\infty} \prod_{i=1}^{t} \left[ \Phi(C - \delta(i - 1)/\sigma) - \Phi(-C - \delta(i - 1)/\sigma) \right].
\]

### 3.2.2 CUSUM Control Chart Procedure

Let

\[
S_H(t) = \max \{0, (e_t - k) + S_H(t-1)\},
\]

\[
S_L(t) = \max \{0, (-e_t - k) + S_L(t-1)\}.
\]

\(S_H(0) = 0\) and \(S_L(0) = 0\) are initialized. Then, stopping times

\[
N_{C}^H = \inf \{t \geq 1 : S_H(t) \geq h\}
\]

and

\[
N_{C}^L = \inf \{t \geq 1 : S_L(t) \geq h\}
\]

are the higher-side and lower-side CUSUM control chart procedures applied to \(\{e_1, e_2, \cdots\}\), respectively. \(k > 0\) is the CUSUM reference value, and \(h > 0\) is the CUSUM threshold value.

We note that

\[
\{S_H(t), t = 1, 2, \cdots\}
\]

and

\[
\{S_L(t), t = 1, 2, \cdots\}
\]

form two non-homogeneous continuous-state Markov chains.

For \(i \geq 0\), we define \(ARL_C^H(s, i)\) and \(ARL_C^L(s, i)\) the ARLs of the upper-side and lower-side CUSUM control chart procedure applied to \(\{e_{i+1}, e_{i+2}, \cdots\}\) given that \(S_H(i) = s\) and \(S_L(i) = s\), respectively. Then, for \(S_H(0) = 0\) and \(S_L(0) = 0\),

\[
E_{r=1}(N_{C}^H) = ARL_C^H(0, 0),
\]
By conditioning and the Markov property,

\[ ARL_C^H(s, i) \]

\[ = 1 P(s + (e_{i+1} - k) \geq h) + \int_{\{e_{i+1}: 0 + (e_{i+1} - k) < h\}} [1 + ARL_C^H(s + e_{i+1} - k, i + 1)]f_N(\delta(i), \sigma^2)(e_{i+1})de_{i+1} \]

\[ + \int_{\{e_{i+1}: 0 + (e_{i+1} - k) \leq 0\}} [1 + ARL_C^H(0, i + 1)]f_N(\delta(i), \sigma^2)(e_{i+1})de_{i+1} \]

\[ = 1 + \int_0^h ARL_C^H(x, i + 1)f_N(\delta(i) - k, \sigma^2)(x - s)d\]

\[ + ARL_C^H(0, i + 1)\Phi((-s - \delta(i) + k)/\sigma). \quad (3.1) \]

Using \(-\delta(\cdot)\) instead of \(\delta(\cdot)\) in (3.1), we obtain the equation for \(ARL_C^L(s, i)\).

Since \(\delta(t - 1) \rightarrow \delta(\infty)\) when \(t \rightarrow \infty\), from (3.1),

\[ ARL_C^H(0, \infty) = \frac{1 + \int_0^h ARL_C^H(x, \infty)f_N(\delta(\infty) - k, \sigma^2)(x)d\}

\[ = \frac{1}{1 - \Phi(k - \delta(\infty)/\sigma)} , \]

and

\[ ARL_C^H(s, \infty) \]

\[ = \left[ \frac{\Phi((k - \delta(\infty) - s)/\sigma)}{1 - \Phi(k - \delta(\infty)/\sigma)} \right] + \int_0^h ARL_C^H(x, \infty)[\frac{\Phi((k - \delta(\infty) - s)/\sigma)}{1 - \Phi(k - \delta(\infty)/\sigma)}f_N(\delta(\infty) - k, \sigma^2)(x) + f_N(\delta(\infty) - k, \sigma^2)(x - s)]d\}

which is a Fredholm equation of the second kind that can be solved numerically by using standard procedures such as Gaussian quadrature.

For an \(N\)-point Gaussian quadrature applied to \([0, h]\), let \(q_1, q_2, \cdots, q_N\) be the abscissas and \(w_1, w_2, \cdots, w_N\) the corresponding weights. Then, adapting the idea of the interpolatory formula observed by Nystrom, which is described in Press, Teukolsky, Vetterling and Flannery
we can accurately approximate (3.1) by

\[
ARL_H^T(s, i) = 1 + \sum_{j=1}^{N} w_j f_N(x(i) - k_j \sigma)(q_j - s)ARL_H^T(q_j, i + 1)
+ \Phi((-s - \delta(i) + k)/\sigma)ARL_H^T(0, i + 1).
\] (3.3)

Usually, $\delta(t)$ converges exponentially fast. For a sufficiently large $T$, say 100, we can initialize $ARL_H^T(s, T)$ to be $ARL_H^T(s, \infty)$ and then compute $ARL_H^T(s, i)$ recursively from $i = T - 1$ to $i = 0$ by using (3.3). For the initialization and recursions, only the function values at $s = 0, q_1, q_2, \ldots, q_N$ need to be evaluated.

In summary, to compute $ARL_H^T(0, 0)$, we carry out the following numerical procedure.

- Solve the integral equation (3.2) for $ARL_H^T(\cdot, \infty)$.
- For a sufficiently large integer $T \geq 0$, initialize $ARL_H^T(\cdot, T)$ to be $ARL_H^T(\cdot, \infty)$.
- For $i = T - 1, T - 2, \ldots, 0$, recursively obtain $ARL_H^T(\cdot, i)$ from $ARL_H^T(\cdot, i + 1)$ by using the interpolatory formula (3.3).
- Obtain $E_{r=1}(N_H^T) = ARL_H^T(0, 0)$, the final value of our interest.

### 3.2.3 EWMA Control Chart Procedure

For a chosen coefficient $\lambda \in (0, 1]$, let

\[
Z_i = \lambda e_t + (1 - \lambda)Z_{i-1}.
\]

$Z_0 = 0$

is initialized, and

\[
\sigma_2^2 = \sigma^2 / (2 - \lambda)
\]
is assumed. For $C > 0$, the stopping time

$$N_E = \inf\{t \geq 1 : |Z_t| \geq C \sigma_Z\}$$

is the EWMA control chart procedure applied to $\{e_1, e_2, \cdots\}$. $\lambda$ is the EWMA coefficient, and $C$ is the EWMA control limit.

$$\{Z_t, t = 1, 2, \cdots\}$$

is a non-homogeneous continuous-state Markov chain.

For $i \geq 0$, define $ARL_E(z, i)$ the ARL of $N_E$ applied to $\{e_{i+1}, e_{i+2}, \cdots\}$ given that $Z_i = z$.

Then, for $Z_0 = 0$,

$$E_{r=1}(N_E) = ARL_E(0, 0).$$

By conditioning and the Markov property,

$$ARL_E(z, i)$$

$$= 1P((1 - \lambda)z + \lambda e_{i+1} \geq C \sigma_Z)$$

$$+ \int_{\{e_{i+1} : (1 - \lambda)z + \lambda e_{i+1} < C \sigma_Z\}}[1 + ARL_E((1 - \lambda)z + \lambda e_{i+1}, i + 1)]f_{N(\delta(i), \sigma^2)}(e_{i+1})de_{i+1}$$

$$= 1 + \frac{1}{\lambda} \int_{-C \sigma_Z}^{C \sigma_Z} ARL_E(x, i + 1)f_{N(\delta(i), \sigma^2)}(\frac{x - (1 - \lambda)z}{\lambda})dx.$$

Then, a similar numerical procedure as in § 3.2.2 can be developed to obtain

$$E_{r=1}(N_E) = ARL_E(0, 0).$$
3.3 Numerical Analysis

In this section, we will numerically investigate the performance of the CUSUM, EWMA and Shewhart control chart procedures applied to the residuals. Obviously, for different types of $\delta(\cdot)$, the mean of residual after change, a change detection procedure may show different performance. Therefore, performance of the control chart procedures should be investigated for various types of $\delta(\cdot)$. To compare the performance of different control chart procedures under the ARL criterion, their in-control ARLs will be fixed at the same value.

Both the CUSUM and EWMA control chart procedures have two design parameters. CUSUM has its reference value $k$ and threshold value $h$. EWMA has its coefficient $\lambda$ and control limit $C$. When their in-control ARLs are fixed, the degree of freedom of choice of their design parameters will be reduced from 2 to 1. That is, once one of the two parameters is chosen, the other will be determined.

For a fixed in-control ARL value, at a particular size of change, we minimize the ARLs of the CUSUM and EWMA control chart procedures. This can be done by optimizing one of their design parameters. We choose to optimize the reference value $k$ for CUSUM and the coefficient $\lambda$ for EWMA.

In the rest of this section, we will present numerical results in three typical cases of $\delta(\cdot)$. In each case, we will first present data in a table, where the CUSUM reference values $k$ and EWMA coefficients $\lambda$ optimal for different sizes of change are given, together with the ARL values for each set of optimal design parameters. The optimum ARLs of CUSUM and EWMA at each considered size of change are superscripted with * in the tables. We will then plot the optimum ARLs versus sizes of change, and the ARL curves for chosen design parameters. In this section, we will assume that
the standard deviation of residual is $\sigma = 1$.

### 3.3.1 Results on the Machining Process Subject to a Step Shift

Consider the steady-state machining process subject to an additive step shift of size $\Delta$ in its state equation, that is, $\Delta(i) \equiv \Delta$ for $i \geq 0$. Then the mean of residual after change is

$$\delta(i) = (1 - w)^i \Delta$$

for $i \geq 0$, where $w = 0.553$.

<table>
<thead>
<tr>
<th>$\Delta$</th>
<th>0</th>
<th>0.5</th>
<th>1</th>
<th>1.5</th>
<th>2</th>
<th>2.5</th>
<th>3</th>
<th>3.5</th>
<th>4</th>
<th>4.5</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUSUM</td>
<td>100</td>
<td>96.0*</td>
<td>87.8*</td>
<td>73.5*</td>
<td>53.9*</td>
<td>33.5*</td>
<td>17.2*</td>
<td>7.4*</td>
<td>3.0*</td>
<td>1.5*</td>
<td>1.1*</td>
</tr>
<tr>
<td>$k = 0.9$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EWMA</td>
<td>100</td>
<td>96.6*</td>
<td>84.5*</td>
<td>62.0*</td>
<td>35.0*</td>
<td>14.6*</td>
<td>4.9*</td>
<td>2.0</td>
<td>1.3</td>
<td>1.1</td>
<td>1.0</td>
</tr>
<tr>
<td>$\lambda = 0.35$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda = 0.45$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda = 0.70$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shewhart</td>
<td>100</td>
<td>98.8</td>
<td>94.3</td>
<td>84.9</td>
<td>69.6</td>
<td>50.1</td>
<td>30.7</td>
<td>15.8</td>
<td>6.9</td>
<td>2.9</td>
<td>1.5</td>
</tr>
</tbody>
</table>

Table 3.1: Optimum ARLs and parameters of the CUSUM, EWMA and Shewhart control chart procedures applied to the residuals of the machining process with $w = 0.553$ subject to an additive step shift of size $\Delta$ in its state equation when their in-control ARLs are 100.

From Table 3.1 and Figure 3.1, we have the following observations on the performance of the CUSUM, EWMA and Shewhart control chart procedures when the mean of residual after change is $\delta(i) = (1 - w)^i \Delta$.

1. Roughly, EWMA performs the best in this case.
2. Only when the size of change is very small, CUSUM performs better than EWMA.

3. Shewhart performs the worst for any considered size of change.
Figure 3.1: Optimum ARLs versus sizes of change (upper panel) and the ARL curves with the chart parameters optimizing the ARLs at $\Delta = 2$ (lower panel) of the CUSUM, EWMA and Shewhart control chart procedures applied to the residuals of the machining process with $w = 0.553$ subject to an additive step shift of size $\Delta$ in its state equation, when the fixed in-control ARL is 100.
3.3.2 Results on the Machining Process Subject to a Drift Rate Change

Consider the steady-state machining process subject to a drift rate change of $c'$ in its state equation. Then the mean of residual after change is

$$\delta(i) = \frac{[1 - (1 - w)^{i+1}] c'}{w}$$  \hspace{1cm} (3.5)

for $i \geq 0$, where $w = 0.553$.

<table>
<thead>
<tr>
<th>$c'$</th>
<th>0</th>
<th>0.5</th>
<th>1</th>
<th>1.5</th>
<th>2</th>
<th>2.5</th>
<th>3</th>
<th>3.5</th>
<th>4</th>
<th>4.5</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUSUM</td>
<td>$k = 0.44$</td>
<td>300</td>
<td>10.3*</td>
<td>4.7</td>
<td>3.3</td>
<td>2.6</td>
<td>2.2</td>
<td>2.0</td>
<td>1.9</td>
<td>1.8</td>
<td>1.6</td>
</tr>
<tr>
<td></td>
<td>$k = 0.81$</td>
<td>300</td>
<td>12.1</td>
<td>4.3*</td>
<td>2.6*</td>
<td>2.2</td>
<td>1.9</td>
<td>1.7</td>
<td>1.5</td>
<td>1.3</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>$k = 1.10$</td>
<td>300</td>
<td>15.4</td>
<td>4.5</td>
<td>2.8</td>
<td>2.1*</td>
<td>1.7*</td>
<td>1.5</td>
<td>1.3</td>
<td>1.2</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>$k = 2.00$</td>
<td>300</td>
<td>27.4</td>
<td>6.3</td>
<td>3.1</td>
<td>2.1</td>
<td>1.7</td>
<td>1.4*</td>
<td>1.2</td>
<td>1.1*</td>
<td>1.0*</td>
</tr>
<tr>
<td>EWMA</td>
<td>$\lambda = 0.12$</td>
<td>300</td>
<td>12.0*</td>
<td>5.7</td>
<td>4.1</td>
<td>3.3</td>
<td>2.8</td>
<td>2.5</td>
<td>2.2</td>
<td>2.0</td>
<td>1.9</td>
</tr>
<tr>
<td></td>
<td>$\lambda = 0.28$</td>
<td>300</td>
<td>13.6</td>
<td>5.2*</td>
<td>3.6</td>
<td>2.9</td>
<td>2.4</td>
<td>2.1</td>
<td>1.8</td>
<td>1.6</td>
<td>1.4</td>
</tr>
<tr>
<td></td>
<td>$\lambda = 0.43$</td>
<td>300</td>
<td>17.0</td>
<td>5.4</td>
<td>3.5*</td>
<td>2.7*</td>
<td>2.2</td>
<td>1.8</td>
<td>1.6</td>
<td>1.3</td>
<td>1.2</td>
</tr>
<tr>
<td></td>
<td>$\lambda = 0.71$</td>
<td>300</td>
<td>28.2</td>
<td>6.7</td>
<td>3.8</td>
<td>2.7</td>
<td>2.1*</td>
<td>1.7*</td>
<td>1.4*</td>
<td>1.2*</td>
<td>1.1*</td>
</tr>
<tr>
<td>Shewhart</td>
<td>300</td>
<td>48.5</td>
<td>9.3</td>
<td>3.9</td>
<td>2.5</td>
<td>1.9</td>
<td>1.5</td>
<td>1.3</td>
<td>1.1</td>
<td>1.1</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Table 3.2: Optimum ARLs and parameters of the CUSUM, EWMA and Shewhart control chart procedures applied to the residuals of the machining process with $w = 0.553$ subject to a drift rate change of $c'$ in its state equation when their in-control ARLs are 300.

From Table 3.2 and Figure 3.2, we have the following observations on the performance of the CUSUM, EWMA and Shewhart control chart procedures when the mean of residual after
change is (3.5).

1. Optimized CUSUM control chart procedures produce the smallest optimum ARL values.

2. The CUSUM control chart procedure optimizing the ARL at a considered size of change produces a lower ARL curve than the EWMA optimizing the ARL at the same size of change, and also a lower curve than Shewhart.

3. Shewhart performs almost as well as CUSUM for large size of changes.

4. Shewhart performs better than EWMA for medium and large size of changes.

5. EWMA does not perform as well as CUSUM.

6. EWMA performs better than Shewhart for small size of changes.
Figure 3.2: Optimum ARLs versus sizes of change (upper panel) and the ARL curves with the chart parameters optimizing the ARLs at $c' = 2$ (lower panel) of the CUSUM, EWMA and Shewhart control chart procedures applied to the residuals of the machining process with $w = 0.553$ subject to a drift rate change of $c'$ in its state equation, when the fixed in-control ARL is 300.
3.3.3 Results on the Cutting Process

Consider the AR(2) cutting process with parameters in (1.4) subject to an observational additive step shift of size $\Delta$, that is, $\Delta(i) \equiv \Delta$ for $i \geq 0$. Then the mean of residual after change is

$$\delta(i) = (1 - 0.45B - 0.2B^2)\Delta(i)$$

$$= \begin{cases} 
\Delta, & \text{for } i = 0, \\
0.55\Delta, & \text{for } i = 1, \\
0.35\Delta, & \text{for } i > 1.
\end{cases} \quad (3.6)$$

From Table 3.3 and Figure 3.3, we have the following observations on the performance of the CUSUM, EWMA and Shewhart control chart procedures when the mean of residual after change is (3.6).

1. CUSUM and EWMA perform better than Shewhart, especially for small and medium size of changes.

2. CUSUM performs better than EWMA for smaller size of changes.

3. EWMA performs better than CUSUM for larger size of changes.
<table>
<thead>
<tr>
<th>$\Delta$</th>
<th>0</th>
<th>0.5</th>
<th>1</th>
<th>1.5</th>
<th>2</th>
<th>2.5</th>
<th>3</th>
<th>3.5</th>
<th>4</th>
<th>4.5</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$k = 0.1$</td>
<td>100</td>
<td>39.5*</td>
<td>20.9</td>
<td>13.0</td>
<td>8.9</td>
<td>6.5</td>
<td>4.9</td>
<td>3.8</td>
<td>3.1</td>
<td>2.6</td>
</tr>
<tr>
<td></td>
<td>$k = 0.2$</td>
<td>100</td>
<td>40.4</td>
<td>20.4*</td>
<td>11.9</td>
<td>7.7</td>
<td>5.3</td>
<td>3.8</td>
<td>2.9</td>
<td>2.3</td>
<td>1.9</td>
</tr>
<tr>
<td>CUSUM</td>
<td>$k = 0.3$</td>
<td>100</td>
<td>42.1</td>
<td>20.8</td>
<td>11.6*</td>
<td>7.0</td>
<td>4.6</td>
<td>3.2</td>
<td>2.3</td>
<td>1.8</td>
<td>1.5</td>
</tr>
<tr>
<td></td>
<td>$k = 0.5$</td>
<td>100</td>
<td>46.1</td>
<td>22.9</td>
<td>12.1</td>
<td>6.6*</td>
<td>3.8</td>
<td>2.4</td>
<td>1.7</td>
<td>1.3</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>$k = 0.7$</td>
<td>100</td>
<td>50.0</td>
<td>25.7</td>
<td>13.3</td>
<td>6.9</td>
<td>3.7*</td>
<td>2.1*</td>
<td>1.5</td>
<td>1.2</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>$k = 1.0$</td>
<td>100</td>
<td>55.1</td>
<td>30.2</td>
<td>16.0</td>
<td>8.1</td>
<td>4.0</td>
<td>2.1</td>
<td>1.4*</td>
<td>1.1*</td>
<td>1.0*</td>
</tr>
<tr>
<td></td>
<td>$\lambda = 0.01$</td>
<td>100</td>
<td>49.8*</td>
<td>23.2</td>
<td>13.7</td>
<td>9.0</td>
<td>6.3</td>
<td>4.6</td>
<td>3.5</td>
<td>2.8</td>
<td>2.3</td>
</tr>
<tr>
<td></td>
<td>$\lambda = 0.05$</td>
<td>100</td>
<td>51.8</td>
<td>22.3*</td>
<td>11.8</td>
<td>6.9</td>
<td>4.4</td>
<td>3.1</td>
<td>2.4</td>
<td>2.1</td>
<td>2.0</td>
</tr>
<tr>
<td>EWMA</td>
<td>$\lambda = 0.10$</td>
<td>100</td>
<td>55.7</td>
<td>23.5</td>
<td>11.3*</td>
<td>6.0</td>
<td>3.6</td>
<td>2.5</td>
<td>2.1</td>
<td>1.9</td>
<td>1.7</td>
</tr>
<tr>
<td></td>
<td>$\lambda = 0.20$</td>
<td>100</td>
<td>62.6</td>
<td>27.2</td>
<td>12.0</td>
<td>5.6*</td>
<td>3.0</td>
<td>2.1</td>
<td>1.7</td>
<td>1.5</td>
<td>1.3</td>
</tr>
<tr>
<td></td>
<td>$\lambda = 0.35$</td>
<td>100</td>
<td>70.0</td>
<td>33.4</td>
<td>14.3</td>
<td>6.0</td>
<td>2.9*</td>
<td>1.8</td>
<td>1.4</td>
<td>1.2</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>$\lambda = 0.55$</td>
<td>100</td>
<td>77.2</td>
<td>41.8</td>
<td>19.0</td>
<td>7.7</td>
<td>3.2</td>
<td>1.7*</td>
<td>1.3*</td>
<td>1.1*</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>$\lambda = 0.75$</td>
<td>100</td>
<td>82.6</td>
<td>50.5</td>
<td>25.4</td>
<td>10.8</td>
<td>4.2</td>
<td>1.9</td>
<td>1.3</td>
<td>1.1</td>
<td>1.0*</td>
</tr>
<tr>
<td>Shewhart</td>
<td>100</td>
<td>88.7</td>
<td>64.5</td>
<td>40.9</td>
<td>23.3</td>
<td>12.0</td>
<td>5.7</td>
<td>2.8</td>
<td>1.5</td>
<td>1.1</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Table 3.3: Optimum ARLs and parameters of the CUSUM, EWMA and Shewhart control chart procedures applied to the residuals of the $AR(2)$ cutting process subject to an observational additive step shift of size $\Delta$ when their in-control ARLs are 100.
Figure 3.3: Optimum ARLs versus sizes of change (upper panel) and the ARL curves with the chart parameters optimizing the ARLs at $\Delta = 2$ (lower panel) of the CUSUM, EWMA and Shewhart control chart procedures applied to the residuals of the $AR(2)$ cutting process subject to an observational additive step shift of size $\Delta$, when the fixed in-control ARL is 100.
3.3.4 Summary of the Performance of Control Charts

Under the ARL Criterion, we have studied the performance of the CUSUM, EWMA and Shewhart control chart procedures under three different situations, where the mean of residual after change represented by $\delta(i)$ for $i \geq 0$ decreases to 0 rapidly (3.4), decreases at first and then stabilizes at a nonzero level (3.6), and gradually increases to a constant level (3.5).

From the analyses, we have the following results.

1. When $\delta(i)$ for $i \geq 0$ decreases to 0 rapidly, EWMA roughly performs the best among the three control chart procedures.

2. Otherwise, CUSUM roughly performs the best.

3. Generally, Shewhart performs well for detection of very large size of changes.
Chapter 4

Sequential Testing on Process Mean Levels
4.1 Introduction

Consider a situation for the residual process \( \{e_t, t = 1, 2, \ldots \} \). For a fixed time point \( r_0 \geq 1 \), either the additive change occurs at \( r_0 \), or no change occurs at all. That is, the change point \( r \in \{r_0, \infty \} \). We want to discriminate \( r = r_0 \) from \( r = \infty \) by sequentially observing \( e_t \) until a decision based on the observations can be made. Therefore, we sequentially test \( H_0: r = \infty \) versus \( H_1: r = r_0 \). Before \( r_0 \), \( e_t \) behaves the same under both \( H_0 \) and \( H_1 \), and starting at \( r_0 \), \( e_t \) behaves differently under \( H_0 \) and \( H_1 \). Specifically, starting at \( r_0 \), \( E(e_t) \) is time-varying according to \( \delta(\cdot) \) under \( H_1 \), and \( E(e_t) \) is constant zero under \( H_0 \). Therefore, starting at \( r_0 \), we actually sequentially test

\[
H_0: E(e_i) = 0, \quad \text{for } i = 1, 2, \ldots, \quad \text{versus} \quad H_1: E(e_i) = \delta(i - 1), \quad \text{for } i = 1, 2, \ldots,
\]

based on the observations \( e_t \). We call these two hypotheses the mean level hypotheses. Since \( \{r_0, \infty\} \subset \{1, 2, \ldots, \infty\} \), the mean level hypothesis testing is a special situation of the general change detection problem formulated in § 2.8 and originally in § 1.3.

Another situation that leads to formulation of the mean level hypothesis testing is when it is suspected that an additive change has taken place before observation is started. If we want to clarify this suspicious situation, we will have to test it. If the suspicion is true, then \( E(e_t) \) is time-varying according to \( \delta(\cdot) \) starting at the beginning of the observation. Therefore, we will need to test the mean level hypotheses.

The problem of sequential testing on process mean levels also stands on its own. It has already been a problem widely accepted and studied in the area of signal detection. Basically, in the circumstances of signal detection, we use the mean level hypothesis testing formulation to test
of a known signal in a noise process. The problem has been studied by many authors. Liptser and Shiryaev (1978) in their §17.6 studied sequential testing of two hypotheses that observations follow a noise process or the observations follow the noise process containing a known signal additively. Cochlar and Vrana (1978) studied sequential testing of two hypotheses that observations have a joint probability density function, say, \( f(x^n_1|0) \), or the observations have a different joint probability density function, say, \( f(x^n_1|1) \). They also particularly studied sequential testing on presence or absence of a known signal in a Gaussian noise process. Liu and Blostein (1992) studied sequential testing of two hypotheses that independent observations have a joint probability density function \( \prod_{i=1}^{t} f_0(x_i) \), or they have a joint probability density function \( \prod_{i=1}^{t} f_1(x_i) \). They also discussed the applications of this sequential testing problem in communications. Recently, Kailath and Poor (1998) has comprehensively reviewed the related topics of this sequential testing problem.

Other applications of the mean level hypothesis testing problem may arise in system identification. Consider that the system status or process mean level is unknown to the observer, but can only be in two possible states, say, normal or abnormal. Regard the normal state as an in-control state. Then, the means of residuals are all zero if the system or process is in the normal state, and are time-varying following a pattern if in the abnormal state. Therefore, the unknown system status or process mean level forms a set of mean level hypotheses. By sequentially testing the mean level hypotheses, we can have the unknown state identified.

We note that the classical Wald's SPRT problem, where \( \delta(i) \) for \( i \geq 0 \) is a constant, is a special case of the mean level hypothesis testing problem. SPRT has been applied in different areas. Since the time-varying feature is more prevalent than the time-invariant feature, the
mean level hypothesis testing is more applicable.

Our objective here is to derive an optimal sequential test for the mean level hypotheses. When $\delta(i - 1)$ is a constant for $i = 1, 2, \ldots$, the optimal sequential test is the SPRT proposed by Wald (1945), and its optimality was proved by Wald and Wolfowitz (1948), and later by Lehmann (1959), Chow and Robbins (1963) and others. For $\delta(\cdot)$ taking a general time-varying pattern, only a few works on the optimal sequential tests exist in the literature. Cochlar and Vrana (1978) derived a Bayes optimum sequential test minimizing the expected sample size over the a posteriori probability distribution of the two hypotheses for given type-I and type-II errors. Liu and Blostein (1992) derived the same Bayes optimum sequential test, and also obtained some properties of the optimal stopping boundaries. Liptser and Shiryaev (1978) derived an optimal sequential test under a different criterion (see the (17.115) in their Theorem 17.8).

In this chapter, we will derive the optimal sequential test that minimizes the expected sample size when the process is in the out-of-control state for given upper bounds of type-I and type-II errors. We will first formulate the optimal sequential testing problem as an optimal stopping problem, then derive the optimal stopping rule, study some important properties of the optimal stopping boundaries, and develop an algorithm for computation of the optimal stopping boundaries, and present an example.

In this chapter, we assume that

- $e_t$ has a normal distribution with standard deviation $\sigma = 1$. 


4.2 Formulation as an Optimal Stopping Problem

For \( n = 0, 1, 2, \cdots \), let \( \mathcal{F}_n \) be the minimum \( \sigma \)-algebra induced by \( \{ e_t, t = 1, 2, \cdots, n \} \), where specially, \( \{ e_t, t = 1, \cdots, 0 \} = \emptyset, \mathcal{F}_0 = \{ \emptyset, \mathcal{R} \} \). Let \( e_t^0 \) be a null vector. A sequential test \( S = (\tau, d) \) consists of a stopping time \( \tau \geq 0 \) with respect to \( \{ \mathcal{F}_n, n = 0, 1, 2, \cdots \} \) and a terminal decision rule \( d \) that is a function

\[
d(e^n_\tau) = \begin{cases} 
0, & \text{to accept } H_0, \\
1, & \text{to accept } H_1,
\end{cases}
\]

given \( \tau = n \) and \( e^n_\tau \). \( \tau = 0 \) represents no sampling at all. A fixed sample size test is a special case of sequential tests, for which \( \tau \equiv m \), the fixed sample size. The type-I and type-II errors of a sequential test \( S \) are

\[
\alpha = P(d = 1|H_0 \text{ is true}) \quad \text{and} \quad \beta = P(d = 0|H_1 \text{ is true}),
\]

respectively. When it is necessary, we use \( \tau(S), \alpha(S) \) and \( \beta(S) \) denote the stopping time, type-I and type-II errors associated with a sequential test \( S \), respectively.

Our objective here is to find an optimal sequential test minimizing

\[
E(\tau(S)|H_1)
\]

subject to the constraints \( \alpha(S) \leq \alpha_0 \) and \( \beta(S) \leq \beta_0 \), where \( 0 < \alpha_0, \beta_0 < 1 \) are specified values. That is, we are interested in finding an optimal sequential test, satisfying the prescribed upper bounds of decision errors, which will stop sampling most quickly on average when \( H_1 \) is true.

For convenience, let \( E_1(\cdot) \) denote \( E(\cdot|H_1) \).

This objective leads us to consider the following auxiliary problem: for \( c_0, c_1 > 0 \), finding the optimal sequential test \( S^* \) attaining

\[
\inf_{S \in \{ \text{all sequential tests} \}} \{ c_0 \alpha(S) + c_1 \beta(S) + E_1(\tau(S)) \}.
\]

(4.2)
\(c_0\) and \(c_1\) can be interpreted as the loss incurred by falsely rejecting \(H_0\) and \(H_1\), respectively. The sufficiency of the auxiliary problem for our objective (4.1) results from the fact that, for any other sequential test \(S'\), if \(\alpha(S') \leq \alpha(S^*)\) and \(\beta(S') \leq \beta(S^*)\), then we must have \(E_1(\tau(S')) \geq E_1(\tau(S^*))\). Otherwise, it contradicts (4.2). Therefore, if we have \(\alpha(S^*) = \alpha_0\) and \(\beta(S^*) = \beta_0\) by properly choosing \(c_0\) and \(c_1\), then \(S^*\) will be the optimal sequential test we want. Similar auxiliary problems were first used by Wald and Wolfowitz (1948) to prove the optimality of Wald’s SPRT, and subsequently used and analyzed by some other authors in optimality proofs.

For convenience, for \(t = 1, 2, \ldots\), let

\[
\mu_t = \delta(t - 1).
\]

Let \(f_{0,t}(\cdot)\) and \(f_{1,t}(\cdot)\) denote the density functions of \(e_t\) under \(H_0\) and \(H_1\), respectively. Then

\[
f_{0,t}(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{x^2}{2}}
\]

and

\[
f_{1,t}(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-(x - \mu_t)^2}{2}}.
\]

Let

\[
L_0(e^n_t) = \prod_{t=1}^{n} f_{0,t}(e_t)
\]

and

\[
L_1(e^n_t) = \prod_{t=1}^{n} f_{1,t}(e_t),
\]

which are the likelihood functions. For convenience, denote \(L_0(e^n_1) = 1\) and \(L_1(e^n_1) = 1\).

For any stopping time \(\tau\), the optimal terminal decision rule \(d^*\) minimizing \(c_0\alpha + c_1\beta\), and

61
therefore minimizing $c_0 \alpha + c_1 \beta + E_1(\tau)$, is when $\tau = n$,

$$d^n(e_1^n) = \begin{cases} 
0 \ (\text{to accept } H_0), & \text{if } c_0 L_0(e_1^n) \geq c_1 L_1(e_1^n), \\
1 \ (\text{to accept } H_1), & \text{if } c_0 L_0(e_1^n) < c_1 L_1(e_1^n). 
\end{cases} \quad (4.3)$$

This is because of the following inequality:

$$c_0 \alpha + c_1 \beta = \sum_{n=0}^{\infty} \left[ c_0 \int_{\tau=n, d(e_1^n)=1} L_0(e_1^n)de_1^n + c_1 \int_{\tau=n, d(e_1^n)=0} L_1(e_1^n)de_1^n \right]$$

$$\geq \sum_{n=0}^{\infty} \left[ \int_{\tau=n, c_0 L_0(e_1^n) < c_1 L_1(e_1^n)} c_0 L_0(e_1^n)de_1^n + \int_{\tau=n, c_0 L_0(e_1^n) \geq c_1 L_1(e_1^n)} c_1 L_1(e_1^n)de_1^n \right]$$

$$= \sum_{n=0}^{\infty} \int_{\{\tau=n\}} \min\{c_0 L_0(e_1^n), c_1 L_1(e_1^n)\}de_1^n$$

$$= \sum_{n=0}^{\infty} \int_{\{\tau=n\}} \min\{c_0 \frac{L_0(e_1^n)}{L_1(e_1^n)}, c_1\}dL_1(e_1^n).$$

Therefore, to find the optimal sequential test of (4.2), we will find an optimal stopping time $\tau^*$ minimizing

$$E_1(\tau + \min\{c_0 \frac{L_0(e_1^n)}{L_1(e_1^n)}, c_1\}). \quad (4.4)$$

The optimal stopping time $\tau^*$, together with the optimal terminal decision rule $d^*$ (4.3), will render an optimal sequential test $S^* = (\tau^*, d^*)$ of our objective (4.1).

In the next section, we derive the form of the optimal stopping time $\tau^*$. 

62
### 4.3 Derivation of the Optimal Stopping Rule

Define the stopping loss at stage $n = 0, 1, \cdots$ as

$$Z_n = n + \min\left\{c_0 \frac{L_0(e^n_i)}{L_1(e^n_i)}, c_1\right\}.$$

Denote $\Lambda_n$ the likelihood ratio at stage $n$, which is

$$\Lambda_n = \frac{L_0(e^n_i)}{L_1(e^n_i)} = \exp\left\{-\sum_{i=1}^{n} \mu_i(e_i - \frac{\mu_i}{2})\right\}. $$

Then,

$$Z_n = n + \min\{c_0 \Lambda_n, c_1\}.$$

Let $z_n$ and $\lambda_n$ be the realizations of $Z_n$ and $\Lambda_n$, respectively. By the definition of $Z_n$ and (4.4), the optimal stopping time $\tau^*$ is the optimal stopping time with respect to $\{(Z_n, \mathcal{F}_n), n = 0, 1, \cdots\}$, which satisfies

$$E_1(Z_{\tau^*}) = \inf_{\tau} E_1(Z_\tau).$$

For $K = 0, 1, \cdots$ and $n = 0, 1, \cdots, K$, by backward recursion, define

$$\beta_n^K = \min\{z_n, E_1(\beta_{n+1}^K | \mathcal{F}_n)\}$$

with

$$\beta_{K+1}^K \equiv \infty.$$

$\beta_n^K$ can be interpreted as the minimum expected loss at stage $n$ for a finite stopping time $\tau \leq K$, which means that we must stop sampling and make a decision before or at stage $K$. By the definition, for any $n \leq K$, $\beta_n^K \geq 0$, and by backward induction, $\beta_{n+1}^K \leq \beta_n^K$. Therefore, for $n = 0, 1, \cdots$,

$$\beta_n = \lim_{K \to \infty} \beta_n^K$$
exists. Since in

\[ Z_n = n + \min\{c_0A_n, c_1\}, \]

\[ \min\{c_0A_n, c_1\} \] is bounded between 0 and \( c_1 \) for all \( n = 0, 1, \cdots \), and \( \{n\} \) is an increasing sequence, by Theorem 2 of Chow and Robbins (1963), the optimal stopping time \( \tau^* \) with respect to \( \{(Z_n, \mathcal{F}_n)\} \) exists and

\[ \tau^* = \min\{n : z_n = \beta_n\}. \]  \hspace{1cm} (4.5)

For \( \lambda \geq 0 \), let

\[ h(\lambda) = \min\{c_0\lambda, c_1\}. \]

Then, for \( n = 0, 1, \cdots \),

\[ z_n = n + h(\lambda_n), \]

and

\[ z_{n+1} = n + 1 + h(\lambda_n) + h(e_{n+1}) \]

Hence, by the definition, for given \( n, K \) and the deterministic sequence \( \{\mu_i\} \), \( \beta^K_n \) is a function of \( \lambda_n \) only, and we may denote it by \( \beta^K_n(\lambda_n) \).

For \( n = 0, 1, \cdots, K + 1 \), define

\[ g^K_n(\lambda_n) = \beta^K_n(\lambda_n) - n. \]

By the definitions, for \( n = 0, 1, \cdots, K \),

\[ g^K_n(\lambda_n) = \min\{h(\lambda_n), G^K_{n+1}(\lambda_n) + 1\}; \]  \hspace{1cm} (4.6)

where

\[ G^K_{n+1}(\lambda_n) = \int_{-\infty}^{\infty} g^K_{n+1}(\lambda_n \exp(-\mu_{n+1}(e_{n+1} - \frac{\mu_{n+1}}{2})) \right) f_{1, (n+1)}(e_{n+1}) de_{n+1} \]

\[ = \int_{-\infty}^{\infty} g^K_{n+1}(\lambda_n \exp(-\mu_{n+1}(x + \frac{\mu_{n+1}}{2})) \right) f_{N(0,1)}(x) dx, \]  \hspace{1cm} (4.7)
$f_{N(0,1)}(\cdot)$ is the standard normal density function. For $n = 0, 1, \cdots$, because $\lim_{K \to \infty} \beta_n^K$ exists,

$$g_n(\lambda_n) = \lim_{K \to \infty} \beta_n^K(\lambda_n)$$

(4.8)

exists. Since

$$0 \leq \beta_n^K(\lambda_n) \leq h(\lambda_n) \leq c_1,$$

by the Lebesgue theorem of dominated convergence,

$$g_n(\lambda_n) = \min\{h(\lambda_n), G_{n+1}(\lambda_n) + 1\},$$

(4.9)

where

$$G_{n+1}(\lambda_n) = \int_{-\infty}^{\infty} g_{n+1}(\lambda_n \exp\{-\mu_n \frac{(x + \frac{\mu_n}{2})}{2}\}) f_{N(0,1)}(x) dx.$$  

(4.10)

By the definitions and (4.5), the optimal stopping time is

$$\tau^* = \min\{n : g_n(\lambda_n) = h(\lambda_n)\}.$$  

(4.11)

The functions $g_n(\lambda)$ and $G_n(\lambda)$ will provide us convenience for studying the structure of $\tau^*$. We first derive some important properties of them, which are the following proposition.

**Proposition 1** For any $n = 0, 1, \cdots$,

(i) both $g_n(\lambda)$ and $G_n(\lambda)$ are concave in $\lambda \geq 0$;

(ii) $g_n(0) = 0$, $g_n(\lambda) = c_1$ for sufficiently large $\lambda$, and $g_n(\lambda)$ is increasing in $\lambda \geq 0$;

(iii) $G_n(0) = 0$, $G_n(\infty) = c_1$, and $G_n(\lambda)$ is increasing in $\lambda \geq 0$.

**Proof of Proposition 1:**

(i) By the definitions, for any integer $K \geq 0$, $g_n^K(\lambda) = h(\lambda)$ is obviously concave in $\lambda \geq 0$. For any $n = 0, 1, \cdots, K$, if $g_n^K(\lambda)$ is concave, then by the definition of concavity and (4.7), it can be verified that $G_n^K(\lambda)$ is concave in $\lambda \geq 0$. Hence, by (4.6) and backward induction, $g_n^K(\lambda)$
and $G_n^K(\lambda)$ are concave in $\lambda \geq 0$. Then, for all $n = 0, 1, \cdots$, by the definition of concavity and (4.8), $g_n(\lambda)$ is concave in $\lambda \geq 0$, so is therefore $G_n(\lambda)$ by (4.10).

(ii) $g_n(0) = h(0) = 0$ is obvious.

If $c_i \leq 1$, then by (4.9), for $\lambda \geq c_i/c_0$ and $n = 0, 1, \cdots$,

$$g_n(\lambda) = c_i.$$ 

Suppose $c_i > 1$. By (4.9), for all $n = 0, 1, \cdots$ and $\lambda \geq c_i/c_0$,

$$g_n(\lambda) \geq 1.$$ 

Let $[c_i]$ denote the integer part of $c_i$. For any $n = 0, 1, \cdots$, consider $m = n + [c_i] + 1$ and $\epsilon \in (0, 1)$.

$$g_m(\lambda) \geq 1,$$

for $\lambda \geq c_i/c_0$. Hence, by (4.10), there exists $v_{m-1} \geq c_i/c_0$, such that for $\lambda \geq v_{m-1}$,

$$G_m(\lambda) \geq 1 - \frac{\epsilon}{2^{m-n}},$$

and by (4.9),

$$g_{m-1}(\lambda) \geq \min\{c_i, 2 - \frac{\epsilon}{2^{m-n}}\}.$$ 

Once again, by (4.10), there exists $v_{m-2} \geq c_i/c_0$, such that for $\lambda \geq v_{m-2}$,

$$G_{m-1}(\lambda) \geq \min\{c_i, 2 - \frac{\epsilon}{2^{m-n}}\} - \frac{\epsilon}{2^{m-n-1}},$$

and by (4.9),

$$g_{m-2}(\lambda) \geq \min\{c_i, 3 - \frac{\epsilon}{2^{m-n}} - \frac{\epsilon}{2^{m-n-1}}\}.$$ 

Thus, by backward recursion, there exists $v_n \geq c_i/c_0$, such that for $\lambda \geq v_n$,

$$g_n(\lambda) = g_{m-([c_i]+1)}(\lambda)$$
\[
\geq \min\{c_1, [c_1] + 2 - \sum_{i=0}^{[c_1]} \frac{\epsilon}{2^{m-n-i}} \}
\]

\[
= \min\{c_1, [c_1] + 2 - \sum_{i=1}^{[c_1]+1} \frac{\epsilon}{2^i} \}
\]

\[
\geq \min\{c_1, [c_1] + 2 - \epsilon \}
\]

\[
= c_1.
\]

That is, for \( \lambda \geq v_n, g_n(\lambda) = c_1. \)

For any \( n = 0, 1, \cdots, \) since \( g_n(0) = 0 \) and \( g_n(\lambda) = c_1 \) for \( \lambda \geq v_n, \) by its concavity, \( g_n(\lambda) \) must be increasing in \( \lambda \geq 0. \) Therefore, (ii) is proved.

(iii) follows (ii) by (4.10). \hspace{1cm} \text{Proposition 1 is proved.}

Now, we can establish the structure of the optimal stopping time \( \tau^*. \)

\textbf{Theorem 1} \hspace{0.5cm} \text{The optimal stopping time has the form}

\[
\tau^* = \min\{n \geq 0: \lambda_n \in (A_n, B_n)\},
\]

where \( \{A_n, n = 0, 1, \cdots, \} \) and \( \{B_n, n = 0, 1, \cdots, \} \) are two deterministic sequences. If \( c_1 > 1, \) then for \( n = 0, 1, \cdots, \)

\[
0 < \frac{1}{c_0} < A_n \leq \frac{c_1}{c_0} \leq B_n,
\]

and \( A_n = c_1/c_0 \) if and only if \( B_n = c_1/c_0. \) If \( 0 < c_1 \leq 1, \) then for any \( n = 0, 1, \cdots, \)

\( A_n = B_n \equiv c_1/c_0. \)

\textbf{Proof of Theorem 1:}

From (4.11) and (4.9), for \( n = 0, 1, \cdots, \) we need to compare \( h(\lambda) \) with \( G_{n+1}(\lambda) + 1 \) for \( \lambda \geq 0. \)

Suppose that \( c_1 > 1. \) By Proposition 1,

\[
G_{n+1}(0) + 1 = 1 > h(0) = 0.
\]
From the proof of Proposition 1 (ii), there exists $v_n \geq c_1/c_0$, such that for $\lambda \geq v_n$,

$$G_{n+1}(\lambda) + 1 > h(\lambda) = c_1.$$ 

Therefore, if there exists $A_n \in (0, c_1/c_0]$, such that

$$G_{n+1}(A_n) + 1 = c_0 A_n = h(A_n),$$

(4.12)

then, by the properties of $G_{n+1}(\lambda)$ in Proposition 1 (i) and (iii), firstly $A_n > 1/c_0$, secondly such an $A_n$ is unique, and thirdly there uniquely exists $B_n \in [c_1/c_0, \infty)$ such that

$$G_{n+1}(B_n) + 1 = c_1 = h(B_n).$$

If $A_n = c_1/c_0$, then $B_n = c_1/c_0$, and vice versa. Furthermore, for $\lambda \in (A_n, B_n)$,

$$G_{n+1}(\lambda) + 1 < h(\lambda),$$

and for $\lambda \in (A_n, B_n)$,

$$G_{n+1}(\lambda) + 1 \geq h(\lambda).$$

When $A_n = B_n$, $(A_n, B_n) = \emptyset$.

If there does not exist such an $A_n \in (0, c_1/c_0]$ that (4.12) holds, that is, for all $\lambda \in (0, c_1/c_0]$,

$$G_{n+1}(\lambda) + 1 > h(\lambda),$$

then by the increasing property of $G_{n+1}(\lambda)$, for all $\lambda \geq 0$,

$$G_{n+1}(\lambda) + 1 > h(\lambda).$$

In this case, $(c_1/c_0, c_1/c_0) = \emptyset$, we let $A_n = B_n = c_1/c_0$ for convenience.

Suppose $0 < c_1 \leq 1$ instead. By Proposition 1 (iii), for all $\lambda \geq 0$,

$$G_{n+1}(\lambda) + 1 \geq G_{n+1}(0) + 1 = 1 \geq h(\lambda).$$
Hence, the optimal stopping time is $\tau^* \equiv 0$ in this case. We then let $A_n = B_n = c_1/c_0$ for convenience as well.

Therefore, under all situations, for $n = 0, 1, \cdots$,

$$g_n(\lambda_n) = \begin{cases} 
G_{n+1}(\lambda_n) + 1 & \text{if } \lambda_n \in (A_n, B_n), \\
h(\lambda_n) & \text{if } \lambda_n \not\in (A_n, B_n),
\end{cases}$$

and the optimal stopping time is

$$\tau^* = \min\{n \geq 0 : \lambda_n \not\in (A_n, B_n)\}.$$ 

Theorem 1 is proved.

Illustrated in Figure 4.1 are the relationships among $A_n$, $B_n$, the functions $h(\cdot)$, $g_n(\cdot)$ and $G_n(\cdot)$, and the decision parameters $c_0$ and $c_1$, as well as the properties of the functions shown in Proposition 1.

We call $\{A_n, B_n\}$ the optimal stopping boundaries at stage $n = 0, 1, \cdots$, and $(A_n, B_n)$ the sampling continuation interval of the optimal sequential test. If $(A_n, B_n) = \emptyset$, sampling must be stopped before or at stage $n$. In the case of $c_1 \leq 1$, that is, the loss incurred by falsely rejecting $H_1$ is not greater than the cost of taking a sample, $\tau^* \equiv 0$, that is, it is optimal not to observe at all. When sampling is stopped at stage $n$, the optimal terminal decision rule by (4.3) is

$$d^*(e^n) = \begin{cases} 
1 & \text{to accept } H_1, \text{ if } \lambda_n \leq A_n, \\
0 & \text{to accept } H_0, \text{ if } \lambda_n \geq B_n.
\end{cases}$$

In case of $\lambda_n = A_n = B_n$, decisions to accept $H_1$ or $H_0$ can be randomized in any way. For
simplicity, we let \( d^*(e^n_1) = 1 \) in this case. In summary, the optimal sequential test is

\[
S^* = \begin{cases} 
\text{stop sampling and accept } H_1 \text{ at stage } n, & \text{if } \lambda_n \leq A_n, \\
\text{continue sampling at stage } n, & \text{if } \lambda_n \in (A_n, B_n), \\
\text{stop sampling and accept } H_0 \text{ at stage } n, & \text{if } \lambda_n \geq B_n.
\end{cases}
\]

Figure 4.1: A typical illustration of \( g_n(\lambda) \) and the related functions and quantities.
4.4 Properties of the Optimal Stopping Boundaries

The optimal stopping boundaries \( \{A_n, B_n, n = 0, 1, \cdots \} \) are determined by \( c_0, c_1 \) and \( \{\mu_i, i = 1, 2, \cdots \} \). Generally, \( A_n \) and \( B_n \) vary as \( n \) changes. In this section, we study the properties of the optimal stopping boundaries, especially, how \( \{A_n\} \) and \( \{B_n\} \) vary when \( \{\mu_i\} \) exhibits certain patterns.

If for \( t_0 \geq 1, \{\mu_t, t = t_0, t_0 + 1, \cdots \} \) is a constant sequence, then \( \{g_n(\cdot), n = t_0 - 1, t_0, \cdots \} \) is a constant sequence of functions. Therefore, \( \{A_n, n = t_0 - 1, t_0, \cdots \} \) and \( \{B_n, n = t_0 - 1, t_0, \cdots \} \) are two constant sequences. When \( t_0 = 1 \), this is the special situation of Wald's SPRT.

If \( \{\mu_t, t = 1, 2, \cdots \} \) is periodically varying in \( t \), that is, there exists a positive integer \( I \) such that
\[
\mu_{t+I} = \mu_t
\]
for any \( t = 1, 2, \cdots \), then both \( \{A_n, n = 0, 1, \cdots \} \) and \( \{B_n, n = 0, 1, \cdots \} \) are periodically varying in \( n \) with the same period \( I \), that is,
\[
A_{n+I} = A_n
\]
and
\[
B_{n+I} = B_n
\]
for any \( n = 0, 1, \cdots \). This is because for any \( n \leq K \), by the definitions,
\[
g^K_n(\cdot) = g^{K+I}_n(\cdot),
\]
and then
\[
g_n(\cdot) = g_{n+I}(\cdot)
\]
for any \( n = 0, 1, \cdots \).
The following proposition shows that when \( \{\mu_i\} \) converges to 0, the optimal sequential test is *truncated*. That is, there exists \( M \), such that \( A_n = B_n = c_1/c_0 \) for all \( n \geq M \).

**Proposition 2** Assume

\[
\lim_{i \to \infty} \mu_i = 0.
\]

The optimal sequential test is truncated. That is, there exists an integer \( M \geq 0 \), such that

\[
\tau^* \leq M.
\]

**Proof of Proposition 2:**

It will be sufficient to show that under the condition of the proposition, there exists an integer \( M \), such that for any \( K \geq n \geq M + 1 \) and \( \lambda \geq 0 \),

\[
g_n^K(\lambda) = h(\lambda)
\]  

(4.13)

and

\[
|G_n^K(\lambda) - h(\lambda)| < 1.
\]  

(4.14)

For any \( \epsilon \in (0, 1) \), there exists \( \delta_\epsilon > 0 \), such that

\[
\int_{|x| \geq \delta_\epsilon} f_{N(0,1)}(x)dx \leq \frac{1}{c_1}(1 - \epsilon).
\]

For any positive integer \( K \), by the definition, for \( \lambda \geq 0 \),

\[
g_K^K(\lambda) = h(\lambda).
\]

Since \( \lim_{i \to \infty} \mu_i = 0 \), there exists \( m_\epsilon \), such that for any \( K \geq m_\epsilon \), \( |x| < \delta_\epsilon \), and \( \lambda \in (0, \frac{\epsilon}{c_0}) \),

\[
|\lambda \exp\{-\mu_K(x + \frac{\mu_K}{2})\} - \lambda| = \lambda|\exp\{-\mu_K(x + \frac{\mu_K}{2})\} - 1| 
\leq \frac{1}{c_0} \epsilon.
\]  

(4.15)
\[
\frac{d}{d\lambda} h(\lambda) = \begin{cases} 
  c_0, & \lambda \in (0, \frac{c_1}{c_0}), \\
  0, & \lambda \in (\frac{c_1}{c_0}, \infty)
\end{cases}
\]

is bounded. Therefore, for \( K \geq m_\epsilon, |x| < \delta_\epsilon, \) and \( \lambda \in (0, \frac{c_1}{c_0}), \)

\[
|h(\lambda \exp\{-\mu_K(x + \frac{\mu_K}{2})\}) - h(\lambda)| \leq c_0 \frac{\epsilon}{c_0} = \epsilon. \tag{4.16}
\]

For \( K \geq m_\epsilon \) and \( |x| < \delta_\epsilon, \)

\[
|\exp\{-\mu_K(x + \frac{\mu_K}{2})\} - 1| \leq \frac{\epsilon}{c_0} \left(\frac{c_1}{c_0}\right) = \frac{\epsilon}{c_1},
\]

and

\[
\exp\{-\mu_K(x + \frac{\mu_K}{2})\} \geq 1 - \frac{\epsilon}{c_1}.
\]

Hence, for \( K \geq m_\epsilon, |x| < \delta_\epsilon \) and \( \lambda \geq c_1/c_0, \)

\[
\lambda \exp\{-\mu_K(x + \frac{\mu_K}{2})\} \\
\geq \frac{c_1}{c_0} \left(1 - \frac{\epsilon}{c_1}\right) \\
= \frac{c_1}{c_0} - \frac{\epsilon}{c_0},
\]

and therefore,

\[
|h(\lambda \exp\{-\mu_K(x + \frac{\mu_K}{2})\}) - h(\lambda)| \\
\leq |c_0 \left(\frac{c_1}{c_0} - \frac{\epsilon}{c_0}\right) - c_1| \\
= \epsilon. \tag{4.17}
\]

The non-equality of (4.17) is because of that \( h(\cdot) \) is an increasing function and \( h(\lambda) \equiv c_1 \) for \( \lambda \geq c_1/c_0. \) Since \( h(0) = 0, \) from (4.16) and (4.17), for any \( K \geq m_\epsilon, |x| < \delta_\epsilon, \) and \( \lambda \geq 0, \)

\[
|h(\lambda \exp\{-\mu_K(x + \frac{\mu_K}{2})\}) - h(\lambda)| \leq \epsilon.
\]
Thus, for any $K \geq m_\epsilon$ and $\lambda \geq 0$,

$$|G^K_K(\lambda) - h(\lambda)|$$

$$= \left| \int_{-\infty}^{\infty} h(\lambda \exp\{ -\mu_K(x + \frac{\mu_K}{2}) \}) f_N(0,1)(x) dx - h(\lambda) \right|$$

$$\leq \int_{|x| \geq \delta_\epsilon} |h(\lambda \exp\{ -\mu_K(x + \frac{\mu_K}{2}) \}) - h(\lambda)| f_N(0,1)(x) dx$$

$$+ \int_{|x| < \delta_\epsilon} |h(\lambda \exp\{ -\mu_K(x + \frac{\mu_K}{2}) \}) - h(\lambda)| f_N(0,1)(x) dx$$

$$\leq \int_{|x| \geq \delta_\epsilon} c_1 f_N(0,1)(x) dx + \int_{|x| < \delta_\epsilon} \epsilon f_N(0,1)(x) dx$$

$$< c_1 \frac{1 - \epsilon}{c_1} + \epsilon$$

$$= 1.$$

Hence, by (4.6),

$$g^K_K(\lambda) = h(\lambda).$$

If $K - 1 \geq m_\epsilon$, then by a similar development,

$$|G^K_{K-1}(\lambda) - h(\lambda)| < 1.$$
This means that the optimal sequential test is truncated. That is, sampling must be stopped no later than \( M \).

Proposition 2 is proved.

We call such an \( M \) in Proposition 2 a truncation point. (4.15) indicates a way for finding a truncation point when \( \lim_{i \to \infty} \mu_i = 0 \). (4.15) also indicates that \( \lim_{i \to \infty} \mu_i = 0 \) is not a necessary condition for the optimal sequential test to be truncated. In fact, if \( \lim_{i \to \infty} |\mu_i| \) is sufficiently small but not zero, then the optimal sequential test can be still truncated. Proposition 2 substantially strengthens the result of Liu and Blostein (1992), where it was shown that the optimal stopping boundaries were truncated when \( \mu_t \equiv 0 \) for \( t \) beyond a certain stage \( M \).
4.5 Computation of the Optimal Stopping Boundaries

Generally, to compute $A_n$ and $B_n$, we need to compute $g_n(\cdot)$ and $G_{n+1}(\cdot)$. To compute $g_n(\cdot)$ and $G_{n+1}(\cdot)$, we can first compute $g_n^K(\cdot)$ and $G_{n+1}^K(\cdot)$ for a sufficiently large $K > n$. By the convergence properties, $g_n(\cdot)$ and $G_{n+1}(\cdot)$ can be well approximated by $g_n^K(\cdot)$ and $G_{n+1}^K(\cdot)$. Once $g_n(\cdot)$ has been computed, $g_{n-1}(\cdot), g_{n-2}(\cdot), \cdots$ can be computed backward recursively. In this way, $A_n$ and $B_n$ can be found for any $n = 0, 1, \cdots$.

Major effort is required in the computation of $G_n^K(\cdot)$ functions, that is, the integration in (4.7). The integration can be done by using Gaussian quadrature eventually. However, caution must be taken to avoid computational singularity in the integral kernels, as well as in the integral intervals. By using variable transformation, the kernel singularity can be removed, and an expression of $G_n^K(\cdot)$ suitable for computation can be obtained as follows.

$$C_n^K(\lambda) = \int_{-\infty}^{\infty} g_n^K(\lambda \exp\{-\mu_n(x + \frac{\mu_n}{2})\}) f_{N(0,1)}(x) dx$$

$$= \int_0^\infty g_n^K(y) f_{N(0,1)}(\frac{\log(\lambda/y)}{\mu_n} - \frac{\mu_n}{2} \frac{1}{\mu_n y}) dy$$

$$= \int_{B_n} c_1 f_{N(0,1)}(\frac{\log(\lambda/y)}{\mu_n} - \frac{\mu_n}{2} \frac{1}{\mu_n y}) dy$$

$$+ \int_{A_n} [G_{n+1}(y) + 1] f_{N(0,1)}(\frac{\log(\lambda/y)}{\mu_n} - \frac{\mu_n}{2} \frac{1}{\mu_n y}) dy$$

$$+ \int_0^\Lambda c_0 y f_{N(0,1)}(\frac{\log(\lambda/y)}{\mu_n} - \frac{\mu_n}{2} \frac{1}{\mu_n y}) dy$$

$$+ \int_{\log(\lambda/A_n) - \frac{\mu_n}{2}}^{\log(\lambda/B_n) - \frac{\mu_n}{2}} G_{n+1}(\lambda \exp\{-\mu_n[x + \frac{\mu_n}{2}]\}) + 1] f_{N(0,1)}(x) dx$$

$$= c_1 \Phi\left(\frac{\log(\lambda/B_n) - \frac{\mu_n}{2}}{\mu_n}\right)$$

76
\[ + c_0 \lambda \exp\left(-\frac{\mu_n^2}{2}\right) \int_{-\infty}^{\infty} \exp(-\mu_n x) f_{N(0,1)}(x) \, dx \]
\[ + \int_{-\log(\mu_n)}^{\log(\mu_n)} \left[ G'^{K}_{n+1}(\exp(-\mu_n x)) + 1\right] f_{N(0,1)}(x + \frac{\log(\lambda)}{\mu_n} - \frac{\mu_n}{2}) \, dx, \quad (4.20) \]

where \( \log(\cdot) \) is the natural logarithm function. Variable transformation was first made to utilize the properties of \( g_n^{K}(\cdot) \) in (4.18). Then, the variable was transformed back to avoid the nearly kernel singularities in (4.19).

An efficient computer algorithm for computation of the optimal stopping boundaries by using the procedure described in the beginning of this section and the expression of \( G^{K}_{n}(\lambda) \) of (4.20) has been developed. In the computer algorithm, we have used some programming techniques to avoid forward recursive calling of \( G^{K}_{n}(\cdot) \) in (4.20), which can cause exponentially increasing memory consumption and slow down computation. The computer algorithm has been successfully implemented in Turbo C++. In the following section, we will present an example of the computed optimal stopping boundaries.
4.6 Numerical Results

Consider

\[ \mu_t = \delta(t - 1) = 0.9^{t-1}, \text{ for } t \geq 1. \]

Using decision parameters \( c_0 = 10 \) and \( c_1 = 10 \), we have computed the optimal stopping boundaries that are presented in Table 4.1 and are plotted in Figure 4.2.

<table>
<thead>
<tr>
<th>( n )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>( \geq 12 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A_n )</td>
<td>0.47</td>
<td>0.53</td>
<td>0.59</td>
<td>0.65</td>
<td>0.71</td>
<td>0.76</td>
<td>0.81</td>
<td>0.85</td>
<td>0.89</td>
<td>0.92</td>
<td>0.95</td>
<td>0.98</td>
<td>1.00</td>
</tr>
<tr>
<td>( B_n )</td>
<td>4.11</td>
<td>3.04</td>
<td>2.39</td>
<td>1.97</td>
<td>1.68</td>
<td>1.49</td>
<td>1.34</td>
<td>1.24</td>
<td>1.16</td>
<td>1.11</td>
<td>1.06</td>
<td>1.03</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 4.1: The optimal stopping boundaries when \( \delta(i) = 0.9^i \) for \( i \geq 0 \).

Figure 4.2: The optimal stopping boundaries of the example.
A distinct feature of the optimal stopping boundaries of the example is shown in Figure 4.2. When \( \mu_i = \delta(t - 1) \) decreases to 0 for \( t \geq 1 \), the continuation region, where the sampling of the sequential test is continued, is bounded. That is, the optimal stopping boundaries are truncated, and the test must be stopped before or at the truncation point, which is 12 in this example.
Chapter 5

Likelihood Ratio Testing Procedures
5.1 Introduction

For change detection in autocorrelated processes, there have been two sets of tools. The first set is the control chart procedures. We have studied the performance of the CUSUM, EWMA and Shewhart control chart procedures for change detection in autocorrelated processes. Now, we examine their alternatives, the second set of tools of LR testing procedures.

A classical LR testing procedure signals occurrence of a change when the maximum of the likelihood ratio over unknown parameters, such as the change point, exceeds a constant threshold value. Basseville and Nikiforov (1993) described a unified likelihood ratio framework for design and performance analysis of change detection procedures, and studied the LR testing procedure in details. Because of the fine intuition behind them and their ability to deal with complex situations, LR testing procedures have been widely used for change detection in autocorrelated processes. When the magnitude of change is unknown, the LR testing procedure is a generalized likelihood ratio testing procedure proposed by Willsky and Jones (1976) and studied recently by Lai (1995) and Siegmund and Venkatraman (1995). When the change point is the only unknown parameter, the LR testing procedure is a likelihood ratio CUSUM procedure by using residuals, and it is optimal under a minimax criterion in the i.i.d. case. However, the optimality may not hold in general, and the performance of the LR testing procedure in comparison with the control chart procedures is unknown.

In this chapter, we will study the structure of the LR testing procedure, extend the LR testing procedure, propose a combined CUSUM and Shewhart control chart procedure, derive integral equations for computation of their ARLs in some cases, and study their performance.
5.2 The Structure of LR Testing Procedures

Consider the change detection problem on the residual process \( \{e_t, t = 1, 2, \cdots\} \). Assume
\( e_t \) has a continuous distribution, and let \( f_\mu(\cdot) \) be the probability density function of \( e_t \) if
\( E(e_t) = \mu \). Because of the mutual independence, the joint density function of \( e_t^i \) is
\[
\prod_{i=1}^{t} f_0(e_i) \quad (5.1)
\]
under \( H_0 \), and
\[
\left[ \prod_{i=1}^{r-1} f_0(e_i) \right] \left[ \prod_{i=r}^{t} f_{\delta(i-r)}(e_i) \right] \quad (5.2)
\]
under \( H_1 \). The likelihood ratio is
\[
\prod_{i=r}^{t} \frac{f_{\delta(i-r)}(e_i)}{f_0(e_i)}, \quad (5.3)
\]
and the LR testing procedure is a stopping time
\[
N_{LR} = \inf\{t \geq 1 : \max_{1 \leq r \leq t} \sum_{i=r}^{t} \log \frac{f_{\delta(i-r)}(e_i)}{f_0(e_i)} \geq h\}, \quad (5.4)
\]
where \( h \) is a constant threshold value.

As we have discussed in the last part of § 2.3, the joint densities of \( e_t^i \) (5.1) under \( H_0 \) and
(5.2) under \( H_1 \) are equal to the joint densities of \( Y_t^i \) under \( H_0 \) and \( H_1 \), respectively. Therefore,
the likelihood ratio (5.3) based on the residuals \( e_t^i \) is equal to the likelihood ratio based on
the observations \( y_t^i \), and the LR testing procedure (5.4) based on the residuals has the same
detection properties as that based on \( y_t^i \). However, because of the mutual independence, the
LR testing procedure based on residuals has a simpler expression than its counterpart based
on observations for an autocorrelated process.

The stopping time \( N_{LR} \) can be rewritten in the following form.
\[
N_{LR} = \min_{r \geq 1}(N_{LR}^r + r - 1), \quad (5.5)
\]
where

\[ N_{LR} = \inf \{ n \geq 1 : \sum_{i=r}^{n-1} \log \frac{f_{\theta_i}(e_i)}{f_0(x_i)} \geq h \} \]

\[ = \inf \{ n \geq 1 : \sum_{i=0}^{n-1} \log \frac{f_{\theta_i}(e_{r+1})}{f_0(x_{r+1})} \geq h \}, \]

for \( r = 1, 2, \ldots \). \( N_{LR}^1 \) is the stopping rule of a one-side SPRT on the mean level hypotheses presented in § 4.1, and \( N_{LR}^r, r = 2, 3, \ldots \), is the same stopping rule applied to \( \{e_r, e_{r+1}, \ldots \} \).

(5.5) suggests that in the LR testing procedure, we actually activate a common stopping rule at each possible change point \( r = 1, 2, \ldots \), and whenever one of the activated tests signals that \( E(e_i) \) has started to change, the whole LR testing procedure \( N_{LR} \) is stopped, and \( H_1 : r \leq t \) in the change point hypotheses is accepted.

Lorden (1971) pointed out the relationship (5.5) between the likelihood ratio CUSUM procedure and the one-side SPRT. By using it, he obtained the asymptotic optimality of the likelihood ratio CUSUM procedure in the i.i.d. case. The relationship (5.5) has also been used for design and analysis of change detection procedures by Basseville and Nikiforov (1993) and Lai (1995).
5.3 Extended LR Testing Procedures

In the i.i.d. case, that is, when $\delta(i)$ is a constant for $i \geq 0$, SPRT is optimal for testing the change profile hypotheses, and the likelihood ratio CUSUM procedure (5.4) is optimal for change detection – testing the change point hypotheses. In both SPRT and the classical LR testing procedure (5.4), their threshold values are a constant. When $\delta(i)$ is generally time-varying for $i \geq 0$, we have shown in Chapter 4 that the optimal sequential test on the mean level hypotheses is a likelihood ratio test with time-varying thresholds. These let us propose the following extended LR testing procedure by replacing the constant threshold in (5.4) with time-varying thresholds.

$$N_{ELR} = \inf \{ t \geq 1 : \max_{1 \leq r \leq t} \left[ \sum_{i=r}^{t} \log \frac{f_{\delta(i-r)}(e_i)}{f_0(e_i)} - h_{t-r} \right] \geq 0 \}, \quad (5.6)$$

where $\{h_i, i = 0, 1, \cdots \}$ is a time-varying sequence.

The classical LR testing procedure $N_{LR}$ can be rewritten as

$$N_{LR} = \inf \{ t \geq 1 : \max_{1 \leq r \leq t} \left[ \sum_{i=r}^{t} \log \frac{f_{\delta(i-r)}(e_i)}{f_0(e_i)} - h \right] \geq 0 \}.$$

For a $\delta(\cdot)$, let $\{N_{ELR}\}$ denote the class of all extended LR testing procedures. Clearly,

$$N_{LR} \in \{N_{ELR}\}.$$

Therefore, the best change detection procedure from the class $\{N_{ELR}\}$ may improve the performance of the classical LR testing procedure $N_{LR}$.

To find the optimal threshold sequence $\{h_i, i = 0, 1, \cdots \}$ for the extended LR testing procedure (5.6) needs some effort. However, we can let

$$h_i = -\log A_{i+1}, \quad \text{for} \quad i = 0, 1, \cdots, \quad (5.7)$$
where \( \{A_i, i = 1, 2, \cdots \} \) is the lower threshold sequence of the optimal sequential likelihood ratio test \( \tau^* \) on the mean level hypotheses studied in Chapter 4, which takes at least one observation.

By doing the value assignment (5.7), we actually make

\[
\tilde{N}_{ELR} = \min_{r \geq 1} (\tilde{N}_{ELR}^r + r - 1),
\]

(5.8)

where \( \tilde{N}_{ELR} \) is the one-side optimal sequential test \( \tau^* \) but applied to \( e_r, e_{r+1}, \cdots \), and \( \tilde{N}_{ELR} \) is the extended LR testing procedure with the threshold sequence \( \{-\log A_{i+1}, i = 0, 1, \cdots \} \),

\[
\tilde{N}_{ELR} = \inf\{t \geq 1 : \max_{1 \leq r \leq t} \left[ \sum_{i=r}^{t} \log \frac{f_\delta(i-r)(e_i)}{f_\delta(e_i)} + \log A_{t-r+1} \right] \geq 0\}.
\]

We expect that the extended LR testing procedure \( \tilde{N}_{ELR} \) of (5.8) performs better than the classical LR testing procedure \( N_{LR} \) of (5.5).

In Chapter 4, we have studied the structure of \( \tau^* \) and obtained some properties of \( \{A_i, i = 0, 1, \cdots \} \). When \( \delta(i) \) converges to zero for \( i = 0, 1, \cdots \), for example, when \( \{e_i\} \) are the residuals of an \( IMA(1,1) \) process subject to an observational additive step shift, the optimal sequential test \( \tau^* \) on the mean level hypotheses is truncated and takes at most \( M \) observations. Therefore, the extended LR testing procedure \( \tilde{N}_{ELR} \) is

\[
\tilde{N}_{ELR} = \inf\{t \geq 1 : \max_{t-M+1 \leq r \leq t} \left[ \sum_{i=r}^{t} \log \frac{f_\delta(i-r)(e_i)}{f_\delta(e_i)} + \log A_{t-r+1} \right] \geq 0\},
\]

(5.9)

which turns out to be a window-limited procedure.

When \( \{e_i\} \) are the residuals of an \( AR(p) \) process subject to an observational additive step shift, \( \delta(i) \) varies at the first \( p \) time points \( i = 0, 1, \cdots, p-1 \), and is a constant for \( i = p, p+1, \cdots \). In this case, we have shown that \( A_i \) is a constant for \( i = p, p+1, \cdots \). Therefore, in \( \tilde{N}^* \), the threshold \( h_i \) is a constant for \( i = p-1, p, p+1, \cdots \), and varies only at the first \( p-1 \) time points \( i = 0, 1, \cdots, p-2 \).
5.4 A Combined CUSUM and Shewhart Control Chart Procedure

To approximate the best change detection procedure from the class \( \{N_{ELR}\} \), we propose the following combined CUSUM and Shewhart (CCS) control chart procedure.

\[
N_{CCS} = \inf\{t \geq 1 : S_H(t) \geq h \text{ or } |e_t| \geq C\},
\]

where

\[
S_H(t) = \max\{0, S_H(t-1) + (e_t - k)\}
\]

is the usual higher-side CUSUM statistic, \( k > 0 \) is the CUSUM reference value, \( h > 0 \) is the CUSUM threshold value, \( C > 0 \) is the Shewhart control limit. Note that we assume

\[ C > k. \]

Otherwise, the CUSUM part will be dominated by the Shewhart part.

CCS procedures first appeared in Westgard, Groth, Aronsson and de Verdier (1977), and were studied by Lucas (1982) for detection of step shifts of unknown size in the i.i.d. case. Shewhart chart was capable of detecting changes of large size, CUSUM was superior in detecting changes of small size, and their combination combined the strength of both.

Our intention here is to approximate the time-varying threshold sequence of the extended LR testing procedure by adding the Shewhart control limit to the threshold of CUSUM. In comparison with the LR or extended LR testing procedures, the structure of the CCS procedure is relatively simple. The CCS procedure is non-parametric, that is, its composition (not performance) is independent of process model parameters. For a general \( \delta(\cdot) \), we can establish integral equations to solve the ARLs of the CCS procedure.
5.5 Performance Analysis

In this section, under the ARL criterion, we will investigate the performance of three change detection procedures: the LR testing procedure applied to the residuals of an $AR(1)$ process for detection of an observational additive step shift, the extended LR testing procedure applied to the residuals of an $IMA(1,1)$ process for detection of an observational additive step shift, and the CCS procedure. For these procedures, because of the Markov chain nature of their testing statistics, we can manage to establish integral equations for their ARLs, and solve them accurately and precisely. The involved Markov chains will be non-homogeneous.

Singularity, primarily the discontinuity of the integral kernels, will be present in the integral equations of the three procedures. This imposes much difficulty in solving the integral equations, care and effort are needed to handle the singularity in order to design valid and efficient computer algorithms. Gaussian quadrature for solving standard Fredholm equations of the second kind cannot be applied due to the singularity. A technique we will use to overcome the singularity is to construct quadrature on a uniform grid by using cubic spline approximation of the unknown function. Once the singularity is handled, similar recursive computational procedures can be developed to solve the ARL values as we did for CUSUM and EWMA in Chapter 3.

In the following subsections, we will describe the three change detection procedures, derive the integral equations, and present numerical results. All the computation has been successfully implemented in Turbo C++. As in § 3.2, for performance analysis of a change detection procedure $N$, we will be only interested in $E_{r=0}(N)$ and $E_{r=1}(N)$, and we can assume that

- $r = 1$.

We will assume that
• $e_t$ has a normal distribution with standard deviation $\sigma = 1$.

### 5.5.1 The LR Testing Procedure Applied to $AR(1)$ Processes

For an $AR(1)$ process with the autoregressive parameter $a_1$ subject to an observational additive step shift, the mean of residual after change is

$$
\delta(i) = \begin{cases} 
\Delta, & i = 0, \\
(1 - a_1)\Delta, & i \geq 1,
\end{cases}
$$

(5.10)

and the threshold in (5.7) is a constant for all $i \geq 0$. Hence, the extended LR testing procedure $\tilde{N}_{ELR}$ in (5.8) reduces to the usual LR testing procedure with a constant threshold.

Let $\Delta_0 > 0$ be the size of an observational additive step shift we intend to detect, $\mu_0 = \Delta_0$ and $\mu_1 = (1 - a_1)\Delta_0$. For $t \geq 1$, let

$$S_t = \max\{\mu_0(e_t - \frac{\mu_0}{2}), S_{t-1} + \mu_1(e_t - \frac{\mu_1}{2})\}.$$

$$\{S_t, t = 1,2,\ldots\}$$

is a non-homogeneous Markov chain. Let $h > 0$ be a constant threshold value. Let $S_0 = 0$, then

$$N_{LR} = \tilde{N}_{ELR} = \inf\{t \geq 1 : S_t \geq h\}.$$

For $i \geq 0$, define $ARL(s,i)$ the ARL of the detection procedure $\tilde{N}_{ELR}$ applied to $\{e_{i+1}, e_{i+2}, \ldots\}$ given that $S_i = s$. Obviously, $s < h$. Then, for $S_0 = 0$,

$$E_{r=1}(N_{LR}) = ARL(0,0).$$

For $s < h$, we can derive an integral equation suitable for computation and computer implementation as follows.

$$ARL(s, i)$$
\[ P(S_{i+1} \geq h) = 1 + \int_{\{ e_{i+1}, S_{i+1} < h \}} [1 + ARL(S_{i+1}, i + 1)] f_N(\delta(i), 1)(e_{i+1}) \, de_{i+1} \]

\[ = 1 + \int_{\{ e_{i+1}, S_{i+1} < h \}} ARL(S_{i+1}, i + 1) f_N(\delta(i), 1)(e_{i+1}) \, de_{i+1} \]

\[ = 1 + \int_{\{ e_{i+1}, S_{i+1} < h, \mu_0(e_{i+1} - \frac{\mu_0}{2}) \geq s + \mu_1 (e_{i+1} - \frac{\mu_1}{2}) \}} ARL(S_{i+1}, i + 1) f_N(\delta(i), 1)(e_{i+1}) \, de_{i+1} \]

\[ + \int_{\{ e_{i+1}, S_{i+1} < h, \mu_0(e_{i+1} - \frac{\mu_0}{2}) \geq s + \mu_1 (e_{i+1} - \frac{\mu_1}{2}) \}} ARL(S_{i+1}, i + 1) f_N(\delta(i), 1)(e_{i+1}) \, de_{i+1} \]

\[ = 1 + \int_{\{ e_{i+1}, \mu_0(e_{i+1} - \frac{\mu_0}{2}) \geq s + \mu_1 (e_{i+1} - \frac{\mu_1}{2}) \}} ARL(\mu_0(e_{i+1} - \frac{\mu_0}{2}), i + 1) f_N(\delta(i), 1)(e_{i+1}) \, de_{i+1} \]

\[ + \int_{\{ e_{i+1}, \mu_0(e_{i+1} - \frac{\mu_0}{2}) \geq s + \mu_1 (e_{i+1} - \frac{\mu_1}{2}) \}} ARL(\mu_0(e_{i+1} - \frac{\mu_0}{2}), i + 1) f_N(\delta(i), 1)(e_{i+1}) \, de_{i+1} \]

\[ = 1 + \int_{\{ x, x > \frac{\mu_0 - \mu_1}{\mu_0 - \mu_1} \}} ARL(x, i + 1) f_N(\mu_0(\delta(i) - \frac{\mu_0}{2}), \mu_0)(x) \, dx \]

\[ + \int_{\{ x, x < \frac{\mu_0 - \mu_1}{\mu_0 - \mu_1} \}} ARL(x, i + 1) f_N(s + \mu_1(\delta(i) - \frac{\mu_1}{2}), \mu_1)(x) \, dx \]

\[ = 1 + \int_{-\infty}^{h} ARL(x, i + 1) f_N(\mu_0(\delta(i) - \frac{\mu_0}{2}), \mu_0)(x) I_{\{ x \geq \frac{\mu_0 - \mu_1}{\mu_0 - \mu_1} \}} \, dx \]

\[ + \int_{-\infty}^{h} ARL(x, i + 1) f_N(s + \mu_1(\delta(i) - \frac{\mu_1}{2}), \mu_1)(x) I_{\{ x < \frac{\mu_0 - \mu_1}{\mu_0 - \mu_1} \}} \, dx, \quad (5.11) \]

where \( f_N(\mu, \sigma^2)(\cdot) \) is the \( N(\mu, \sigma^2) \) density function, and \( I_{set} \) is the indicator function. Clearly, the integral kernel is discontinuous, and therefore, singular. Besides, there is another singularity, the infinite integral range. To solve this integral equation, special techniques and major effort more than those for solving the integral equations in § 3.2 are required.
procedures applied to the residuals of an AR(1) process with autoregressive coefficient \( a_1 = 0.45 \) subject to an observational additive step shift of size \( \Delta \) when their in-control ARLs are 100. ARLs are optimized at \( \Delta = 1 \) in the upper panel, and at \( \Delta = 3 \) in the lower panel.

For the fixed in-control ARL of 100 and \( a_1 = 0.45 \), Table 5.1 presents the ARLs, \( E_{r=1}(N) \), of the LR procedure in comparison to those of CUSUM and EWMA, and Figure 5.1 shows the ARL curves.

Figure 5.1 shows that when the mean of residual after change is (5.10), the LR testing procedure performs roughly the same as the CUSUM control chart procedure, and both of them perform better than the EWMA control chart procedure.
Figure 5.1: ARLs of the LR testing procedure and the CUSUM and EWMA control chart procedures applied to the residuals of an $AR(1)$ process with autoregressive coefficient $a_1 = 0.45$ subject to an observational additive step shift of size $\Delta$. For the fixed in-control ARL, the chosen chart parameters optimize the ARL at $\Delta = 1$ in the upper panel and at $\Delta = 3$ in the lower panel.
5.5.2 The Extended LR Procedure Applied to $IMA(1,1)$ Processes

For an $IMA(1,1)$ process subject to an observational additive step shift, the mean of residual after change is decaying exponentially fast:

$$\delta(i) = \beta_1^i \Delta$$

(5.12)

for $i \geq 0$. We will study the situations where $|\beta_1|$ is small so that the optimal sequential test $\tau^*$ on the mean level hypotheses is truncated at two observations. Therefore, the window size of $\tilde{N}_{ELR}$ in (5.9) is $M = 2$ and integral equations for computation of ARLs can be established and solved.

Let $\Delta_0 > 0$ be the size of an observational additive step shift we intend to detect, $\mu_0 = \Delta_0$ and $\mu_1 = \beta_1 \Delta_0$. Let $h_0$ and $h_1 > 0$ be two threshold values.

$$\tilde{N}_{ELR} = \inf\{t \geq 1 : \max_{t-1 \leq r \leq t} \log \frac{f_{\mu_1} (e_t)}{f_{\mu_0} (e_t)} - h_{t-r} \geq 0\}$$

$$= \inf\{t \geq 1 : \log \frac{f_{\mu_0} (e_t)}{f_{\mu_0} (e_{t-1})} \geq h_0 \text{ or } \log \frac{f_{\mu_0} (e_{t-1}) + \log f_{\mu_1} (e_t)}{f_{\mu_0} (e_t)} \geq h_1\}$$

$$= \inf\{t \geq 1 : \mu_0 (e_t - \frac{\mu_0}{2}) \geq h_0 \text{ or } \mu_0 (e_{t-1} - \frac{\mu_0}{2}) + \mu_1 (e_t - \frac{\mu_1}{2}) \geq h_1\}. $$

We assume that $e_0 = 0$.

$$\{(e_{t-1}, e_t), t = 1, 2, \cdots\}$$

is a non-homogeneous Markov chain.

For $i \geq 0$, define $ARL(e, i)$ the ARL of the extended LR procedure $\tilde{N}_{ELR}$ applied to $\{e_{i+1}, e_{i+2}, \cdots\}$ given that $e_i = e$. For $e_0 = 0$,

$$E_{\tau=1} (\tilde{N}_{ELR}) = ARL(0, 0).$$

We can derive the following integral equation.

$$ARL(e, i)$$
\[
= 1P(\mu_0(e_{i+1} - \frac{\mu_0}{2})) \geq h_0, \text{ or } \mu_0(e - \frac{\mu_0}{2}) + \mu_1(e_{i+1} - \frac{\mu_1}{2} \geq h_1)
\]
\[
+ \int_{\{e_{i+1} : \mu_0(\mu_1 + e_{i+1} - \frac{\mu_1}{2}) + \mu_1(\mu_1 - \frac{\mu_1}{2} < h_1)\}} [1 + ARL(e_{i+1}, i + 1)] f_N(\delta(i), 1)(e_{i+1}) de_{i+1}
\]
\[
= 1 + \int_{\{e_{i+1} : \mu_0(\mu_1 + e_{i+1} - \frac{\mu_1}{2}) + \mu_1(\mu_1 - \frac{\mu_1}{2} + \frac{\mu_2}{2}) \}} ARL(e_{i+1}, i + 1) f_N(\delta(i), 1)(e_{i+1}) de_{i+1}
\]
\[
= 1 + \int_{-\infty}^{\frac{b_0}{\mu_0} + \frac{\mu_0}{2}} ARL(x, i + 1)[f_N(\delta(i), 1)(x) I_{\{x < \frac{\mu_0}{\mu_1} e + \frac{\mu_1}{2} + \frac{\mu_2}{2}\}}] dx,
\]
(5.13)

for \( e < \frac{b_0}{\mu_0} + \frac{\mu_0}{2} \). The same kinds of singularity as in (5.11) are present in (5.13).

For the fixed in-control ARL of 100 and \( b_1 = 0.1 \), Table 5.2 presents the ARLs of the extended LR detection procedure in comparison with the classical LR detection procedure and the CUSUM and EWMA control chart procedures and Figure 5.2 shows the ARL curves.

Figure 5.2 shows that when the mean of residual after change is (5.12), the extended LR procedure improves the performance of both CUSUM and the classical LR detection procedure slightly. For detection of this rapidly decreasing change, the EWMA control chart procedure offers the best performance.
$\Delta = 2.4$ XLR ($h_1 = 1.77$)

$\Delta = 4.2$ XLR ($h_1 = 2.29$)

Table 5.2: ARLs of the extended LR (XLR) testing procedure in comparison with the classical LR testing procedure and the CUSUM and EWMA control chart procedures applied to the residuals of an $IMA(1,1)$ process with moving average coefficient $b_1 = 0.1$ subject to an observational additive step shift of size $\Delta$ when their in-control ARLs are 100. ARLs are optimized at $\Delta = 1$ in the upper panel, and at $\Delta = 3$ in the lower panel.
Figure 5.2: ARLs of the extended LR (XLR) testing procedure in comparison with the classical LR testing procedure and the CUSUM and EWMA control chart procedures applied to the residuals of an \(IMA(1,1)\) process with moving average coefficient \(b_1 = 0.1\) subject to an observational additive step shift of size \(\Delta\). For the fixed in-control ARL, the chosen chart parameters optimize the ARL at \(\Delta = 1\) in the upper panel and at \(\Delta = 3\) in the lower panel.
5.5.3 The CCS Procedure

For a general $\delta(\cdot)$,

\[
\{(S_H(t), e_i), t = 1, 2, \ldots\}
\]

is a non-homogeneous Markov chain.

For $i \geq 0$, define $ARL(s, i)$ the ARL of the CCS procedure applied to $\{e_{i+1}, e_{i+2}, \ldots\}$ given that $S_H(i) = s$. For $S_H(0) = 0$,

\[
E_{r=1}(N_{CCS}) = ARL(0, 0).
\]

For $0 < s < h$, we can derive the following integral equations for computation.

\[
ARL(s, i)
\]

\[
= 1P(|e_{i+1}| \geq C \text{ or } s + (e_{i+1} - k) \geq h)
\]

\[
+ \int_{\{e_{i+1} : |e_{i+1}| < C, s + (e_{i+1} - k) < h\}} [1 + ARL(s + (e_{i+1} - k), i + 1)] f_N(\delta(i), 1)(e_{i+1}) de_{i+1}
\]

\[
+ \int_{\{e_{i+1} : |e_{i+1}| < C, s + (e_{i+1} - k) \leq 0\}} [1 + ARL(0, i + 1)] f_N(\delta(i), 1)(e_{i+1}) de_{i+1}
\]

\[
= 1 + \int_0^h ARL(x, i + 1)[f_N(s - k + \delta(i), 1)(x) I_{s - k - C < x < s - k + C}] dx
\]

\[
+ ARL(0, i + 1)[\Phi(-\delta(i) + \min\{C, k - s\}) - \Phi(-\delta(i) - C)]
\]

\[
= 1 + \int_0^h ARL(x, i + 1)[f_N(s - k + \delta(i), 1)(x) I_{s - k - C < x < s - k + C}] dx
\]

\[
+ ARL(0, i + 1)[\Phi(-\delta(i) + k - s) - \Phi(-\delta(i) - C)]
\]

Therefore,

\[
ARL(0, \infty) = \frac{1 + \int_0^h ARL(x, \infty) f_N(-k + \delta(\infty), 1)(x) I_{s - k - C < s < s + k + C} dx}{1 - \Phi(-\delta(\infty) + k) + \Phi(-\delta(\infty) - C)},
\]

and

\[
ARL(s, \infty)
\]
for the purpose of initialization. Singularity of discontinuous integral kernels also appears in the above integral equations.

\[
\begin{align*}
&= 1 + \frac{\Phi(-\delta(\infty) + k - s) - \Phi(-\delta(\infty) - C)}{1 - \Phi(-\delta(\infty) + k) + \Phi(-\delta(\infty) - C)} \\
&\quad + \int_0^h ARL(x, \infty)|f_{N(s-k+\delta(\infty),1)}(x)|I_{\{s-k-C<x<s-k+C\}} \Phi(-\delta(\infty) + k - s) - \Phi(-\delta(\infty) - C) \\
&\quad \times \frac{f_{N(-k+\delta(\infty),1)}(x)I_{\{x<s-k+C\}}}{1 - \Phi(-\delta(\infty) + k) + \Phi(-\delta(\infty) - C)} dx
\end{align*}
\]

Table 5.3: ARLs of a CCS procedure in comparison with the LR testing procedure and the CUSUM control chart procedure applied to the residuals of an AR(1) process with autoregressive coefficient \(a_1 = 0.45\) subject to an observational additive step shift of size \(\Delta\) when their in-control ARLs are 100.

<table>
<thead>
<tr>
<th>(\Delta)</th>
<th>0</th>
<th>0.5</th>
<th>1</th>
<th>1.5</th>
<th>2</th>
<th>2.5</th>
<th>3</th>
<th>3.5</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCS</td>
<td>(C = 3.80)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(k = 0.28)</td>
<td>(h = 4.17)</td>
<td>100</td>
<td>28.58</td>
<td>12.65</td>
<td>7.31</td>
<td>4.87</td>
<td>3.50</td>
<td>2.62</td>
</tr>
<tr>
<td>LR</td>
<td>(h = 2.72)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>29.04</td>
<td>12.66</td>
<td>7.07</td>
<td>4.46</td>
<td>3.01</td>
<td>2.13</td>
<td>1.59</td>
<td>1.28</td>
</tr>
<tr>
<td>CUSUM</td>
<td>(k = 0.28)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>28.47</td>
<td>12.62</td>
<td>7.31</td>
<td>4.90</td>
<td>3.59</td>
<td>2.78</td>
<td>2.24</td>
<td>1.84</td>
</tr>
</tbody>
</table>

A recursive computational procedure similar to that for CUSUM chart in § 3.2.2 has been developed. As an example, we compute the ARL values of the CCS procedure and compare its performance with the CUSUM and the LR testing procedure when the mean of residual after change is (5.10). For the fixed in-control ARL of 100 and \(a_1 = 0.45\), Table 5.3 presents the computed ARL values, \(E_{\tau=1}(N)\), and Figure 5.3 shows the ARL curves. The CCS procedure performs roughly the same as CUSUM and the LR procedure, and it performs a little better.
than CUSUM for detection of larger size of changes.

---

**Figure 5.3:** ARLs of a CCS detection procedure in comparison with the classical LR testing procedure and the CUSUM and EWMA control chart procedures applied to the residuals of an AR(1) process with autoregressive coefficient $a_1 = 0.45$ subject to an observational additive step shift of size $\Delta$. 
5.6 Summary of the Performance Analyses

Combining the performance analyses in Chapter 3 and this chapter, we have the following results on the performance of the control chart procedures, the classical and extended LR testing procedures, and the CCS procedure for detection of an additive change in an autocorrelated process.

1. The classical and extended LR testing procedures and the CCS procedure roughly perform as well as CUSUM. The LR testing procedure shows slightly better performance than CUSUM for detection of a large size of observational additive step shift in an AR(1) process. The extended LR testing procedure can improve the performance of the classical one.

2. When the mean of residual after change decreases to 0 rapidly, EWMA roughly performs the best among all the considered change detection procedures.

3. Otherwise, CUSUM roughly performs the best.

4. Generally, Shewhart performs well for detection of very large size of changes.
Chapter 6

Conclusion and Future Research
6.1 Contributions

This thesis has studied several issues on change detection in autocorrelated processes. The contributions and major findings of this thesis are as follows.

We have systematically investigated the properties of process residuals for change detection. We have shown that process residuals are sufficient, and equivalent to process observations, for the problem of change detection. Therefore, change detection can be done by using process residuals. We have shown that process residuals are mutually uncorrelated with zero means when the process is in-control. We have developed a general procedure for specifically deriving the forms of process residuals. Using the procedure, we have derived the forms of residuals of general ARIMA processes and state space models, and studied the properties of the residuals under situations of additive changes, nonadditive changes, steady state, and nonsteady state. Under the Gaussian assumption, for an ARIMA process or a steady-state state space model subject to an additive change, the residuals are i.i.d. with zero means before the occurrence of change, and are still mutually independent with the same variance as before but with time-varying means after the occurrence of change. For an ARIMA process or a steady-state state model subject to an additive change, we have found that the properties of residuals are independent of the feedback control applied to the process. The independence of feedback control of process residuals makes the combination of process control and process monitoring (change detection) easier, because change detection procedures based on residuals can be designed independently of the feedback control policies.

Because of the properties of process residuals, detection of an additive change in an ARIMA process or a steady-state state space model is transformed to detection of an additive change in the residual process with detection properties preserved. We have studied the performance of
the CUSUM, EWMA and Shewhart control chart procedures applied to the residuals under the ARL criterion. For computation of the ARLs of the control chart procedures, we have derived an explicit formula for Shewhart, established integral equations for CUSUM and EWMA, and developed numerical procedures for solving the integral equations. We have investigated three typical situations to study the performance of the control chart procedures applied to the residuals.

We have also studied the LR testing procedure applied to residuals for detection of an additive change in an autocorrelated process. We have extended the classical LR testing procedure by replacing its constant threshold with a time-varying threshold sequence. We have proposed a combined CUSUM and Shewhart procedure for approximation of the extended LR testing procedure. We have developed numerical procedures based on integral equations for computation of the ARLs of these detection procedures, and numerically investigated their performance in some cases.

We have found that the LR testing procedures perform roughly as well as the CUSUM control chart procedure. The extended LR testing procedure can improve the performance of the classical one. When the mean of residual after change decreases rapidly to zero, the EWMA control chart procedure roughly performs the best among all considered change detection procedures. Otherwise, CUSUM roughly performs the best. Shewhart performs well only for detection of very large size of changes.

We have studied the problem of optimal sequential testing on process mean levels. The problem is a special situation of change detection, is a generalization of Wald’s SPRT problem, and has wide applications in signal processing and other areas. We have formulated the problem as an optimal stopping problem, derived the optimal stopping rule, obtained some
important properties of the optimal stopping boundaries, which are two sequence of time-
varrying threshold values, developed an algorithm for computation of the optimal stopping
boundaries, and presented a numerical example. When the mean of residual after change
decreases to zero (or a sufficiently small value), the optimal stopping boundaries are truncated.
That is, to be cost-efficient, the sampling and the test must be stopped before or at the
truncation point. When the mean of residual after change is stabilized at a constant level, the
optimal stopping boundaries are also stabilized. When the mean of residual after change is
periodically varying, the optimal stopping boundaries are also periodically varying with the
same period.

6.2 Future Research

Future research in the area of change detection in autocorrelated processes can be done in the
following directions.

- Detection of nonadditive changes.
- Change detection in multivariate processes.
- Combination of detection and estimation of the change point in the change detection
  problem.
Bibliography


