Chart Allocation and Control Techniques for Multistage and Run-to-Run Processes

by

JIN, Ming

A Thesis Submitted to
The Hong Kong University of Science and Technology
in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy
in Industrial Engineering and Engineering Management

June 2008, Hong Kong
Authorization

I hereby declare that I am the sole author of the thesis.

I authorize the Hong Kong University of Science and Technology to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize the Hong Kong University of Science and Technology to reproduce the thesis by photocopying or by other means, in total or in part, at the request of other institution or individuals for the purpose of scholarly research.

JIN, Ming
Chart Allocation and Control Techniques for Multistage and Run-to-Run Processes

by

JIN, Ming

This is to certify that I have examined the above PhD thesis and have found that it is complete and satisfactory in all respects, and that any and all revisions required by the thesis examination committee have been made.

Dr. Fugee Tsung
Supervisor

Professor Chung-Yee Lee
Head

Department of Industrial Engineering and Logistics Management

23 June 2008
Acknowledgements

My foremost thank goes to my advisor, Professor Fugee Tsung. I thank him for his insights and suggestions that helped to shape my research skills, and for his thoughtful and considerate advices that inspired my life. His visionary thoughts and energetic working style have influenced me greatly.

I also thank the thesis committee members, Professor Kwai-Sang Chin at the City University of Hong Kong, Professor Andrew Lim and Professor Richard So in the Department of Industrial Engineering and Logistics Management at the Hong Kong University of Science and Technology, and Prof. Dit-Yan YEUNG in the Department of Computer Science and Engineering at the Hong Kong University of Science and Technology, for their helpful discussions and valuable comments to the thesis, which has helped improve the quality of this work greatly.

I am also grateful to all the faculty members in the department of Industrial Engineering and Logistics Management at The Hong Kong University of Science. Special thanks to Dr. Kaijie Zhu, Professor Mitchell Tseng and Professor Chung-Yee Lee. I have learned from them both knowledge and the right attitude toward work.

I would like to thank my classmates and friends in Hong Kong. Although I cannot list all their names here, their help made my study here more colorful.

Words cannot adequately describe the support I received from my family; the understanding and selfless support from my parents and my girlfriend, Xin DU, made it possible for me to pursue my dream. I am deeply indebted to all of them.
TABLE OF CONTENTS

Title Page ....................................................................................................................... i
Authorization Page ........................................................................................................ ii
Signature Page ............................................................................................................. iii
Acknowledgements ...................................................................................................... iv
Table of Contents .......................................................................................................... v
List of Figures ............................................................................................................ viii
List of Tables ................................................................................................................ x
Acronyms ..................................................................................................................... xi
Abstract ...................................................................................................................... xii

CHAPTER 1 INTRODUCTION .................................................................................... 1
1.1 Motivation .............................................................................................................. 1
   1.1.1 Quality and statistical process control .......................................................... 1
   1.1.2 Multistage manufacturing processes (MMP) in modern industries ............ 3
   1.1.3 Run-to-run control for processes with measurement delays .................... 6
1.2 Literature review ................................................................................................... 9
   1.2.1 A review of conventional SPC techniques ................................................. 9
   1.2.2 Conventional monitoring of multistage processes ....................................... 12
   1.2.3 EWMA Run-to-run control schemes .......................................................... 14
1.3 Organization of dissertation .................................................................................. 16

CHAPTER 2 CHART ALLOCATION STRATEGY FOR SERIAL MMP .... 18
2.1 Multistage process modeling and charting methods ............................................ 19
   2.1.1 Multistage modeling .................................................................................. 19
   2.1.2 Charting methods ...................................................................................... 20
2.2 A chart allocation strategy for output monitoring ............................................... 21
2.3 A Chart allocation strategy for residual monitoring .......................................... 24
2.4 Illustrative Examples .......................................................................................... 26
   2.4.1 A univariate case ....................................................................................... 27
   2.4.2 A multivariate case .................................................................................... 31
2.5 The impact of uncertainty in the parameters ..................................................... 34
   2.5.1 Uncertainty analysis in univariate cases .................................................... 34
   2.5.2 Uncertainty analysis in multivariate cases ................................................ 38
CHAPTER 3 CHART ALLOCATION STRATEGY FOR SERIAL PARALLEL MULTISTAGE MANUFACTURING PROCESSES ........................................... 44
3.1. SP-MMP modeling and charting methods ........................................ 46
3.2 Chart allocation strategy for output monitoring .................................. 52
3.3. Numerical example ....................................................................... 59
3.4. Summary ..................................................................................... 65

CHAPTER 4 INTEGRATED MONITORING STRATEGY FOR MMP ........ 66
4.1 Introduction ..................................................................................... 66
4.2 Problem formulation ........................................................................ 67
  4.2.1 Review of chart allocation strategy ........................................... 67
  4.2.2 Integrated monitoring strategy formulation .............................. 69
4.3 Integrated monitoring strategy max-min problem formulation .......... 71
4.4 Numerical example ........................................................................ 78
4.5 Summary .......................................................................................... 81

CHAPTER 5 SMITH-EWMA RUN-TO-RUN CONTROL SCHEMES FOR A PROCESS WITH MEASUREMENT DELAY .................................... 82
5.1 Process model and Smith-EWMA ..................................................... 83
  5.1.1 Smith Predictor .......................................................................... 83
  5.1.2 Smith-EWMA model ................................................................... 85
5.2 Stability Properties .......................................................................... 89
  5.2.1 No unit measurement delay ......................................................... 90
  5.2.2 One unit measurement delay ....................................................... 91
  5.2.3 More than one unit measurement delay ...................................... 91
5.3 Performance Comparisons .............................................................. 94
5.4 Summary ......................................................................................... 100

CHAPTER 6 CONCLUSIONS ................................................................. 101
6.1 Summary ......................................................................................... 101
6.2 Future research areas ...................................................................... 103

Appendix A. Noncentrality parameters of output monitoring for multivariate cases ................................................................................................. 105
Appendix B. Noncentrality parameters of residual monitoring for multivariate cases.................................................................106
Appendix C. Parameters of univariate example for Chart allocation...........107
Appendix D. Parameters of multivariate example for Chart allocation.........108
Appendix E. Impact of parameter uncertainty for univariate cases in chart allocation...................................................................................................................109
Appendix F. Stability analysis for the Smith-EWMA controller with no measurement delay.................................................................110
Appendix G. Stability analysis for the Smith-EWMA controller with one unit of measurement delay.........................................................111
Bibliography.............................................................................................................112
Vita............................................................................................................................120
LIST OF FIGURES

Figure 1.1 An Example of Multistage Processes: A typical PCB Manufacturing Process .......................................................................................................................... 5
Figure 1.2 Classification of conventional SPC charts .................................................. 10
Figure 1.3 The IMC structure of EWMA controllers ...................................................... 16
Figure 2.1 Comparison of ATS with different fault magnitudes of 1, 3 and 5 with different lambda values (0.05, 0.1, 0.2, and 0.4). .......................................................... 29
Figure 2.2 Comparison of the noncentrality parameter intervals of different stages under uncertainty ........................................................................................................ 35
Figure 2.3 Comparison of ATS intervals of different stages under parameter uncertainty when the processing time equals 5 min. .................................................. 36
Figure 2.4 Comparison of ATS intervals of different stages under parameter uncertainty when the processing time equals 20 min. .................................................. 36
Figure 2.5 Comparison of ATS intervals of different stages under parameter uncertainty when the processing time equals 5 min using the residual monitoring method ........................................................................................................ 37
Figure 3.1 Process of block part. .................................................................................. 49
Figure 3.2 Example of convergence in SP-MMP .......................................................... 53
Figure 3.3 Example of multiple convergences in SP-MMP ......................................... 56
Figure 3.4 Hood fit process ....................................................................................... 60
Figure 5.1 The structure of a Smith predictor controlled system............................... 84
Figure 5.2 An equivalent IMC structure of a Smith predictor controlled system..... 85
Figure 5.3 The transformed structure of the EWMA controller .................................. 86
Figure 5.4 The simplified structure of EWMA controller .......................................... 86
Figure 5.5 The structure of the Smith-EWMA controller ............................................ 87
Figure 5.6 Development of Smith-EWMA and equivalent transformations ............ 88
Figure 5.7 Stability comparison between Smith-EWMA and EWMA controllers .. 92
Figure 5.8 The stability comparison between two controllers with a specific $\lambda$=0.2, 0.3 and 0.4 ...................................................................................................................... 93
Figure 5.9 Comparison of stability improvement with different discount factors $\lambda$=0.2 (dash-dot), 0.3 (dotted line), and 0.4(solid) ....................................................... 94
Figure 5.10 Performance comparison with metrology delay when b=1.5 (solid line for Smith-EWMA, dash-dotted line for EWMA, dotted line for RLS-RWD ....... 96
Figure 5.11  Performance comparison with metrology delay when $b=1$ (solid line for Smith-EWMA, dash-dotted line for EWMA, dotted line for RLS-RWD)............ 97
Figure 5.12  Performance comparison with metrology delay when $b=1.5$ (solid line for Smith-EWMA, dash-dotted line for EWMA, dotted line for RLS-RWD)............ 98
Figure 5.13  Performance comparison with metrology delay when $b=1$ (solid line for Smith-EWMA, dash-dotted line for EWMA, dotted line for RLS-RWD)............ 99
LIST OF TABLES

Table 2.1  ARL of EWMA chart in univariate example with $U_2 = \sigma_{\epsilon 1}$. .................... 28
Table 2.2  ATS (min) of EWMA chart in univariate example with $U_2 = \sigma_{\epsilon 1}$. ........... 28
Table 2.3  ARL of EWMA chart in univariate example with $U_2 = 5\sigma_{\epsilon 1}$. .................. 30
Table 2.4  ATS (min) of EWMA chart in univariate example with $U_2 = 5\sigma_{\epsilon 1}$. ........... 30
Table 2.5  ARL of residual EWMA chart in univariate example with $U_3 = \sigma_{\epsilon 2}$. ............. 31
Table 2.6  ATS (min) of residual EWMA chart in univariate example with $U_3 = \sigma_{\epsilon 2}$. .................................................................................................................................... 31
Table 2.7  Noncentrality Parameters of multivariate example with mean shift =0.5. 32
Table 2.8  ARL of MEWMA chart of multivariate example (mean shift =0.5 and 0.1, lambda=0.1). s-1, s-2, and s-3 stands for stage1,2 and 3 respectively. Shift only occurs in one of the twelve directions.................................................................32
Table 2.9  ATS (min) of MEWMA chart of multivariate example (mean shift =0.5 and 0.1, lambda=0.1). s-1, s-2, and s-3 stands for stage1,2 and 3 respectively. Shift only occurs in one of the twelve directions.................................................................33
Table 2.10 Noncentrality Parameters of multivariate example with mean shift =0.1. .................................................................................................................................... 33
Table 2.11 Factorial DOE table. ............................................................................................. 39
Table 3.1 Scenarios in SP-MMP.................................................................................. 47
Table 3.2 Fault propagation patterns and corresponding ARL for output monitoring. .................................................................................................................................... 62
Table 3.3 ARL of multiple potential faults for output monitoring. ......................... 63
Table 3.4 ATS of multiple potential faults for output monitoring.......................... 63
Table 3.5 Optimal solution of the chart allocation strategy for the hood fit process.64
Table 4.1 The decision space of integrated monitoring strategy. ............................... 70
Table 4.2 Noncentrality parameters for $U_2$............................................................... 79
Table 4.3 Noncentrality parameters for $U_3$............................................................... 79
Table 4.4 $\tilde{\beta}_{k,i,j}$ for $U_2$ and $U_3$................................................................................... 80
Table 4.5 Optimal solution of the integrated monitoring strategy......................... 80
### ACRONYMS

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR</td>
<td>Autoregressive</td>
</tr>
<tr>
<td>ARL</td>
<td>Average Run Length</td>
</tr>
<tr>
<td>ATS</td>
<td>Average Time to Single</td>
</tr>
<tr>
<td>CL</td>
<td>Control Limit</td>
</tr>
<tr>
<td>COPQ</td>
<td>Cost of Poor Quality</td>
</tr>
<tr>
<td>CMP</td>
<td>Chemical Mechanical Polishing</td>
</tr>
<tr>
<td>CUSUM</td>
<td>Cumulative Sum chart</td>
</tr>
<tr>
<td>DOE</td>
<td>Design of Experiments</td>
</tr>
<tr>
<td>EM</td>
<td>Expectation Maximization</td>
</tr>
<tr>
<td>EWMA</td>
<td>Exponentially Weighted Moving Average</td>
</tr>
<tr>
<td>IMA</td>
<td>Integrated Moving Average</td>
</tr>
<tr>
<td>KIRC</td>
<td>Knowledge-based interactive run-to-run controllers</td>
</tr>
<tr>
<td>MCUSUM</td>
<td>Multivariate CUSUM chart</td>
</tr>
<tr>
<td>MEWMA</td>
<td>Multivariate EWMA chart</td>
</tr>
<tr>
<td>MIMO</td>
<td>Multiple Input Multiple Output</td>
</tr>
<tr>
<td>MMP</td>
<td>Multistage Manufacturing Processes</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean Squared Error</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal component Analysis</td>
</tr>
<tr>
<td>PI</td>
<td>Proportional Integral</td>
</tr>
<tr>
<td>PLS</td>
<td>Partial Least Squares</td>
</tr>
<tr>
<td>RtR</td>
<td>Run-to-Run</td>
</tr>
<tr>
<td>RIE</td>
<td>Reactive Ion Etching</td>
</tr>
<tr>
<td>SISO</td>
<td>Single Input Single Output</td>
</tr>
<tr>
<td>S-MMP</td>
<td>Serial Multistage Manufacturing Processes</td>
</tr>
<tr>
<td>SPC</td>
<td>Statistical Process Control</td>
</tr>
<tr>
<td>SP-MMP</td>
<td>Serial Parallel Multistage Manufacturing Processes</td>
</tr>
</tbody>
</table>
Chart Allocation and Control Techniques for Multistage and Run-to-Run Processes

by JIN, Ming
Department of Industrial Engineering and Logistics Management, The Hong Kong University of Science and Technology

Abstract

Nowadays, quality is one of the very basic and crucial factors for any products or services in business world. It determines the success of an organization or product to a large extent. Large variability is believed to be the biggest enemy of excellent quality, especially, the variability caused by certain assignable causes. In practice, Statistical Process Control (SPC) techniques have been widely applied in industries to detect assignable causes and reduce variability. Most SPC charts are designed to detect assignable causes in single stage processes encountered in industrial practice. Due the complexity of modern techniques, more and more tasks and products require more than one stages/operations to accomplish. This is what we call multistage manufacturing processes (MMP). However, applications of conventional charts to such multistage processes usually result in an unsatisfactory performance due to two main obstacles.

1. Limited resources, such as measurement systems and manpower, it is usually too costly to place SPC charts at every stage, especially for the processes which have a large number of stages.

2. Incomplete information. Even if we have the capability to set up control charts for all the stages, it is still possible that the whole monitoring performance (in terms of Average Run Length and Average Time to Signal) is
not good enough, because each stage is treated as a single stage and inherent correlations among stages are ignored.

Therefore, we try to improve the monitoring performance and efficiency by investigating the inherent characteristics of multistage processes and introducing new monitoring strategy.

In this thesis, we first propose a strategy to properly allocate control charts in a serial multistage manufacturing process (S-MMP) that can enhance the fast detection of out of control behaviors of conventional SPC. Based on our chart allocation strategy, information of inherent interrelationship among stages is involved in decision making to achieve quicker detection of a potential fault. Two automotive assembly examples are used to demonstrate the applications of the chart allocation strategy. It shows that the rational strategy may differ from industrial common sense in some cases. The impact of uncertainty in the structural parameters is also considered, which may allow practitioners to make more realistic decisions in serial multistage manufacturing processes when they face such uncertainty or error. Guidelines of application procedures are also provided.

After investigating the chart allocation strategy in S-MMP, we carried on our research on extending the strategy to a more complicated and realistic situation-serial parallel multistage manufacturing processes (SP-MMP). The three special scenarios of SP-MMP are analyzed and modeled respectively. One of most important features of SP-MMP is that the mean shift in upstream stages may cause not only the mean shift but also the change of variance in downstream stages. Corresponding modifications in process modeling and optimization formulation are made in this work. Furthermore, the limited resource problem is also considered in the optimization formulation so that it can help us to determine appropriate amount of monitoring resource for a processes.

Besides the chart allocation strategy in a MMP, it is also important for people to decide which kind of control charts is appropriate for a MMP. We tackle this problem together with the chart allocation in order to form a monitoring strategy for a MMP. Control charts are divided into two classes due to the data
form: 1. Output monitoring chart; 2. Residual monitoring chart. The whole monitoring strategy is formulated to a nonlinear integer programming problem. However, when the number of stages and charts become large, it is usually hard for people to obtain the optimal solution for such problem. Therefore, we further simplify the problem to a max-min problem to achieve good solution if not the best with less effort.

Whenever a signal appears in the monitoring system of a process, people need to take corresponding actions to pull the process to the right track again. In the final part of the thesis, we also tackle this problem by studying the control policy for run-to-run processes which are common in semiconductor manufacturing and chemical engineering. Smith predictor structure is introduced together with the EWMA controllers to deal with the inherent metrology delays in run-to-run processes.
Chapter 1

Introduction

1.1 Motivation

Quality, it is one of the most basic and crucial factors for any products or services in today’s business world. Producers can enlarge their profit margins by reducing cost of poor quality (COPQ), which measures the cost due to poor quality products or services. Besides the cost savings due to high quality products, a brand with high quality image has inestimable value for any business organizations as it can greatly affect customers’ buying behavior and brand loyalty. In this section, we introduce the importance for studying SPC, and problems of implementing SPC in modern MMP.

1.1.1 Quality and statistical process control

Due to importance of quality, people have studied it for a long period of time, and defined it in multiple dimensions to cover all the related aspects. Based on Juran and Gryna (1988)’s definition, quality includes three different aspects: quality of design, quality of conformance and quality of performance (Montgomery (2005), Summers (2003)). The quality of design represents the different expected quality levels which are determined by designers intentionally; the quality of conformance reflects the degree of conformance-to-requirements; and the quality of performance evaluates how well the product performs its function. Quality can be interpreted as "Customer's expressed and implied requirements are met fully". This is a core statement from which some eminent definitions of quality have been derived. They include: "the totality of features and characteristics of a product or service that bears on its ability to meet a stated or implied need" (ISO, 1994), "fitness for use" (Juran, 1988), and "conformance to requirement" (Crosby, 1979). It is important to note that satisfying the customers' needs and expectations is the main factor in all these definitions. Therefore it is necessary for a company to identify such needs early in the product/service development cycle. The ability to define correct needs related to
design, performance, safety, and other business processes will place a firm ahead of its competitors in the market. In 1992 Crosby broadened his definition for quality adding an integrated notion to it: "Quality meaning getting everyone to do what they have agreed to do and to do it right the first time is the skeletal structure of an organization, finance is the nourishment, and relationships are the soul." Some Japanese companies find that "conformance to a standard" is too narrow compared to the actual meaning of quality and have used a newer definition of quality as "providing extraordinary customer satisfaction”.

In order to achieve extraordinary customer satisfaction, reducing variation becomes a necessity in modern industries. Nowadays, people put more and more emphasis on the root of poor quality, or the way of gaining good quality. Montgomery (2005) also noted the modern definition of quality as: quality is inversely proportional to variability. Variation is inevitable in all processes due to the fact that no two items can be produced exactly the same. However, variation shall be treated differently by different kinds of root causes. We can divide the variations into two categories:

1. chance cause variation
2. assignable cause variation

Chance cause variation means the natural variability that causes processes vary with acceptable tolerance. The chance causes may include the fluctuation of temperature and power supply, which are usually unavoidable in reality. If a process shows only chance cause variation, the process is said to be in statistical control, or in-control. In contrast, assignable cause variation is the variation caused by certain identifiable and removable reasons, such as failure of components and defects in products. When assignable causes appear in a process, the process is said to be out-of-control.

When a process is out-of-control, the produced components will not achieve their target values, and moreover, it will result in poor quality. Therefore, the early detection and removal of assignable causes are crucial to quality assurance. The detection relies on data gathering and statistical analysis. A major tool that is being widely utilized is Statistical Process Control (SPC).
Statistical process control (SPC) involves using statistical techniques to measure and analyze the variation in processes. Most often used for manufacturing processes, the intent of SPC is to monitor product quality and maintain processes to targets. SPC is used to monitor the consistency of processes which manufacture a product as designed. It aims to get and keep processes under control. No matter how good or bad the design, SPC can ensure that the product is being manufactured as designed.

Control chart is a main tool used for SPC. It is a graphical representation of certain descriptive statistics for quality measurements of a process. The descriptive statistics are displayed in the control chart in comparison to their "in-control" sampling distributions. If points are found outside the control limits, an out-of-control alarm will be issued, and operational intervention occurs to check assignable causes. After the assignable causes have been identified and removed, the process will operate at target levels again. People developed different descriptive statistics and different types of control charts which can be applied in different variation causes, such as large or small mean shifts. By using appropriate statistics, we are able to realize early detection of assignable causes and reduce the cost of poor quality.

1.1.2 Multistage manufacturing processes (MMP) in modern industries

In modern industries, operations have become very complicated. In most cases, a product cannot be produced by a single operational stage. Multistage manufacturing processes can be found in various industries, such as automobile assembly and semiconductor manufacturing (see Figure 1). Likewise, in service industries, such as financial services, medical care services, and call center industries, the thriving demand for professional and meticulous service asks for more and more detailed labor division, which can be also similarly viewed as a multistage process. In most cases, outputs from operations at upstream stages can be affected by operations at downstream stages. In addition, a product part or service transferring from one stage to the next stage in a multistage process may introduce extra variations that do not occur in a single stage process.
Due to recent development in sensing and information technologies, automatic data acquisition techniques are used increasingly in complicated processes with multiple stages, and a large amount of data and information related to quality measurements has become available. Thus, engineering and statistical approaches to make use of the multistage data and information have become possible in both industrial and service practices.

A key problem in monitoring multistage manufacturing is how to describe such a multistage process. Historically, most of the researchers tried to modeling multistage process with statistical model, such as a linear regression model. However, it is essential to incorporate engineering knowledge in multistage modeling and analysis for more effective process control and monitoring, statistical model-based methods usually cannot describe the relationship among stages explicitly due to the lack of engineering background and knowledge, and a large variety of current literature adopts multistage engineering models in a linear state space model structure based on physical laws and engineering knowledge that describe the quality information of a multistage process. See e.g., Jin and Shi (1999) and Ding, Shi and Ceglarek (2002) for rigid-part assembly processes. Djurdjanovic and Ni (2001), Huang, Zhou and Shi (2002) and Zhou, Huang and Shi (2003) for multistage machining processes. Also, Lawless, MacKay and Robinson (1999) and Agrawal, Lawless and MacKay (1999) discuss an AR (1) model, which could be put in a linear state space form, for representing the variation transmission in both multistage assembly and machining processes. The linear state space model structure provides an analytical engineering tool for modeling, analyzing and diagnosing a multistage process. An extensive review of state-space model can be found in Basseville and Nikiforov (1993) and Ding, et al (2002).

Many practical applications in multistage manufacturing processes can be found in recent literatures. Zhou, Huang and Shi (2003) discussed an example of the 2-D panel auto body assembly process, which contains multiple stages of assembly operations and product inspection for surface finish, joint quality and dimensional defects. The authors also report another example for the engine-head machining. Djurdjanovic and Ni (2001) studied a machining process that is a combination of multiple machining operation stages.
A large variety of statistical process control (SPC) schemes was developed for quality and productivity improvement since 1960s, SPC utilizes statistical methods to monitor manufacturing processes with an aim to maintain and improve the product quality while decreasing the variance. Much research has been conducted on the issues of SPC and the resulting developments are readily available in the literatures, see surveys of the literatures on SPC by Lowry and Montgomery (1995), Woodall and Montgomery (1999) and Stoumbos et al. (2000). Nevertheless, conventional SPC methods typically restricted to a single process stage in applications.

Figure 1.1. An Example of Multistage Processes: A typical PCB Manufacturing Process.

A natural application of traditional SPC techniques to modern multistage processes is to use the conventional control charts directly at stages of an MMP, considering each stage is a single stage process. Nevertheless, real applications of the above idea are not so satisfactory due to the following reasons:

1. Due to limited resources, such as measurement systems and manpower, it is usually too costly to place SPC charts at every stage, especially for the processes which have a large number of stages.

2. Even if we have the capability to set up control charts for all the stages, it is still possible that the whole monitoring performance (in terms of Average Run Length
and Average Time to Signal) is not good enough, because each stage is treated as a single stage and inherent correlations among stages are ignored.

Recently, some researchers have noticed problems in the multistage quality control area and investigated methods to improve the situation. However, most of the research does not tackle the above questions of how to apply traditional SPC techniques more efficiently in an MMP.

No matter which control chart we use, where to allocate the charts in a multistage process is always an important issue. This is because cost and resources are always limited in reality and it is not always possible to set up control charts to measure and monitor the process outputs at every single stage in a complex multistage process. When dealing with such cases, decisions about chart allocation are usually made based on common sense and experience. Because of the lack of a systematic way to support the decisions, many of the decisions cause the monitoring performance to be inefficient and costly. However, little attention has been given to this problem in SPC research.

1.1.3 Run-to-run control for processes with measurement delays

Besides the monitoring for multistage processes, the consequential control actions whenever a chart alarms also draw great attention. Run-to-run control is a popular control methodology in multistage processes, such as semiconductor manufacturing and chemical engineering processes. It is a form of discrete process and machine control. Its aim is to force the output of the process to reach a desired value (the target value) by adjusting the input of the process according to previous deviations in each run. Adjustment can be made only after the previous run and before the next run, and no adjustment can be done during a run. Run-to-run control is widely used in semiconductor manufacturing processes, such as in reactive ion etching (RIE) or chemical mechanical polishing (CMP) processes.

Many control schemes and algorithms have been developed for run-to-run control. Ingolfsson and Sachs (1993) proposed the EWMA controller for process control. The double EWMA was proposed by Butler and Stefani (1994), (1997) to eliminate
process drift, and Del Castillo, (1999), (2001), and (2002) and Ning, et al. (1996) also further studied the double EWMA controller. Moreover, other run-to-run controllers, such as self tuning controllers (Astrom and Wittenmark (1973), Clarke and Gawthrop (1975), Astrom et al. (1977), and Astrom and Wittenmark (1989)), and knowledge-based interactive run-to-run controllers (KIRC) (Hankinson et al. (1997)), have also been proposed.

Among the studies on run-to-run control, notable work has concentrated on the applications of EWMA controllers in the semiconductor manufacturing industry (Ingolfsson and Sachs (1993); Sachs et al. (1995); Del Castillo and Hurwitz (1997)). The stability conditions of EWMA controllers were presented by Ingolfsson and Sachs (1993) under the conditions of deterministic first-order and second-order systems and then drift noise and wander noise were added.

In applications, it is quite often that there exists measurement delay in the feedback control loop. In run-to-run control, we measure the outputs when a run is finished and use the measurement of this run to adjust the inputs for future runs. That is, the measurement is also run by run. Take the wafer fabrication processes as an example. Wafers are placed in a closed chamber, some critical parameters, such as wafer thickness, are impossible to measure within the processing period and can only be obtained when the whole process has finished one run. Because of this post process measurement strategy, measurement delay is inevitable in the process. This means that the measurements of the outputs cannot be fed back into the system to adjust the inputs for the next run immediately. The measurement delay will cause the adjustment to be late. As a result, the output will not immediately meet the target, which may lead to poor stability properties of the system and also bad performance. Because the measurement delay problem exists widely in run-to-run control, it is an important topic for the semiconductor industry and other industries in which this type of control is widely used.

Several researchers have discussed the performance and stability properties of EWMA controllers with such measurement delay: Good and Qin (2004) investigated the stability region both for SISO and MIMO EWMA controllers dealing with different metrology delays. The stability properties of double EWMA controllers
have been well studied in Good and Qin (2002). A comparison between run-to-run algorithms, including the EWMA with measurement delay, has been conducted by Chamness et al. (2001). Not surprisingly, all the papers we mentioned show that both the stability properties and performance of EWMA controllers are worse when there is a delay in the measurement. Therefore, researchers started to investigate and improve the EWMA type run to run controllers under metrology delay. Wang et al. (2005) proposed a run-to-run controller based on recursive least squares (RLS) estimation in order to obtain better performance in handling the measurement delay and measurement noise. Adivikolanu and Zafiriou (1997) mainly emphasized improvement of performance under disturbances; their paper also investigated stability conditions for EWMA controller systems with one unit of measurement delay. Further extensions of EWMA controllers have been done by Adivikolanu and Zafiriou (2000). They have made some modifications to the EWMA controller using internal model control to improve the performance of the controller under disturbances and have also analyzed the performance and stability properties of the modifications. However, their modifications only improved the performance of the controllers under disturbances instead of the stability. The stability property appears to be no better than with the original EWMA controllers.

Based on Adivikolanu, and Zafiriou (1997) and Good and Qin (2004)’s results, the stability properties would be worse under measurement delay. Moreover, because the model parameters in EWMA controller are always deviated from the real value to some extent, the worse stability region under measurement delay may cause the system to become unstable. To deal with this problem, people may use statistical tools, such as design of experiments and regression, to obtain more accurate model parameters, or modify the EWMA controller to enlarge the stability region. Nevertheless, most of the improvement in the literature focused on the control performance of controllers rather than the stability properties. Not much attention is paid on how to enlarge the stability region of the EWMA type controllers under measurement delay. In the thesis, we mainly focus on how to improve the stability properties and performance of the processes with measurement delays based on EWMA controllers.
1.1.4 Objectives and Scope

According to the problems we have mentioned in the multistage monitoring and Run-to-Run control, such as limited monitoring resources, lack of interrelationship considerations, and measurement delays, the objectives and scope of this research can be divided into two main parts: 1. multistage process monitoring; 2. Run-to-Run control for processes with measurement delays. In the first part, due to the limited resources or control charts in a MMP, we tackle the problems of chart allocation for both serial and serial parallel multistage processes based on conventional SPC techniques and the interrelationships among stages; furthermore, we extend the strategy to a more general form to determine both the chart allocations and charting methods to achieve greatest monitoring efficiency. In the second part of this research, we focus on the measurement delay problem in Run-to-Run processes to investigate a new control scheme which can achieve better stability and performance properties than the conventional controllers.

1.2 Literature review

1.2.1 A review of conventional SPC techniques

The single stage SPC techniques are developed for various situations, which are summarized in the following figure. A brief review and discussion of some of the most popular charts are given in follows.

For the univariate cases, $\bar{x}$-chart is one of the easiest and most widely used techniques. Let $x_i, i=1,\ldots,n$ be a sample of size $n$ collected from a process in which $x_i \sim N(0,\sigma^2)$. Then, the sample mean, which is given by $\bar{x} = (x_1 + \cdots + x_n)/n$, follows a normal distribution with mean 0 and variance $\sigma^2/n$. Suppose a series of sample averages are calculated, an $\bar{x}$ chart can be set up to monitor the process, and an out-of-control alarm will be released if the following condition is satisfied

$$|\bar{x}_t| > h,$$  \hspace{1cm} (1.1)

where $\bar{x}_t$ is the average of the $t$th sample. $h$ is the control limit. For the $\bar{x}$-chart, which is also called Shewhart chart, it uses only the latest sample and
ignores information about the process. Due to the ignorance of the historical information about processes, Shewhart control charts are relatively insensitive to small process shifts.

In order to overcome this drawback, Page (1954) proposed a Cumulative Sum (CUSUM) chart, which takes the advantage of historical information. This method was later discussed by Page (1961), Gan (1993), Lucas (1976). A thorough analysis of the CUSUM scheme can be found in Basseville and Nikiforov (1993). A one-sided CUSUM chart takes the form:

\[ C_i = \max[0, x_i - k + C_{i-1}] \quad (1.2) \]

where \( k \) is a reference value of the chart that depends on the size of interested shifts. Suppose the process is subject to a constant shift of magnitude \( \mu \), the optimal CUSUM scheme is obtained by setting \( k = \mu / 2 \). \( C_i \) is a cumulative statistic with initial value \( C_0 = 0 \). This chart alarms when the cumulative statistic exceed the control limit, \( C_i > h \).

The Exponentially Weighted Moving Average (EWMA) chart is another alternative which considers historical information of processes and also sensitive to small shifts (Aparisi and Garcia-Diaz (2004)). The EWMA chart takes the follow form:

\[ z_i = (1 - \lambda)z_{i-1} + \lambda x_i \quad (1.3) \]

where \( \lambda \) is a smoothing parameter. The statistic \( z_i \) is computed as the weighted average of the previous value and the latest observation. A two-sided EWMA chart alarms whenever \( |z_i| > h \) happens, where \( h \) is the control limit.
The above charts we have discussed are all for univariate cases. However, due to fact that the multivariate states of nature is an important feature of modern industries. Control charts for multivariate cases are also developed.

Let $\mathbf{x}_t$ be an observational vector sampled from a process at time $t$. The vector follows a $P$ -dimensional normal distribution with mean $\mathbf{\mu}$ and variance $\mathbf{\Sigma}$, $\mathbf{x}_t \sim N(\mathbf{\mu}, \mathbf{\Sigma})$. The probability density function of $\mathbf{x}_t$ is given by:

$$f(\mathbf{x}_t) = \frac{1}{\sqrt{(2\pi)^P |\mathbf{\Sigma}|}} \exp\left(-\frac{1}{2} (\mathbf{x}_t - \mathbf{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{x}_t - \mathbf{\mu})\right).$$  \hspace{1cm} (1.4)

In order to monitor the multivariate process against possible faults, the $T^2$ chart proposed by Hotelling (1947) has gained wide attention. Recent discussions are referred to Hawkins (1991), Mason et al. (1995), Lowry and Montgomery (1995) and Mason et al. (2003). Hotelling’s $T^2$ chart takes the form:

$$T^2 = (\mathbf{x}_t - \mathbf{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{x}_t - \mathbf{\mu}) > h.$$  \hspace{1cm} (1.5)

When the covariance matrix, $\mathbf{\Sigma}$, is known, the $T^2$ statistic follow a chi-square distribution with $P$ degrees of freedom. If the covariance matrix is estimated from samples, the control limit needs a slight modification (Lowry and Montgomery (1995), Montgomery (2005)).

Hotelling’s $T^2$ chart considers only the most recent observation and is a Shewhart-type control chart. Therefore, it gives satisfactory larger shift performance, but suffers from poor performance to small shifts.

Therefore, the multivariate versions of EWMA and CUSUM charts have been developed (Crosier (1988), Pignatiello and Runger (1990), Lowry et al. (1992)). Lowry et al. (1992) developed the MEWMA control chart based on the EWMA control chart.
Suppose that we monitor \( p \) variables simultaneously. We have the variables in a \( p \times 1 \) vector, \( y_{i,k} \) where \( i \) is the time point and \( k \) is the stage number. Then, the exponentially weighted moving average statistic, \( Z_{i,k} \), is obtained as:

\[
Z_{i,k} = \lambda y_{i,k} + (1 - \lambda)Z_{i-1,k},
\]

(1.6)

Where \( 0 < \lambda \leq 1 \) and \( Z_0 = 0 \). Since \( Z_i \) is a vector and is difficult to plot directly on a control chart, \( T^2 \) is used to construct the charting statistic:

\[
T^2 = Z'_{i,k} \Sigma_{Z,i,k}^{-1} Z_{i,k},
\]

(1.7)

where the covariance matrix is

\[
\Sigma_{Z,i,k} = \frac{\lambda}{2 - \lambda} \left[ 1 - (1 - \lambda)^2 \right] \Sigma
\]

(1.8)

and \( \Sigma \) is the covariance matrix of the original data, \( y_{i,k} \). Here, we can see that the univariate EWMA control chart is a special case of the MEWMA when there is only one variable.

Similarly, the multivariate CUSUM (MCUSUM) chart has been developed in parallel (Crosier (1988), Pignatiello and Runger (1990)). A review and discussion on the features of the above charts are referred to Lowry and Montgomery (1995).

### 1.2.2 Conventional monitoring of multistage processes

One way to monitor a multistage process is to apply the well developed conventional SPC techniques directly into multistage monitoring: people may monitor final stage’s outputs using SPC techniques (Xiang and Tsung, 2008), such as Shewhart charts, cumulative sum (CUSUM) charts and exponentially weighted moving average (EWMA) charts for univariate cases, and Hotelling’s \( T^2 \) charts, multivariate EWMA (MEWMA) and multivariate CUSUM (MCUSUM) charts for multivariate cases. However, the drawback of doing so is that the final-stage monitoring may not perform well due to overlooking interrelationships among stages and missing the opportunities of early detection (Xiang and Tsung (2008)). Intuitively, an alternative way to solve the above problem is to monitor quality measurements at each
individual stage. However, if the number of stage is large, it would be too expensive to use this method in reality.

Besides the direct application of traditional SPC techniques into MMP, some researchers also have investigated methods which can be more fitted in multistage processes monitoring. Hawkins (1991, 1993) developed a regression adjustment method to monitor the residuals from the regression model prediction. A similar approach named cause-selecting chart was proposed by Zhang (1984, 1985, 1989, 1992) to model a two-stage process using simple linear regression and then monitoring the residual of the current stage by subtracting the impacts from the previous stage. We call these control charts residual monitoring charts in this thesis.

Suppose there is a multistage process which has n stages in total, and the quality characteristic of stage i of product t is xi. The statistics used in regression-adjusted variable for monitoring xi is the residual left by regressing xj on all the upstream stages’ outputs. On the other hand, the cause-selecting method only regresses on the last upstream stage’s outputs. The above residual monitoring charts can describe the cascading effect in multistage processes to some extent by regressing a quality characteristic on the upstream quality information. However, they usually cannot describe the relationship among stages explicitly due to the lack of an engineering background and knowledge in their regression model. Various works on multistage processes adopts linear state space model structure as their engineering models because of its capability to incorporate physical laws and engineering knowledge that describe the quality interrelationships among multiple stages. A stream-of-variation model in a state-space form has been applied successfully to describe such interrelationships and variation propagation at the process level of a multistage process (Ding et al, 2002). With this model, physical and engineering knowledge can help to make the inherent structural information explicit. Jin and Shi (1999) and Ding, Shi and Ceglarek (2002) used this structure for a rigid-part assembly processes. Djurdjanovic and Ni (2001), Huang et al. (2002) and Zhou et al. (2003) considered applications of this model in machining processes. Tsung et al. (2008) also reviewed how the state space model was applied in multistage processes. An extensive summary of the exiting methods for multistage processes monitoring and diagnosis can be found in Shi (2007). However, all the research above either only using
data-driven method (SPC) or only focus on static data using engineering model. In SPC area, very few researches integrate the engineering knowledge together with conventional SPC techniques to achieve satisfactory performance in MMP.

As statistical model-based methods usually cannot describe the relationship among stages explicitly due to the lack of engineering background and knowledge, studies about the multistage modeling and monitoring based on engineering models attracts attentions from a lot of researchers.

A stream-of-variation model in a state-space form has been applied successfully to describe such interrelationships and variation propagation at the process level of a multistage process (Ding et al, 2002). With this model, physical and engineering knowledge can help to make the inherent structural information explicit. Djurdjanovic and Ni (2001), Huang et al. (2002) and Zhou et al. (2003) considered applications of this model in machining processes. Tsung et al. (2008) also reviewed how the state space model was applied in multistage processes.

1.2.3 EWMA Run-to-run control schemes

Run-to-run control techniques have been developed and used to control manufacturing processes in the semiconductor and chemical engineering industries. These techniques combine statistical process control, response surface and feedback control techniques. One of the most popular control schemes- EWMA controller was proposed by Ingolfsson and Sachs (1993) and also its sensitivity and stability properties were studied. The double EWMA was proposed by Butler and Stefani (1994), (1997) to eliminate process drift, and Del Castillo, (1999), (2001), and (2002) and Ning, et al. (1996) also further studied the double EWMA controller. Moreover, other run-to-run controllers, such as self tuning controllers (Astrom and Wittenmark (1973), Clarke and Gawthrop (1975), Astrom et al. (1977), and Astrom and Wittenmark (1989)), and knowledge-based interactive run-to-run controllers (KIRC) (Hankinson et al. (1997)), have also been proposed.

The following is how an EWMA controller is formed:
We assume the relation between input and output of the manufacturing process can be modeled as (Del Castillo (2002), Box and Luceno (1997), Sachs et al. (1995)):

$$Y_t = \alpha + \beta X_{t-1} + \epsilon_t$$  \hspace{1cm} (1.9)

where $Y_t$, $X_{t-1}$ denote the output and input for the t-th run, and $\{ \epsilon_t \}$ is a white noise process. Since a linear approximation to the true process model is always valid within a small area in the $X_{t-1}$ space, equation (1.9) is assumed to be the input-output relation of a manufacturing process.

If there is no process dynamics, the optimal control action is

$$X_t = \frac{T - \alpha}{\beta}$$  \hspace{1cm} (1.10)

where T is the desired target. If the parameters $\alpha$ and $\beta$ are constants through time, then we can use their estimates, which can be obtained from off-line experiments. Therefore, what we use is a pure feedback control policy (Del Castillo, Hurwitz (1997)).

However, to protect against dynamic effects not present in equation-(1.10), EWMA controllers assume that the intercept is time-varying (Ingolfsson and Sachs (1993); Sachs et al. (1991)). In this case, we use predicted output values to adjust the control action:

$$\hat{Y}_t = a_{t-1} + bX_{t-1} = T$$  \hspace{1cm} (1.11)

where $b$ is the estimate of $\beta$; and we obtain

$$X_t = \frac{T - a_i}{b}$$  \hspace{1cm} (1.12)

The estimate of the intercept is computed recursively based on the EWMA equation (Ingolfsson and Sachs (1993)):

$$a_i = \lambda(Y_i - bX_{t-1}) + (1-\lambda)a_{t-1}$$

$$= \lambda \left[ Y_i - bX_{t-1} + (1-\lambda)(Y_{t-1} - bX_{t-2}) + (1-\lambda)^2(Y_{t-2} - bX_{t-3}) + \ldots \right]$$  \hspace{1cm} (5)

Without loss of generality, the target value can be set to zero; and we have

$$X_i = -\frac{a_i}{b}$$  \hspace{1cm} (1.13)
In order to make the later development of our new Smith-EWMA controllers easy to understand, we represent the EWMA controller as an Internal-Model-Controller (IMC) (Adivikolanu and Zafiriou, 2000) as shown in Fig.1.3.

![Figure 1.3. The IMC structure of EWMA controllers.](image)

In the figure, \( p(z) = \beta \), \( \bar{p}(z) = b \), and \( g(z) = z^{-d + 1} \) (d units of measurement delay exists) are the transfer functions for a real process, a process model and feedback, respectively. According to Adivikolanu and Zafiriou (1997), the IMC controller’s transfer function is:

\[
q(z) = \frac{\lambda z / b}{z - 1 + \lambda} \tag{1.14}
\]

The IMC controller has the same performance as the EWMA controller if the tuning parameter, \( \lambda \), is consistent.

1.3 Organization of dissertation

The contributions of this thesis can be classified into two parts: first, rational chart allocation and monitoring strategies for different types of MMP are proposed. The guidelines for their implementation in real cases are also studied; second, a new run-to-run scheme which is designed for metrology processes is presented to take proper actions when there is an alarm in a MMP. The outline of the thesis is below:

Chapter 1: This chapter illustrates the motivation for launching this research. Typical example of MMP and detailed review of conventional SPC techniques and run-to-run schemes are presented.
Chapter 2: In this chapter, we propose a strategy to properly allocate control charts in a multistage process to achieve quicker detection of potential faults by using conventional control charts. Interrelationships among stages are involved in determining the chart allocations by introducing the state space model. Two automotive assembly examples are used to demonstrate the applications of the chart allocation strategy. Furthermore, we study the impact of uncertainty in the structural parameters, which may allow practitioners to make more realistic decisions in multistage manufacturing processes.

Chapter 3: This chapter proposes a strategy to properly allocate control charts in serial parallel-multistage manufacturing processes (SP-MMP) considering the interrelationship information between stages. Specialties of SP-MMP are investigated from the perspectives of process modeling and fault propagations. Modifications are made based on the chart allocation strategy in chapter 2. The strategy is formulated to an optimization problem which enables us to make rational chart allocation decisions to achieve quicker detection the whole potential fault set. Performance studies are conducted to prove the advantages of the strategy.

Chapter 4: We propose an integrated monitoring strategy to help people to decide where to place charts and which kind to place in a MMP, considering interrelationships information among stages. The strategy is formulated into a nonlinear optimization problem; moreover, it can also be formulated to a max-min problem which is possible to be linearized and much easier to obtain a good solution compared to the original nonlinear problem. A hood assembly example is used to demonstrate the applications of the chart allocation strategy.

Chapter 5: In order to maintain the satisfactory outcomes of control schemes for MMP with measurement delays, we introduce the Smith predictor control scheme, created particularly for time delay systems in control theory, into EWMA controllers which is a popular run-to-run scheme in semiconductor industries. A modification of the EWMA controller, called the Smith-EWMA run-to-run controller, is proposed. Comparisons between the stability properties of Smith-EWMA and EWMA run-to-run controllers are studied. Moreover, we conduct a performance comparison
with the EWMA and recursive least square (RLS) type controllers under disturbance conditions based on simulation.

Chapter 6: This chapter concludes the thesis with a review of the major contributions of this research and an outlook of future research directions.

CHAPTER 2

Chart Allocation Strategy for Serial MMP

The monitoring for multistage processes attracts great attention due to the popularity of MMP. In modern manufacturing industries, operations are divided into numerous steps due to the complex technology. In order to detect out-of-control conditions in multistage processes, people may monitor process outputs at the final stage using SPC techniques, or place conventional SPC techniques on every stage of a MMP. The final-stage charting may not perform well due to the fact that ignoring inherent structural information (stage-wise correlations) among stages could make the conventional SPC less effective and efficient. For the later case, it is difficult to implement such monitoring plan in a complex MMP due to cost and resources limitations. The decision of chart allocation is usually made based on people’s common sense and experience, which also ignore the interrelationship among stages. The monitoring performance is also not good enough because there is lack of useful information and systematic method in decision procedures.

In this chapter, we tackle the problem of how to rationally determine the chart allocation in MMP using the information of interrelationship among stages. We first introduce a stream-of-variance (SOV) model in a state space form to describe a general multistage process. Based on that, chart allocation strategies for both the output monitoring and residual monitoring methods are proposed in Sections 2.2 and 2.3. We aim at finding the stage in which the critical fault can be detected with greatest efficiency according to the criterion of noncentrality parameter and its
corresponding average time to signal (ATS). In Section 2.4, two real examples are used to demonstrate how the chart allocation strategies can be applied and illustrate the efficiency of the strategies. Section 2.5 investigates the impact of uncertainty in structural parameters on the proposed method. An extension to multiple faults cases is tackled via dynamic programming optimization in Section 2.6.

2.1 Multistage process modeling and charting methods

2.1.1 Multistage modeling

The modeling of multistage manufacturing processes is more complicated than a single-stage process because of the complex interrelationships among stages. For example, the quality of downstream stages can be affected by upstream stages. In such a case, a stream-of-variation model in a state-space form has been applied successfully to describe such interrelationships and variation propagation at the process level of a multistage process (Ding et al, 2002). With this model, physical and engineering knowledge can help to make the inherent structural information explicit. Djurdjanovic and Ni (2001), Huang et al. (2002) and Zhou et al. (2003) considered applications of this model in machining processes. Tsung et al. (2008) also reviewed how the state space model was applied in multistage processes.

Here, we apply the popular model represented in equation (2.1) to describe a multistage process and to develop our chart allocation strategies.

\[
y_k = C_k x_k + w_k
\]

\[
x_k = A_k x_{k-1} + U_k + v_k.
\]

(2.1)

Two kinds of quality information are described in this equation. The first is state vector, \( x_k \), such as the dimensional deviations of parts in an assembly process. The second is the observed quality information, \( y_k \), which is the quality measurement of the process output at the kth stage. Through a cascade process, these two kinds of information are transferred when a product is passed to its downstream stage. In addition, \( A_k \) denotes how the quality information in stage k-1 transfers to the quality information in stage k. \( C_k \) indicates the relationship between the quality
measurement, $y_k$, and the state vector, $x_k$, in stage k. In practice, both $A_k$ and $C_k$ can be obtained from engineering knowledge and product information. Moreover, inherent process noise is considered: $v_k$ represents the process noise such as background disturbance and unmodeled error, while $w_k$ is the measurement error, such as the sensor noise in the process.

Here $U_k$ represents a process fault or an out-of-control condition, such as an unacceptable fixture deviation. It is natural to assume that the process fault has an additive effect on $x_k$. This is consistent with the common practice in quality control, in which the process faults (e.g., the fixture error, machining error, or thermal errors in machining processes, etc.) are considered as system inputs (Zhou et al., 2004). We also assume a prior knowledge about the fault patterns, i.e., their magnitudes and directions. Many researchers have discussed on how to obtain and estimate the fault patterns in various industries and processes (Apley and Shi, 1998, Ceglarek et al. 1994, and Ceglarek and Shi, 1996). For example, in an auto assembly process, the set of potential tooling faults can be prefixed and limited to certain major elements of the fixture according to the CAD data and mechanical structure of the machines. The relationship between the faults and output measurements can well fit a linear model, so that the directions and magnitudes of the faults can be estimated accordingly (Ceglarek and Shi, 1996).

In the following, the EWMA and MEWMA charts are applied for multistage process control because of their popularity. We at first consider the situation with occurrences of a single fault in the monitoring of a multistage process. Extensions to the cases with multiple faults will be discussed in a later section.

### 2.1.2 Charting methods

Nowadays, most multistage processes also involve multiple dimensional quality characteristics. Therefore, in this paper, multivariate control charts are used. Hotelling’s (1947) $T^2$ charts, MEWMA (Lowry et al. 1992), and MCUSUM (Crosiers, 1988) are the most popular and well-known multivariate control charts in
SPC. If the number of variables is large, more advanced methods such as latent structure methods including principal components analysis (PCA) and partial least squares techniques (PLS) can be applied (MacGregor et al., 1994, Smilde et al., 2000). Recently, the MEWMA chart has received great attention from researchers because of its excellent detection power and flexibility (Prabhu and Runger, 1997, Molnau, Montgomery, and Runger, 2001, Stoumbos and Sullivan, 2002, Testik, Runger, and Borror, 2003). Thus, we will particularly emphasize the use of MEWMA in this paper (details of MEWMA chart can be found in chapter 1).

In the univariate case, process shifts are often presented in terms of standard deviation units. Similarly, we can present the shifts’ size in a multivariate case in terms of the covariance matrix:

$$\delta = (\mu^T \Sigma^{-1} \mu)^{1/2},$$

(2.2)

which is the well-known noncentrality parameter, and $\mu$ is the mean of $Y_{i,k}$. Here the larger the noncentrality parameter, the larger the shifts in the process mean. Prabhu and Runger (1997) have provided a thorough analysis of the average run length performance of the MEWMA control chart. They also give tables and charts to guide selection of the upper control limit for the MEWMA. In this chapter, we will use their results to monitor the stages and calculate the out of control ARL. According to Prabhu and Runger (1997)’s work, once the noncentrality parameter and the dimension of $Y_{i,k}$ are known, the out of control ARL is fixed. In the next sections, we develop the chart allocation strategy based on the information about $\delta$ and the out of control ARL.

### 2.2 A chart allocation strategy for output monitoring

Output monitoring charts are widely used in monitoring a multistage process (Zou et al. 2008). It is known that an important and popular criterion for evaluating the performance of control charts is the average run length (ARL). Nevertheless, the situation with multi-stages is different, as the processing (delay) time between stages should be considered in the chart evaluation. In this paper, in order to consider the processing time along with the multistage process, the average time to signal (ATS) criterion is used instead, which is also a widely used performance measure in both
industry and research (Zhang, and Wu, 2006; He, and Grigoryan, 2005; Shamsuzzaman et al, 2005).

\[ ATS = ARL \times t + \text{delay}, \quad (2.3) \]

where \( t \) is the time interval between runs. If the ATS is smaller, it means less defective products are produced (the number of defective products = production rate \( \times \) ATS). Therefore, the production cost can be reduced to some extent.

According to its definition in equation (2.3), we can see that a smaller value of ATS denotes a quicker response to the fault. The definition also shows that we only need to know the out-of-control ARL to obtain the ATS when the delay (processing) time and time interval are known in advance. Knowing the fact that the out-of-control ARL can be determined by the noncentrality parameter and that it is inversely proportional when the control chart is chosen, how to calculate the noncentrality parameters in the process becomes the key issue to obtain the ATS. In multistage processes, the shift is propagated through the downstream stages according to the parameters in the state space model (the \( A_s \) and \( C_s \)), so the shifts in the downstream stages can be calculated based on the state space model. Similarly, the corresponding noncentrality parameters can be obtained in terms of \( A_s \) and \( C_s \) for all the downstream stages. After the calculation of the noncentrality parameters, we can then use the results of Prabhu and Runger (1997) to obtain the out-of-control ARL and corresponding ATS for all the downstream stages. According to equation (2.3), we would recommend allocating the chart at the stage with the minimum ATS.

We start from a simple univariate case in which there is only one potential fault with known magnitude and the output measurements are used directly in the control chart. In such a case, there are a total of \( N \) stages in the process. We further assume that, according to engineering knowledge and historical data, the shift in stage \( i \) is critical to product quality and requires our special attention. Traditionally, we would place a control chart, say, EWMA, at stage \( i \) to monitor the potential fault. However, this may not be the best choice, as the process change may be very difficult to detect at the current stage, but it can be detected more quickly at some downstream stage due to the inherent state-wise correlation structure.
From equation (2.1), we know that the shift will be propagated through the downstream stages. The shift propagation on the downstream stages is either amplified or reduced according to different values of \( A_k \)s and \( C_k \)s in the multistage model. Our aim is to determine the stage in which the shift propagation is amplified most so that we can obtain the quickest detection of the shift (with the consideration of the processing time between stages) by allocating the control chart at this chosen stage. Here, the noncentrality parameter is used to measure how much the shift is amplified through the propagation.

By using the state space model and variation propagation, we obtain the mean shift/sigma for stage \( n \), where \( i \leq n \leq N \):

\[
\delta_n = \frac{C_n \ast (\prod_{k=n}^{i+1} A_k)U_i}{\sqrt{\sum_{l=2}^{n} [C_n A_n ... A_l]^2 \sigma_i^2 + C_n^2 \sigma_1^2 + \sigma_2^2}}.
\]  \hspace{1cm} (2.4)

Since we know that \( \delta_n \) has an inverse proportion to the out-of-control ARL in the EWMA control chart, the stage that has the largest \( \delta_n \) should also have the smallest ARL. Based on the corresponding ARL of \( \delta_n \), we can calculate the ATS accordingly and develop a chart allocation strategy by choosing the stage that has the minimal ATS.

The extended results in multivariate cases are shown in equation (2.5), which can be obtained by applying matrix manipulation on equation (2.4) (see Appendix A in the supplemental file for details):

\[
\delta_n = \sqrt{\mu' \sum_{Y(n)}^{-1} \mu_{Y(n)}} = \sqrt{M'N^{-1}M}
\]  \hspace{1cm} (2.5)

where

\[
M = C(n) \prod_{k=n}^{i+1} A(k)U(i)
\]

\[
N = \sum_{i=2}^{n} C(n)A(n)...A(i) \sum_{k=(i-1)}^{n} A(i)^T A(n)^T C(n)^T + C(n) \sum_{vn} C(n)^T + \sum_{wn}
\]

and \( \sum_{vn} \) and \( \sum_{wn} \) are the covariance matrices of \( v_n \) and \( w_n \) respectively.
By using the algorithm in Molnau et al (2001), we can obtain the corresponding out-of-control ARL and also ATS for each $\delta_k$. The control chart should be placed at the stage that minimizes the ATS.

Here, a step-by-step procedure of the proposed chart allocation strategy for output monitoring is given:

Step 1. Model a multistage process based on the state space model; the parameters (As and Cs) in the model may be obtained from engineering and physical knowledge (Ding et al, 2002; Shu and Tsung, 2003)

Step 2. Determine which fault in which stage needs particular focus. Obtain the fault pattern including the approximate direction and magnitude from historical data.

Step 3. For the fault of interest, calculate noncentrality parameters for all the downstream stages of the faulty stage (include the faulty stage itself) based on the state space model.

Step 4. According to the noncentrality parameters, the out-of-control ARL and also ATS can be obtained for each downstream stage.

Step 5. Determine the stage that has the smallest ATS and place a control chart at that stage.

Step 6. If there is some other fault of interest or any new fault is found in the diagnosis, we may approach it iteratively by going back to step 2 and running the procedure again to update the allocation strategy.

### 2.3 A Chart allocation strategy for residual monitoring

As the correlation structure between stages can be characterized by the multistage model in Section 2, one may consider monitoring the residual after a model-based prediction instead of monitoring the process output directly. Hawkins (1991, 1993) developed a regression adjustment method to monitor the residuals from the regression model prediction. A similar approach named cause-selecting chart was proposed by Zhang (1984, 1985, 1989, 1992) to model a two-stage process using simple linear regression and then monitoring the residual of the current stage by subtracting the impacts from the previous stage. Residual monitoring based on the
process model is a popular alternative because it incorporates the multistage model information and can better diagnose the out-of-control stage (Shu, Apley and Tsung, 2003; Shu, Tsung, and Kapur, 2004; Shu and Tsung, 2003; Li and Tsung, 2008).

We also investigate the chart allocation strategy for residual monitoring. The residuals can be obtained by subtracting the predicted values based on the state space model and the observations from a previous stage. Here, our aim is to find an appropriate stage, $n$, in which we can detect the fault as soon as possible if we monitor the residuals. Assume that we are able to obtain quality measures from stage $j$ that can then be used to predict the downstream stages’ outputs based on the multistage state space model. Here, our focus is on a potential fault at a downstream stage, $i > j$, and we may generate the residuals for stage $i$ and its downstream stages based on the predictions from stage $j$. Similar to what was demonstrated in the previous section, the noncentrality parameter is used to measure the magnitude of the shift propagation except that it is calculated based on the residuals we obtain. Firstly, we derive the noncentrality formula for univariate cases. By using the state space model, the noncentrality parameter is obtained as:

$$
\delta_n = \sqrt{\sigma^{-2}_{Y(n)_{res}} \mu^2_{Y(n)_{res}}}, \quad (2.6)
$$

where

$$
\mu_{Y(n)_{res}} = C(n) \prod_{k=n}^{i-1} A(k) U(i),
$$

and

$$
\Sigma_{Y(n)_{res}} = \sum_{k=j+2}^{n} [C(n)A(n)...A(k)]^2 \sigma^2_{1} + C^2(n)\sigma^2_{1}
+ \sigma^2_{2} + [C(n)A(n)...A(j+1)C^{-1}(j)]^2 \sigma^2_{2}, \quad (2.7)
$$

where $\sigma^2_{1}$ and $\sigma^2_{2}$ are the variances of $v_k$ and $w_k$ respectively. We can then obtain the out-of-control ARL as well as ATS for each $\delta_n$ based on Molnau et al’s (2001) algorithm.

Here, we also extend the residual monitoring strategy to multivariate cases by applying matrix manipulation on equation (2.6) and (2.7) (see Appendix for details). The results are as follows:

$$
\delta_n = \sqrt{\mu^\prime_{Y(n)_{res}} \Sigma_{Y(n)_{res}}^{-1} \mu_{Y(n)_{res}}}, \quad (2.8)
$$

where

$$
\mu_{Y(n)_{res}} = C(n) \prod_{k=n}^{i-1} A(k) U(i) \quad (2.9)
$$
and

\[
\sum_{Y(n)_{res}} = \sum_{k=j+2}^{n} C(n)A(n)\ldots A(k)\Sigma_{j}A(k)^{T}\ldots A(n)^{T}C(n)^{T} + C(n)\Sigma_{j}C(n)^{T} + \sum_{j} + C(n)A(n)\ldots A(j+1)C^{-1}(j)\sum_{j}C^{-1}(j)^{T}A(j+1)^{T}\ldots A(n)^{T}C(n)^{T},
\]  

(2.10)

where \( \Sigma_{1} \) and \( \Sigma_{2} \) are the covariance matrices of \( v_{k} \) and \( w_{k} \) respectively.

When we obtain the noncentrality parameter values, the corresponding out-of-control ARL and ATS can also be obtained. Based on that, a step-by-step procedure of the proposed chart allocation strategy for residual monitoring is given as follows.

Step 1. Model a multistage process based on the state space model; the parameters (As and Cs) in the model may be obtained from engineering and physical knowledge.

Step 2. Determine which fault in which stage, i, needs particular focus. Obtain the fault pattern including approximate direction and magnitude from historical data.

Step 3. Find an upstream stage, j < i, to obtain quality measures and use that to predict the output of stage i and its downstream stages.

Step 4. Based on the predictions, the residuals of stage i and its downstream stages can be calculated.

Step 5. Calculate the noncentrality parameters for the residuals of the downstream stages (including stage i).

Step 6. According to the noncentrality parameters, out-of-control ARL and ATS can be obtained for each downstream stage.

Step 7. Determine the stage with minimal ATS and place a control chart at that stage.

Step 8. If there is some other fault of interest or any new fault is found in the diagnosis, we may approach it iteratively by going back to step 2 and running the procedure again to update the allocation strategy.

2.4 Illustrative Examples

In the above sections, we have proposed a modeling method and theoretical results of our chart allocation strategies. In this section, two multistage assembly processes are used to illustrate how to apply the chart allocation strategy procedure to real
applications. The results indicate that the monitoring performance is at least as good as the common-sense chart allocation and, in many cases, better by using the proposed strategy.

### 2.4.1 A univariate case

We consider the hood assembly example in Lawless et al. (1999), which includes the flushness of 19 cars measured at each of the four plant operations, i.e., the number of operation stages is $N=4$. Lawless et al. (1999) analyzed the variation in a multistage manufacturing process based on a transmission model. Using appropriate transformation, the model can be rewritten as a state space model as in Xiang and Tsung (2008). Details of the example can be found in the appendix.

Step 1. Model the process

Steps 2 and 3. Identify the potential fault as reviewed in Section 2, and calculate the noncentrality parameters accordingly.

Suppose that the mean shift $U_2 = \sigma_{\epsilon_1}$ occurring in stage 2, i.e., $U_2 = \sigma_{\epsilon_1} = 0.13$. We can obtain:

$$\delta_2 = 0.61$$
$$\delta_3 = 0.5$$
$$\delta_4 = 0.363$$

Step 4. Obtain ARL and ATS tables:
We further assume the time interval between runs to be one minute, and the processing time at each stage to be five minutes. Based on the noncentrality parameters, we can then obtain the corresponding ARL and ATS tables of the EWMA chart with different lambda values.
Table 2.1. ARL of EWMA chart in univariate example with $U_2 = \sigma_{e1}$

<table>
<thead>
<tr>
<th>Lambda</th>
<th>$\lambda=0.05$</th>
<th>$\lambda=0.1$</th>
<th>$\lambda=0.2$</th>
<th>$\lambda=0.4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stages</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>27.1</td>
<td>25.7</td>
<td>28.9</td>
<td>41.7</td>
</tr>
<tr>
<td>3</td>
<td>37.3</td>
<td>37.4</td>
<td>44.1</td>
<td>63.6</td>
</tr>
<tr>
<td>4</td>
<td>65.8</td>
<td>70.6</td>
<td>84.8</td>
<td>114.8</td>
</tr>
</tbody>
</table>

Table 2.2. ATS (min) of EWMA chart in univariate example with $U_2 = \sigma_{e1}$

<table>
<thead>
<tr>
<th>Lambda</th>
<th>$\lambda=0.05$</th>
<th>$\lambda=0.1$</th>
<th>$\lambda=0.2$</th>
<th>$\lambda=0.4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stages</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>27.1</td>
<td>25.7</td>
<td>28.9</td>
<td>41.7</td>
</tr>
<tr>
<td>3</td>
<td>42.3</td>
<td>42.4</td>
<td>49.1</td>
<td>68.6</td>
</tr>
<tr>
<td>4</td>
<td>75.8</td>
<td>80.6</td>
<td>94.8</td>
<td>124.8</td>
</tr>
</tbody>
</table>

Step 5. Select the best stage to monitor:

From the above tables, we may conclude that it is better to monitor the second stage for this particular fault as it has the minimal ATS regardless of the EWMA charting parameter lambda. Moreover, the results in Table 2.2 can help us to determine a better weight ($\lambda$) of the EWMA chart to monitor the fault. In this case, $\lambda=0.1$ seems a better choice among all the values in Table 2.2, although the performance may not differ when $\lambda$ is between 0.05 and 0.2.

Since sometimes the magnitude of a fault cannot be estimated accurately, a sensitivity analysis should be conducted over a range of fault magnitudes to test if our decision is robust to the fault magnitude uncertainty. Considering the cases when $U_2 = 3\sigma_{e1}$ and $U_2 = 5\sigma_{e1}$, we compare the results with different fault magnitudes in Figure 2.1.
By comparing the results with different fault magnitudes in Figure 2.1, we can see that the relative order of the ATS in the stages remains the same even when the magnitude varies. Therefore, we may still conclude that it is better to monitor stage 2 in this case even if the fault magnitude is uncertain. In fact, in a univariate case, the wrong assumption about the fault magnitude may not influence the result of chart allocation strategy much, since the relative order of the noncentrality parameters in the downstream stages remains the same regardless of the magnitude of the fault based on equation (2.5).

In this case, the chart allocation coincides with our intuition, i.e., monitoring the faulty stage. However, the proposed strategy does not necessarily always lead to such an intuitive conclusion. Next, we illustrate a counter example.

In the above case, if $\beta_2 = 10$ and the processing time at each stage is assumed to be two minutes, the noncentrality parameter $\delta_3$ will be enlarged. According to our results in Section 3, $\delta_3 = 5.81$.

Therefore, the corresponding ARL and ATS tables are changed to be:
The above tables indicate that monitoring stage 3 may obtain an out-of-control signal quicker than monitoring stage 2 for all the $\lambda$ values. This conclusion differs from our intuition that we should monitor the faulty stage itself and also proves that it is necessary to investigate the chart allocation strategy even if the potential faulty stage is known. Besides the parameters in the state space model, the processing time can also influence the chart allocation decision. If the processing time is long enough, we tend to monitor on the faulty stage itself instead of its downstream stage. In the above case, if the processing time is longer than 5 min, stage 2 would be better no matter which $\lambda$ we choose. However, if the processing time is short, it becomes possible that we should monitor a downstream stage as we can see in Table 2.4.

Next, we demonstrate how to implement the chart allocation strategy for residual monitoring in this example. Suppose that a fault in stage 3 is critical and needs our attention. We have already approximated the magnitude of the fault based on historical data, in which the $U_2^2 = \sigma^2_{e1}$. According to the residual monitoring scenario we set in Section 2.3, based on stage 2’s output we obtain the residuals of stage $n \geq 3$. Based on the results in Section 2.3, we can obtain:

$$\delta_3 = 0.562$$
$$\delta_{\lambda} = 0.388$$

The ARL and ATS tables of EWMA chart can constructed as well:
Table 2.5. ARL of residual EWMA chart in univariate example with \( U_3 = \sigma_e^2 \)

<table>
<thead>
<tr>
<th>( \lambda )</th>
<th>( \lambda = 0.05 )</th>
<th>( \lambda = 0.1 )</th>
<th>( \lambda = 0.2 )</th>
<th>( \lambda = 0.4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stages 3</td>
<td>30.8</td>
<td>29.9</td>
<td>34.4</td>
<td>49.8</td>
</tr>
<tr>
<td>Stages 4</td>
<td>58.2</td>
<td>61.8</td>
<td>74.4</td>
<td>102.6</td>
</tr>
</tbody>
</table>

Table 2.6. ATS (min) of residual EWMA chart in univariate example with \( U_3 = \sigma_e^2 \)

<table>
<thead>
<tr>
<th>( \lambda )</th>
<th>( \lambda = 0.05 )</th>
<th>( \lambda = 0.1 )</th>
<th>( \lambda = 0.2 )</th>
<th>( \lambda = 0.4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stages 3</td>
<td>30.8</td>
<td>29.9</td>
<td>34.4</td>
<td>49.8</td>
</tr>
<tr>
<td>Stages 4</td>
<td>63.2</td>
<td>66.8</td>
<td>79.4</td>
<td>107.6</td>
</tr>
</tbody>
</table>

According to Table 2.6, we may conclude that the best choice is to monitor the third stage of this fault, which is consistent with our intuition, i.e., monitoring the faulty stage.

2.4.2 A multivariate case

Here we consider an application in an automotive assembly process that involves multivariate characteristics (Ding et al., 2002). It is a three stage process in which a fixture is installed at each stage for calibrating.

Step 1. We model this process by using the linear state space model in equation (2.1). The initial state, \( X_0 \), indicates a product error resulting from the process prior to this process and is assumed to be 0 in our case. In this example, the state vector at each stage is a \( 12 \times 1 \) vector, which means that we have 12 state quality variables to deal with at every stage. The coefficient matrices, \( A_k \) and \( C_k \), can be found in Appendix for details. The noise terms, \( V \) and \( W \), can be estimated by the EM algorithm. The algorithm consists of two steps: the E-step and the M-step. Details of both steps can be found in Xiang and Tsung (2008). Based on the EM algorithm, the covariance matrices can be approximated as follows:

\[
\hat{\Sigma}_v = 0.01I_{12 \times 12},
\]

\[
\hat{\Sigma}_{w_1} = diag(0.085^2, 0.06^2)
\]

and

\[
\hat{\Sigma}_{w_2} = \hat{\Sigma}_{w_1} = diag(0.085^2, 0.06^2, 0.01^2, 0.07^2)
\]

Step 2. If the fixture is well maintained, i.e., the fixture locators on the stages are free of error, the process is in a normal operating condition and thus this
in-control process can be described by model (2.1) with $U_k = 0$. The process may be out of control if a fixture fault occurs at any stage.

For demonstration, we assume the potential fault is in stage 1. Since the output of stage 1 has 12 variables (directions), we further assume that only one of them has a shift with $U_k=\sigma_r=0.5$. Such information is usually obtained from historical data and engineering knowledge.

Step 3: Obtain the noncentrality parameters. After the information on the fault is obtained, we can then calculate the noncentrality parameter for every stage with different shift directions.

<table>
<thead>
<tr>
<th>Shift directions</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>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage 1</td>
<td>1.45</td>
<td>7.965</td>
<td>49.99</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Stage 2</td>
<td>0.08</td>
<td>0.055</td>
<td>0</td>
<td>0.08</td>
<td>0.055</td>
<td>49.99</td>
<td>14.35</td>
<td>5.55</td>
<td>35.36</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Stage 3</td>
<td>0</td>
<td>0.02</td>
<td>28.30</td>
<td>0</td>
<td>0.02</td>
<td>9.895</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.30</td>
<td>0.03</td>
<td>35.36</td>
</tr>
</tbody>
</table>

Step 4. Obtain the ARL and ATS tables: Based on Table 2.7, we also can obtain the corresponding ARL and ATS by using equation (2.9) and by Molnau et al.’s (2001) algorithm. Results are shown in Tables 2.8 and 2.9. Here the in-control ARL is assumed to be 200.

Table 2.7. Noncentrality Parameters of multivariate example with mean shift =0.5

<table>
<thead>
<tr>
<th>Shift directions</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>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meanshift=0.5</td>
<td>6.42</td>
<td>1</td>
<td>1</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>s-1</td>
<td>125.39</td>
<td>149.85</td>
<td>200</td>
<td>125.35</td>
<td>149.85</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>200</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>s-2</td>
<td>200</td>
<td>174.97</td>
<td>1</td>
<td>200</td>
<td>174.97</td>
<td>1</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>25.6</td>
<td>170</td>
</tr>
<tr>
<td>s-3</td>
<td>72.4</td>
<td>6.79</td>
<td>1.03</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>Meanshift=0.1</td>
<td>178.88</td>
<td>179.36</td>
<td>200</td>
<td>178.88</td>
<td>179.36</td>
<td>1.03</td>
<td>3.02</td>
<td>10.58</td>
<td>1.85</td>
<td>200</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>s-1</td>
<td>200</td>
<td>179.62</td>
<td>2.02</td>
<td>200</td>
<td>179.62</td>
<td>5.3</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>169.23</td>
<td>179.56</td>
<td>1.85</td>
</tr>
<tr>
<td>s-2</td>
<td>200</td>
<td>179.62</td>
<td>2.02</td>
<td>200</td>
<td>179.62</td>
<td>5.3</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>169.23</td>
<td>179.56</td>
<td>1.85</td>
</tr>
</tbody>
</table>

Table 2.8. ARL of MEWMA chart of multivariate example (mean shift =0.5 and 0.1, lambda=0.1). s-1, s-2, and s-3 stands for stage1,2 and 3 respectively. Shift only occurs in one of the twelve directions.
Table 2.9. ATS (min) of MEWMA chart of multivariate example (mean shift =0.5 and 0.1, lambda=0.1). s-1, s-2, and s-3 stands for stage1,2 and 3 respectively. Shift only occurs in one of the twelve directions.

<table>
<thead>
<tr>
<th>Shift directions</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>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>s-1 Meanshift=0.5</td>
<td>6.42</td>
<td>1</td>
<td>1</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>s-2</td>
<td>130.39</td>
<td>154.85</td>
<td>205</td>
<td>130.35</td>
<td>154.85</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>205</td>
<td>205</td>
<td>205</td>
</tr>
<tr>
<td>s-3</td>
<td>210</td>
<td>184.97</td>
<td>11</td>
<td>210</td>
<td>184.97</td>
<td>11</td>
<td>210</td>
<td>210</td>
<td>210</td>
<td>35.6</td>
<td>180</td>
<td>11</td>
</tr>
<tr>
<td>s-1 Meanshift=0.1</td>
<td>72.4</td>
<td>6.79</td>
<td>1.03</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>s-2</td>
<td>183.88</td>
<td>184.36</td>
<td>205</td>
<td>183.88</td>
<td>184.36</td>
<td>6.03</td>
<td>8.02</td>
<td>15.58</td>
<td>6.85</td>
<td>205</td>
<td>205</td>
<td>205</td>
</tr>
<tr>
<td>s-3</td>
<td>210</td>
<td>189.62</td>
<td>12.02</td>
<td>210</td>
<td>189.62</td>
<td>15.3</td>
<td>210</td>
<td>210</td>
<td>210</td>
<td>179.23</td>
<td>189.56</td>
<td>11.85</td>
</tr>
</tbody>
</table>

Step 5: Find the best stage to allocate the control chart.

With the above tables, we first identify the direction of the fault. For example, if the shift at stage 1 is in its sixth direction, which means \( U = [0 \ 0 \ 0 \ 0 \ 0.5 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0] \), we can find the corresponding rows. After that, we focus on the fault in sixth direction under \( \lambda = 0.1 \). It shows us that ATS1=200, ATS2=6, and ATS3=11. In this case, stage 2 should be a better chart allocation location because of its resulting smaller ATS. With such tables, if we know the fault information in advance, we can choose the best monitoring strategy. We further consider the faults with different magnitudes.

The noncentrality parameters are shown in Table 2.10. Tables 2.8 and 2.9 include the ARL and ATS performances when mean shift =0.1, say, \( u_1 = [0.1,0,0,0,0,0,0,0,0,0,0,0,0] \) with \( \sigma = 0.1 \).

Table 2.10. Noncentrality Parameters of multivariate example with mean shift =0.1.

<table>
<thead>
<tr>
<th>Shift directions</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>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage 1</td>
<td>0.29</td>
<td>1.5927</td>
<td>9.998</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Stage 2</td>
<td>0.016</td>
<td>0.01075</td>
<td>0</td>
<td>0.01581</td>
<td>0.01075</td>
<td>9.9982</td>
<td>3.469</td>
<td>1.11</td>
<td>7.0709</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Stage 3</td>
<td>0</td>
<td>0.00396</td>
<td>5.659</td>
<td>0</td>
<td>0.0039</td>
<td>1.9789</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0598</td>
<td>0.0057</td>
<td>7.06987</td>
</tr>
</tbody>
</table>

By comparing the ARL and ATS for different mean shift magnitudes in Tables 2.8 and 2.9, we can conclude that the relative order of ARLs of different stages remains the same for the same fault, i.e., stage 2 always has the smallest ARL for the ninth direction shift in both tables. If there is only one component in \( U_i \) has a mean shift, say, \( U_i = (1,0,...,0,0,..0)' \), the wrong assumption about the magnitude of the first component will not change the relative order of the noncentrality parameters and the chart allocation decision will remain the same if the
processing delay is not very long. The underlying reason for this phenomenon is that
the noncentrality parameters are proportional to the magnitude of the fault (if the
shift direction is fixed), which means that the relative order of $\delta$ s of different stages
does not change nor does the order of ARLs. However, if the assumption about the
direction of the fault is wrong, say the real fault is $U_i=(0,1,\ldots,0,0,\ldots,0)'$ rather than
$U_i=(1,0,\ldots,0,0,\ldots,0)'$, it is possible that the chart allocation strategy will become
inappropriate. This possibility may also be caused by the misspecification of the
parameters $A_s$ and $C_s$ in the multistage state space model.

2.5 The impact of uncertainty in the parameters

In the above examples, we have shown that our chart allocation strategies achieve
good performance based on the inherent structural information of the processes, since
we always choose to monitor the stage at which we can achieve the shortest ATS for
the potential fault. However, the inherent structural information, such as matrices $A_s$ and $C_s$, may not be always accurate. If that was the case, any conclusion based on
minimum ATS could be in doubt. Here, we present how to analyze the impact of
such uncertainty on our chart allocation strategies in both univariate and multivariate
cases.

2.5.1 Uncertainty analysis in univariate cases

Here, we revisit the hood assembly example in Lawless, MacKay and Robinson
(1999). We develop an interval of the objective function’s value according to the
uncertainty of the $A_s$ and $C_s$.

2.5.1.1 Output data monitoring

The structure information we obtained in model (1) is $\overline{A}_i$ and $\overline{C}_i$ for stage i. $\overline{A}_i$ and
$\overline{C}_i$ are the measured information with errors. Suppose that the true value of
$A_i \in [\overline{A}_i - \Delta_i, \overline{A}_i + \Delta_i]$ , where $\Delta_i \geq 0$ , and $\overline{A}_i - \Delta_i \geq 0$; $C_i \in [\overline{C}_i - \Delta_2, \overline{C}_i + \Delta_2]$ where $\Delta_2 \geq 0$ and $\overline{C}_i - \Delta_2 \geq 0$. According to the state space model in equation (2.1),
the intervals of the noncentrality parameters are obtained (details are given in Appendix):
Based on the results in Appendix, we are able to estimate the upper and lower limits of the noncentrality parameters in every downstream stage. Considering the same faults in Section 5, when $U_2 = \sigma_{e_2}$ occurs in stage 2, the noncentrality parameters in the downstream stages are:

$$\delta_2 = 0.61$$
$$\delta_3 = 0.5$$
$$\delta_4 = 0.363.$$ 

Since errors exist when we estimate the inherent structure information, we set a tolerance interval for each parameter, say,

$$(\beta_1, \beta_2, \beta_3) \in (1.15 \pm 0.1, \ 0.98 \pm 0.1, \ 1.06 \pm 0.1)$$
$$(\alpha_1, \alpha_2, \alpha_3) \in (1 \pm 0.1, \ 1 \pm 0.1, \ 1 \pm 0.1).$$

The limits of the noncentrality parameters can be obtained as follows:

$$\delta_2 \in [0.516, 0.715]$$
$$\delta_3 \in [0.360, 0.681]$$
$$\delta_4 \in [0.229, 0.557].$$

![Figure 2.2. Comparison of the noncentrality parameter intervals of different stages under uncertainty](image)

When a processing time of five minutes is added into our consideration, we can obtain the following ATSs:

$$ATS_2 \in [20.77, 41.28]$$
$$ATS_3 \in [27.976, 91.135]$$
$$ATS_4 \in [45.1, 196.087]$$

35
We may conclude that monitoring stage 2 is better than monitoring stage 4. However, we cannot ensure that monitoring stage 2 can achieve a better performance than monitoring stage 3 since there is an overlap between the ATS intervals. When we face such a situation, some other factors (e.g., monitoring costs) may become important criteria for the chart allocation decision.

Note that the information on the processing time could change our conclusion. If the processing time equals 20 min, then the interval of ATS3 is changed to:

\[ [42.976, 106.135] \]

In this case, monitoring stage 2 is the best choice among all.

### 2.5.1.2. Residual monitoring

Likewise, we investigate the residual monitoring approach. The potential fault is in stage 3, and we use the output of stage 2 to predict the output of the downstream stages.
In our example, if there is no parameters’ uncertainty, the noncentrality parameters of stages 3 and 4 are:

\[
\begin{align*}
\delta_3 &= 0.562 \\
\delta_4 &= 0.388 .
\end{align*}
\]

If we set the tolerance interval for the parameters to

\[
(\beta_1, \beta_2, \beta_3) \in (1.15 \pm 0.1, ~ 0.98 \pm 0.1, ~ 1.06 \pm 0.1)
\]

\[
(\alpha_1, \alpha_2, \alpha_3) \in (1 \pm 0.1, ~ 1 \pm 0.1, ~ 1 \pm 0.1),
\]

then, according to the results of noncentrality parameters’ intervals in Appendix D in the supplemental file, the intervals of the noncentrality parameters can be obtained as follows:

\[
\begin{align*}
\delta_3 &\in [0.878, 1.392] \\
\delta_4 &\in [0.542, 1.077].
\end{align*}
\]

Therefore, the corresponding ATS intervals are:

\[
\begin{align*}
ATS_3 &\in [11.26, 18.8] \\
ATS_4 &\in [19.5, 47.2].
\end{align*}
\]

Since there is no overlap between these two intervals, we can surely claim that stage 3 is better even under the consideration of parameter uncertainty.

From the above examples, we can tell if monitoring at a certain stage is really better in reality by studying the impact of parameter uncertainty. In the above results, for example, if the intervals of \( ATS_3 \) and \( ATS_4 \) do not have any overlap, it means that monitoring one of them obtains a better performance than the other. If overlap exists,
managers need to consider the performance and cost together to determine a suitable monitoring strategy.

2.5.2 Uncertainty analysis in multivariate cases

When we try to extend the uncertainty analysis to multivariate cases, it becomes rather difficult to obtain an analytical relationship between the interval of the objective function’s value and the parameter uncertainty. However, we can still obtain numerical results of the intervals based on the design of experiments (DOE) approach. The idea is similar to performance measure modeling (Wu and Hamada, 2000; Dasgupta and Wu, 2006) with our objective function being the performance measure.

To present the problem in a DOE setting, we would use the noncentrality parameters instead of the ATS as the responses, since more computation error would be introduced when we convert the noncentrality parameters into ATS. Each uncertain parameter is considered as a factor, and the bounds of a factor are determined by the measurement error of the measurement instrument. The lower bound of the factor is set as the low level (-1), while the upper bound is set as the high level (+1). With the factorial design matrix as in Table 2.11 and following standard DOE analysis procedure (Wu and Hamada, 2000, pp. 96-141; Montgomery, 2005, pp. 244-270), we can investigate not only the maximum and minimum of the objective function (i.e., the response) but also the relationship between the objective function and the parameter uncertainty (i.e., the factors).

Let us revisit C1 of the multivariate case in the assembly process application. In this 2×12 matrix of C1, only four elements have nonzero values, and these estimates may suffer from uncertainty as well. Here, we assume that these estimates have ±10% measurement errors:

\[
\begin{align*}
X_1: & \quad C1(1,1) \in [0.9, 1.1] \\
X_2: & \quad C1(1,3) \in [-605, -495] \\
X_3: & \quad C1(2,2) \in [0.9, 1.1] \\
X_4: & \quad C1(2,3) \in [-110, -90].
\end{align*}
\]
Since there are four uncertain parameters in this example, we apply a $2^4$ orthogonal array as in Table 2.11. Based on these combinations, we can obtain the response values (the noncentrality parameters) according to the relation in equation (2.5), with a focus on the potential fault in stage 1, $U_1=[0.5 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0]$. Detailed DOE table can be found in table 2.11.

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$\delta_1$</th>
<th>$\delta_2$</th>
<th>$\delta_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>1.3082</td>
<td>0.079</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>1.5987</td>
<td>0.079</td>
<td>0</td>
</tr>
<tr>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1.5751</td>
<td>0.079</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1.9247</td>
<td>0.079</td>
<td>0</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>1.3016</td>
<td>0.079</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>1.5906</td>
<td>0.079</td>
<td>0</td>
</tr>
<tr>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>1.5674</td>
<td>0.079</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>1.9153</td>
<td>0.079</td>
<td>0</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1.0822</td>
<td>0.079</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1.3226</td>
<td>0.079</td>
<td>0</td>
</tr>
<tr>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>1.3092</td>
<td>0.079</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>1.5998</td>
<td>0.079</td>
<td>0</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>1.0767</td>
<td>0.079</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>1.3158</td>
<td>0.079</td>
<td>0</td>
</tr>
<tr>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1.3026</td>
<td>0.079</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1.5918</td>
<td>0.079</td>
<td>0</td>
</tr>
</tbody>
</table>

After analyzing the results in the DOE table by standard procedures as in Wu and Hamada (2000), we can obtain the following regression model:

$$
\delta_1 = 1.4614 + 0.146X_1 + 0.1368X_2 - 0.0037X_3 - 0.1363X_4 + 0.0136X_1X_2 - 0.0136X_1X_4 - 0.11X_2X_4
$$

It can be seen that $\delta_1$ is maximized with the combination of (1,1,1,-1) and minimized with (-1,-1,1,1). The corresponding maximum and minimum of $\delta_1$ are:
Based on that, we obtain the estimated interval of $\delta_1$ $[1.0767, 1.9153]$, and we can calculate the ATS interval accordingly:

$$ATS_1 \in [4.03, 9.13]$$

Note that the parameter uncertainty has little impact on $ATS_2 = 151.05$ and $ATS_3 = 210$, since, through the stage-wise transformation in the state space model, the impact of the parameter uncertainty has been largely reduced or even eliminated. From the above result, it is concluded that monitoring the stage 1 is the best choice for this particular potential fault with the consideration of the parameter uncertainty.

### 2.6 Extension to multiple faults cases

In the guidelines in Sections 2.2 and 2.3, we consider how to deal with newly discovered faults by iterating the procedures. However, if we already have a set of multiple potential faults before the monitoring, we should take that into consideration.

Here the chart allocation strategy is to ensure the best performance to the whole fault set with $n$ potential faults $\{F_1, F_2, \ldots, F_n\}$ when the number of charts ($m$) is fixed. To extend the chart allocation strategy to the case of multiple potential faults, we may formulate this optimization problem as a dynamic programming problem:

- **Decision variable:**
  
  $S_j$ - binary variable, 1 means a chart allocated at stage $j$.

- **Objective function:**
  
  In order to consider the performance of the chart allocation strategy on the whole fault set, we try to minimize the sum of expected ATS of all the charts:

  $$\min \sum_{k=1}^{N} S_k \times E(ATS_{i,k})$$

  where the expected ATS of a chart on stage $k$ is defined as

  $\delta_{\text{max}} = 1.9153$

  $\delta_{\text{min}} = 1.0767$.
\[ E(\text{ATS}_{i,k}) = EE[\text{ATS}_{i,k} | \text{ith fault in the upstream occurs}] = \sum_{j=1}^{j} \bar{p}_j \times \text{ATS}_{i,k} \]  

(2.12)

and \( \bar{p}_i \) is the probability of ith fault conditional on the faults in the upstream of stage k; \( \text{ATS}_{i,k} \) is the ATS of the ith fault on the kth stage.

Subject to:

\[ S_j = 1 \text{ or } 0 \]

\[ \sum_{j=1}^{N_j} S_j = m \]

\[ \sum_{j=g}^{N} S_j \geq 1 \]

Faults set=\( \{F_1, F_2, \ldots, F_n\} \)

g is the stage index of the last potential fault. This constraint is to ensure that no potential fault is ignored in the chart allocation strategy. Due to the fact that the allocation decisions on stages are sequentially related, which fits a typical dynamic programming scenario (Denardo, 1982, pp. 3, Smith, 1991, pp. 11-13), we then apply the dynamic programming method to obtain the optimal solution.

We shall use the notation \( C(S_k, k) = S_k \times E(\text{ATS}_{i,k}) \), which can be treated as the cost function only in stage k. The state in stage k is the remaining number of charts-c. The optimal cost (sum of expected ATS) from stage k onwards to the end is:

\[ f(c, k, S_k, S_{k+1}, \ldots, S_N) = \min_{j=k}^{N} \sum_{j=k}^{N} S_j \times E(\text{ATS}_{i,j}) \]

The above function also needs to subject to the constraints:

\[ S_j = 1 \text{ or } 0 \]

\[ \sum_{j=k}^{N} S_j = c \]

\[ \sum_{j=g}^{N} S_j \geq 1 \]

With the above definitions, the dynamic programming recursion can be formulated as:

\[ f(c, k, S_k, S_{k+1}, \ldots, S_N) = \min_{S_k} \{ f(c - S_k, k + 1, S_{k+1}, S_{k+2}, \ldots, S_N) + S_k \times E(\text{ATS}_{i,k}) \} \]
By running the above approach recursively, the optimal solution of the objective function can be obtained (Denardo, 1982, pp. 69-85, Smith, 1991, pp. 19-23, Warwick and Phelps, 1986)

Below is a ten-stage univariate example to illustrate the procedure. We follow the similar steps as the guideline in Section 2.2 except that we use the dynamic programming method to replace steps 3-5 in the optimization steps.

Without loss of generality, we assume $C=1$ in all stages, and $A$s are randomly generated from the parameter range of the previous examples: $[0.26, 3.38, 3.58, 2.79, 3.55, 0.39, 1.47, 0.65, 0.97, 1.24]$. The noise variances in all the stages are assumed to be 0.01.

Step 1: Model the process based on the state space model and the above parameters.

Step 2: Determine the fault set and number of charts
Five potential faults are placed in different stages. The fault set vector $F=[0.1, 0, 0, 0.1, 0.5, 0, 0.2, 0, 0]$. The eighth element is 0.2, which means a potential mean shift in the eighth stage. We assume that two charts will be used in the monitoring.

Step 3: Determine the chart allocation by optimization
Using the dynamic programming model we developed, the corresponding optimal solution is found when $S_2=S_{10}=1$, and other $S_j=0$. According to the results, we find that stages 2 and 10 are the best allocations for the charts.

2.7 Summary

In this chapter, we define a new problem for deciding which stage an SPC chart should be applied in the multistage process monitoring. The interrelationship among stages is considered into the decision by modeling the process with the state space model. We also formulate the objective function for the above decision based on the well-known noncentrality parameter in the MEWMA chart. We further extend the
derivation of the noncentrality parameter in the MEWMA chart from the single stage process to the multistage process.

Uncertainty impact of parameters in the state space model is studied to judge if the decision is reliable or not for both univariate and multivariate cases. Design of Experiments and regression are used in the analysis of multivariate cases.

Considering the real industrial applications, which usually involve multiple potential faults, the chart allocation problem is extended via dynamic programming optimization. The objective function is to obtain the minimum sum of expected out of control ATS for control charts.

Note that the proposed chart allocation strategy is a general approach and not restricted to any charting schemes, so the choice of charting schemes are not our focus. Besides the conventional charts such as MEWMA, more advanced charts can surely be applied whenever appropriate. Take the latent structure methods as an example: if the latent structure is known, it is feasible for us to calculate the mean shift propagation on the latent variables and so does the noncentrality parameters using the latent structure. By knowing the above information, the ARL can then be obtained depending on the type of charts used for monitoring the latent variables.

Moreover, we use real-life examples from automotive assembly processes to explore some properties of the proposed strategy and to demonstrate step-by-step how the strategies work in both univariate and multivariate cases. They show that sometimes the resulting strategy could be different from our intuition due to the inherent structure of the multistage process. We also consider the impact of parameter uncertainty. It can further help us to determine the chart allocation strategy when uncertainty is inevitable in practice.
Chapter 3

Chart Allocation strategy for Serial-Parallel Multistage Manufacturing Processes

The chart allocation strategy proposed in chapter 2 outlined how to place conventional control charts on appropriate allocations in a serial multistage processes. In this chapter, we further extend the strategy to a more complex and practical cases: serial parallel multistage manufacturing processes (SP-MMP).

Most of the works on monitoring MMP we mentioned are for serial MMP which has only one workstation at each stage. In reality, in order to meet the productivity and line balance requirements, more than one identical workstation would be placed in the same stage in a MMP. Such a process is called Serial-Parallel MMP (SP-MMP), which is a more general and realistic setting in the manufacturing process. Figure 3.1 shows a typical example of SP-MMP. The finished parts at stage 1 will be split as the inputs for the two identical workstations at the second stage. After the second stage, the two streams of parts will be converged to be the input of the third stage.

Some control charts and methods have been investigated to monitor multiple stream processes (Runger, Alt, and Montgomery, 1996; Nelson, 1986; Mortell, and Runger, 1995), however, most methods assume a single-stage process and do not take multistage complexity into consideration. Ignoring inherent structural information (stage-wise correlations) among stages could make the conventional SPC less effective and efficient. A very limited number of works focus on monitoring of the serial parallel MMP: Huang and Shi (2004) extended the state space modeling approach from serial MMP to SP-MMP, proposed dimension reduction techniques and analyzed their impact on the system diagnosability issues; Wu and Shamsuzzaman (2005) studied the design of integrated control charts for SP-MMP monitoring with the assumption that each workstation is under monitoring; Li and
Zhou (2007) proposed a chain graph method to represent the relationships among key product characteristics in SP-MMP.

In the previous chapter, we investigated how to use the interrelationship between stages to develop an appropriate chart allocation strategy to achieve best performance for a serial MMP due to the limitation of cost and resources in reality. The state space model is used to model the serial MMP, and noncentrality parameter and average time to signal (ATS) are used as the critical criterion for decision. If the process is constructed by several parallel identical MMP which never merge at one stage, we can treat the process as multiple serial MMP and determine their chart allocation strategies based on the results in Jin and Tsung (2008) separately.

We further extend the chart allocation strategy to the general SP-MMP cases in this chapter due to the fact that the SP-MMP is even more complex and practical and has specialities compared with S-MMP: The structure of SP-MMP is more complex than S-MMP; it has three special scenarios- coincidence, divergence, and convergence rather than one scenario in S-MMP. Due to the complex process structure, modifications of the state space model are needed in order to fit the model with SP-MMP. Besides the process model modifications, the fault propagation patterns are also different from what we derived in S-MMP due to the special process scenarios: In S-MMP, a mean shift will only cause mean shifts at the downstream stages along the propagation. Nevertheless, in SP-MMP, a mean shift fault at an upstream stage would cause not only the mean shift but also the change of variance at the downstream stages if the convergence scenario happens in between.

The remaining of this chapter is organized as the follows: We first introduce how to modify the state space model model in a state space form to describe a SP-MMP under different special scenarios, and the charting method in Section 3.1. Based on that, chart allocation strategies for the output monitoring method is proposed in Sections 3.2. In order to achieve the best performance for the whole process, the SP-MMP chart allocation problem is formulated into a dynamic programming
optimization problem. In Section 3.3, a univariate example is used to demonstrate how the chart allocation strategies can be applied and illustrate the efficiency of the strategies. Section 3.4 concludes this chapter with a summary of contributions of the research.

3.1. SP-MMP modeling and charting methods

3.1.1 SP-MMP modeling

The modeling of an SP-MMP is relevant to that of a serial MMP because a serial MMP is a subset of general SP-MMP. In a serial MMP, a stream-of-variation model in a state-space form has been applied successfully to describe the interrelationships and variation propagation at the process level of a multistage process (Ding et al., 2002, Jin and Tsung, 2008). With this model, physical and engineering knowledge can help to make the inherent structural information explicit. Yu et al. (2002a, 2002b) applied this model to rigid-part assembly processes. Djurdjanovic and Ni (2001), Huang et al. (2002) and Zhou et al. (2003) considered applications in machining processes.

We try to apply the state space model in the SP-MMP cases to describe the interrelationships among stages or workstations. Equation (3.1) presents the state space model to describe a serial MMP.

\[
\begin{align*}
y_k &= C_k x_k + w_k \\
x_k &= A_k x_{k-1} + U_k + v_k.
\end{align*}
\]

Two kinds of quality information are described in this equation. The first is the state vector, \( x_k \), such as the dimensional deviations of parts in an assembly process. The second is the observed quality information, \( y_k \), which is the quality measurement of the process output at the \( k \)th stage. Through a cascading process, these two kinds of information are transferred when a product is passed to its downstream stage. In addition, \( A_k \) denotes how the quality information in stage \( k-1 \) transfers to the quality information in stage \( k \). \( C_k \) indicates the relationship between the quality measurement, \( y_k \), and the state vector, \( x_k \), in stage \( k \). In practice, both \( A_k \) and \( C_k \) can be obtained from engineering knowledge and product information. Moreover,
$U_k$ represents a process fault or an out-of-control condition, such as an unacceptable fixture deviation. Inherent process noise is also considered: $v_k$ represents the process noise such as background disturbance and unmodeled errors, while $w_k$ is the measurement error, such as the sensor noise in the process.

However, because of the special properties of SP-MMP, we need to modify the process model under certain scenarios. We list three scenarios that are basic in an SP-MMP in Table 3.1 (Huang and Shi, 2004):

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coincidence:</td>
<td>Serial transformation through stage k-1 and k.</td>
</tr>
<tr>
<td>Divergence:</td>
<td>The outputs of stage k-1 are split at stage k</td>
</tr>
<tr>
<td>Convergence:</td>
<td>The outputs of workstations 1 and 2 of stage k-1 are merged at stage k</td>
</tr>
</tbody>
</table>

The corresponding modeling for each scenario is:

Scenario 1:
From the figure shown in Table 3.1, the coincidence scenario has the same propagation pattern as that of serial MMP. Therefore, the state space model in equation (3.1) can be introduced directly to describe the coincidence cases:

$$x_k = A_{k-1}x_{k-1} + v_k + U_k$$
$$y_k = C_kx_k + w_k$$.

Scenario 2:
In the divergence scenario, we have two streams for the two workstations at stage k, while we only have one stream at stage k-1. In order to fit this fact into the state space model, we use two dummy variables, $\tilde{x}_{k-1,1}$ and $\tilde{x}_{k-1,2}$, to model the quality characteristics at stage k-1:
\[
\begin{align*}
\begin{pmatrix} x_{k,1} \\ x_{k,2} \end{pmatrix} &= \begin{pmatrix} A_{k-1} & 0 \\ 0 & A_{k-1} \end{pmatrix} \begin{pmatrix} \tilde{x}_{k-1,1} \\ \tilde{x}_{k-1,2} \end{pmatrix} + \begin{pmatrix} v_{k,1} \\ v_{k,2} \end{pmatrix} + \begin{pmatrix} U_{k,1} \\ U_{k,2} \end{pmatrix} \\
\begin{pmatrix} y_{k,1} \\ y_{k,2} \end{pmatrix} &= \begin{pmatrix} C_k & 0 \\ 0 & C_k \end{pmatrix} \begin{pmatrix} x_{k,1} \\ x_{k,2} \end{pmatrix} + \begin{pmatrix} w_{k,1} \\ w_{k,2} \end{pmatrix},
\end{align*}
\] (3.2)

where \( \tilde{x}_{k-1,1} \) and \( \tilde{x}_{k-1,2} \) represent the artificial quality characteristics as the sources of \( x_{k,1} \) and \( x_{k,2} \), respectively. Moreover, \( \tilde{x}_{k-1,1} \) and \( \tilde{x}_{k-1,2} \) follow the same distribution, and \( x_{k-1} \) is the combination of these two. If a fault, say, a mean shift, occurs in stage \( k-1 \), it will impact both workstations in stage \( k \).

**Scenario 3:**

In the convergence scenario, stage \( k \) has more than one source of inputs, while the output has only one stream. We use the dummy variables \( \tilde{x}_{k,1}, \tilde{x}_{k,2}, \tilde{y}_{k,1}, \) and \( \tilde{y}_{k,2} \) to solve the problem of unequal numbers of streams of input and output:

\[
\begin{align*}
\begin{pmatrix} \tilde{x}_{k,1} \\ \tilde{x}_{k,2} \end{pmatrix} &= \begin{pmatrix} A_{k-1} & 0 \\ 0 & A_{k-1} \end{pmatrix} \begin{pmatrix} x_{k-1,1} \\ x_{k-1,2} \end{pmatrix} + \begin{pmatrix} v_{k,1} \\ v_{k,2} \end{pmatrix} + \begin{pmatrix} U_k \\ U_k \end{pmatrix} \\
\begin{pmatrix} \tilde{y}_{k,1} \\ \tilde{y}_{k,2} \end{pmatrix} &= \begin{pmatrix} C_k & 0 \\ 0 & C_k \end{pmatrix} \begin{pmatrix} \tilde{x}_{k,1} \\ \tilde{x}_{k,2} \end{pmatrix} + \begin{pmatrix} w_{k,1} \\ w_{k,2} \end{pmatrix},
\end{align*}
\] (3.3)

where \( x_{k-1,1} \) and \( x_{k-1,2} \) are the quality characteristics in the first and second workstations in stage \( k-1 \); \( \tilde{x}_{k,1} \) and \( \tilde{x}_{k,2} \) represent the transformed quality characteristics based on \( x_{k-1,1} \) and \( x_{k-1,2} \), respectively. \( v_{k,1} \) and \( v_{k,2} \) follow the same distribution, and so do \( w_{k,1} \) and \( w_{k,2} \). If the input is not separated in stage \( k \), the actual transformed quality characteristic at stage \( k \) are \( x_k \), which is a combination of \( \tilde{x}_{k,1} \) and \( \tilde{x}_{k,2} \). If no fault exists in the upstream stages of stage \( k \), \( \tilde{x}_{k,1} \) and \( \tilde{x}_{k,2} \) have the same distribution, which implies that the mixture distribution of \( x_k \) is the same as the distribution of \( \tilde{x}_{k,1} \) and \( \tilde{x}_{k,2} \). Therefore, equation (3.1) can be modified to:

\[
\begin{align*}
x_k &= A_{k-1} \tilde{x}_{k-1} + v_k + U_k \\
y_k &= C_k x_k + w_k,
\end{align*}
\] (3.4)
Where \( \tilde{x}_{k-1} \) follows the mixture distribution formed by \( x_{k-1,1} \) and \( x_{k-1,2} \).

Furthermore, \( \tilde{x}_{k-1} \), \( x_{k-1,1} \) and \( x_{k-1,2} \) share the same distribution.

We successfully modify the state space model to fit these three special cases. The block part example in Huang and Shi (2004) is used to illustrate the modeling procedure.

Since the parameters of this example, as seen in Huang and Shi (2003), cannot be found in any database, we use some made-up parameters for illustration. We assume the inputs of three workstations of stage 1 are all from \( x_0 \).
\[
\begin{pmatrix}
\bar{x}_{3,1} \\
\bar{x}_{3,2} \\
\bar{x}_{3,3}
\end{pmatrix}
= \begin{pmatrix}
A_2 & 0 & 0 \\
0 & A_2 & 0 \\
0 & 0 & A_2
\end{pmatrix}
\begin{pmatrix}
\bar{x}_{2,1} & \bar{x}_{2,2} \\
\bar{x}_{2,2} & \bar{x}_{2,2} \\
\bar{x}_{2,2} & \bar{x}_{2,2}
\end{pmatrix}
+ \begin{pmatrix}
\bar{v}_{3,1} \\
\bar{v}_{3,2} \\
\bar{v}_{3,3}
\end{pmatrix}
+ \begin{pmatrix}
\bar{U}_{3,1} \\
\bar{U}_{3,2} \\
\bar{U}_{3,3}
\end{pmatrix}
\]

\[
\begin{pmatrix}
\bar{y}_{3,1} \\
\bar{y}_{3,2} \\
\bar{y}_{3,3}
\end{pmatrix}
= \begin{pmatrix}
C_3 & 0 & 0 \\
0 & C_3 & 0 \\
0 & 0 & C_3
\end{pmatrix}
\begin{pmatrix}
\bar{x}_{3,1} \\
\bar{x}_{3,2} \\
\bar{x}_{3,3}
\end{pmatrix}
+ \begin{pmatrix}
\bar{w}_{3,1} \\
\bar{w}_{3,2} \\
\bar{w}_{3,3}
\end{pmatrix}
\]

### 3.1.2 Charting method

With the development of SPC techniques, various control charts have been invented to deal with different situations. Among all the charts, the Shewhart chart, the EWMA chart and CUSUM chart are commonly used to detect mean shifts in processes. However, in real situations, faults are not only related to mean shifts but also to changes of variance. To detect both the mean shift and change of variance efficiently, other control charts have been developed, which are popular in practice due to their efficiency and simplicity. These include X-bar & R, X-bar & S, and X-bar & $S^2$ charts. The latter two are usually used when the sample size is relatively large while the first chart can be used with small sample size. In this paper, we use the X-bar & $S^2$ chart as the charting method to monitor the workstations in the SP-MMP because X-bar & $S^2$ charts can result accurate estimation of both mean and variance. The control limits for X-bar & $S^2$ charts are:

If $x_1, x_2, \ldots, x_n$ is a sample of size n, the average and the square of the standard deviation of the sample are:

\[
\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}
\]
\[
S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}
\]

Let $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_m$, $S_1^2, S_2^2, \ldots, S_m^2$ and be the average and the square of the standard deviation of each sample. $\bar{x}$, which is the grand average, would be the best estimator of the mean (Montgomery, 2001), whereas $\bar{S}^2$ is an unbiased estimation of $\sigma^2$.
\[ \bar{X} = \frac{\bar{x}_1 + \bar{x}_2 + \ldots + \bar{x}_m}{m} \]

\[ \overline{S^2} = \frac{\sum_{i=1}^{m} S_i^2}{m}. \]

Then the control limits for both X-bar and \( S^2 \) charts are derived as follows:

The control limits for the X-bar chart are

\[ \text{UCL} = \bar{X} + k \sqrt{\overline{S^2}} \]

Center line = \( \bar{X} \) \hspace{1cm} (3.5)

\[ \text{LCL} = \bar{X} - k \sqrt{\overline{S^2}}, \]

where \( k \) determines the sigma level of the chart.

The control limits for the \( S^2 \) chart are

\[ \text{UCL} = \frac{\overline{S^2}}{n-1} \chi^2_{\alpha/2,n-1} \]

Center line = \( \overline{S^2} \) \hspace{1cm} (3.6)

\[ \text{LCL} = \frac{\overline{S^2}}{n-1} \chi^2_{1-(\alpha/2),n-1}, \]

where \( \chi^2_{\alpha/2,n-1} \) and \( \chi^2_{1-(\alpha/2),n-1} \) denote the upper and lower \( \alpha / 2 \) percentage points of the chi-square distribution with \( n \)-degrees of freedom. If \( \sigma^2 \) is known, it can be used to replace \( \overline{S^2} \) in the above equations (Montgomery, 2001). The performance of the charts, say, the out of control average run length (ARL), depends on the magnitude of the fault (a mean shift, a change of variance, or both) when the control limits are fixed.

Output monitoring charts are widely used in monitoring a multistage process. The outputs of a stage are obtained and used directly to plot the corresponding conventional control charts, such as X-bar & R, X-bar & S, and EWMA charts. Another monitoring method is residual monitoring charts. In this case, one may consider monitoring the residual after a model-based prediction instead of monitoring
the process output directly. Compared with residual monitoring, the output monitoring method has advantages in simple data collection and being prediction-free. Moreover, when we apply the residual monitoring in a MMP, the choice of residual base is another issue we need to consider. We will discuss the residual base problem together with the chart allocation in the next chapter. In this chapter, due to the simplicity and popularity of output monitoring in industry, we study how to develop the chart allocation strategy for SP-MMP with the output monitoring method.

### 3.2 The chart allocation strategy

We also use ATS as the main criterion to evaluate the strategy’s performance. According to its definition in previous chapter, we can see that ATS has a direct relationship with the fault propagation. Therefore, we first study how to obtain the mean shift propagation patterns for different scenarios in SP-MMP based on the original fault, process structure, and As and Cs in equation (3.1). In the following paragraphs, we present the derivations of the propagated faults for output monitoring in different situations.

As discussed in section 3.1, the fault propagation patterns in the first scenario are similar to the serial MMP case. The mean shift in the upstream stage can only cause propagated mean shifts in the downstream stages/workstations. The noncentrality parameter $\delta_n$, is used to denote the propagated mean shift patterns. Therefore, Jin and Tsung’s (2007) results on the chart allocation for serial MMP can be directly introduced in the coincidence scenario:

$$\delta_n = \frac{C_n^*(\prod_{k=n}^{i-1} A_k)U_i}{\sqrt{\sum_{j=2}^{n} [C_nA_n...A_i]_j^2 \sigma_j^2 + C_n^2 \sigma_1^2 + \sigma_2^2}}$$

where $\delta_n$ is always known as the noncentrality parameter, which measures the magnitude of the propagated mean shift in stage $n$; $\sigma_1^2$ and $\sigma_2^2$ are the noise levels in equation (3.1).
For the divergence scenario, even when we have multiple workstations at the downstream stage, all of the inputs are from one source at the upstream stage. If there is a fault at the upstream stage, the impacts on all the downstream workstations are the same. Therefore, in terms of fault propagation patterns, we can treat all the downstream workstations as a big artificial workstation because they share the same mean shift fault. Due to the above fact, the mean shift propagation pattern under this scenario can be modeled as a serial MMP again.

Nevertheless, the convergence scenario is different from what we have proposed due to the fact that if there is a fault that occurs in one of the workstations in stage i (see Fig. 3.2), the mean shift will not only cause the mean to be off the target in the downstream stage but it will also increase the variance. Now, we will show how to derive the propagated fault patterns for this particular case.

![Figure 3.2. Example of convergence in SP-MMP](image)

For simplicity, we suppose that there are two workstations in stage i, and workstation 1 has a mean shift of $U_{i,1}$. The proportions of input to stage i+1 from workstation 1 and workstation 2 are p and p-1, respectively. The state space model for stage i is:

\[
\begin{align*}
    x_{i,1} &= A_{i-1} x_{i-1} + U_{i,1} + v \\
    y_{i,1} &= C_i x_{i,1} + w \\
    x_{i,2} &= A_{i-1} x_{i-1} + v \\
    y_{i,2} &= C_i x_{i,2} + w.
\end{align*}
\]

According to the discussion in previous sections, the state space model of stage i+1 becomes:
The grand input stage $i+1-x_{i+1}$ is the combinations of $\tilde{x}_{i+1,1}$ and $\tilde{x}_{i+1,2}$, and $y_{i+1}$ is the total of $\tilde{y}_{i+1,1}$ and $\tilde{y}_{i+1,2}$.

According to the theory of mixture normal distribution, the mean and variance of the mixture distribution from two normal distributions, $N(\mu_1, \sigma)$ and $N(\mu_2, \sigma)$, with mixed proportions of $p$ and $1-p$, respectively, are:

Mean\(= p\mu_1 + (1 - p)\mu_2 \)

Variance\(= p(1-p)(\mu_1 - \mu_2)^2 + \sigma^2. \)

Therefore, the expectations of the propagated fault patterns are measured:

\[
E(y_{i+1}) = pC_{i+1}A_i U_j
\]

\[
V(y_{i+1}) = C^2_{i+1}A^2_i v^2 + C^2_{i+1}v^2 + w^2 + p(1-p)C^2_{i+1}A^2_i U_j^2
\]

\[
V_0(y_{i+1}) = C^2_{i+1}A^2_i v^2 + C^2_{i+1}v^2 + w^2
\]

As shown above, the mean shift in the upstream stage not only causes a mean shift but it also changes the variance in the downstream stage. Thus, we use another index, $\Delta$, which describes the change in variance, together with $\delta$ to depict the fault propagation:

\[
\delta = \frac{pC_{i+1}A_i U_j}{\sqrt{C^2_{i+1}A^2_i v^2 + C^2_{i+1}v^2 + w^2}} \quad (3.7)
\]
If there is no convergence after stage i+1, the general propagated fault patterns in the downstream stage n are derived as:

\[ E(y_n) = pC_nA_{n-1}...A_{U_i} \]

\[ V(y_{i+1}) = C^2_n \sum_{j=1}^{n-2} A^2_{n-1} A^2_j v^2 + C^2_n v^2 + w^2 + p(1-p)(C_nA_{n-1}...A_{U_i})^2 \]

\[ V_0(y_{i+1}) = C^2_n \sum_{j=1}^{n-2} A^2_{n-1} A^2_j v^2 + C^2_n v^2 + w^2 \cdot \]

The noncentrality parameter is:

\[ \delta = \frac{pC_nA_{n-1}...A_{U_i}}{\sqrt{C^2_n \sum_{j=1}^{n-2} A^2_{n-1} A^2_j v^2 + C^2_n v^2 + w^2}} \quad (3.8) \]

\[ \Delta = \frac{\sigma_1}{\sigma_0} = \sqrt{\frac{C^2_n \sum_{j=1}^{n-2} A^2_{n-1} A^2_j v^2 + C^2_n v^2 + w^2 + p(1-p)(C_nA_{n-1}...A_{U_i})^2}{C^2_n \sum_{j=1}^{n-2} A^2_{n-1} A^2_j v^2 + C^2_n v^2 + w^2}} \]

However, if another convergence occurs downstream, we need to further investigate the propagated fault patterns.

Figure 3.3 shows the situation in which convergence occurs in both stages 2 and 3. In such cases, we need to identify the percentage of faulty products that flows into each workstation. In the example, the proportion of faulty products at workstation 1 of stage 3 is \( p = p_1p_2 = 0.25 \). We can apply this new \( p \) into the above equations to obtain \( \delta \) and \( \Delta \).
As soon as we obtain the propagation patterns based on the above analysis in different scenarios, we can use $\delta$ and $\Delta$ to estimate the out of control ARL and ATS for the control charts we use in the process.

After discussing the fault propagation for one fault, we need to notice that with the development of technology, it is quite possible that we have a fault set including multiple potential faults. Two factors are vital in the chart allocation problem: first is the number of faults we need to focus on ($n$) and second is the number of charts in the process ($m$). If the number of potential faults is more than one, the chart allocation strategy is to ensure the best performance for the whole fault set $\{F_1, F_2, \ldots, F_n\}$ rather than to choose an allocation that has the smallest ATS for one particular potential fault.

To formulate the chart allocation strategy for cases of known fault sets with multiple potential faults, we need to solve the optimization problem as follows:

Objective function:

$$\min \frac{1}{m} \sum_{k=1}^{N} \sum_{j=1}^{k} S_{k,j} \times E(ATS_{i,k,j})$$  \hspace{1cm} (3.9)

where $S_{k,j}$ is the binary decision variable on the $j$th workstation of stage $k$ which has $kn$ workstations in total, $S_{k,j}=1$ meaning a chart allocated at this workstation.

$E(ATS_{i,k,j}) = EE[ATS_{i,k,j} | ith \text{ fault in the upstream occurs}] = \sum_{j=1}^{k} \overline{p}_i ATS_{i,k,j}$ and $\overline{p}_i$ is the probability of the $i$th fault conditional on the faults in the upstream of stage $k$;
\( ATS_{i,k,j} \) is the ATS of the ith fault on the jth workstation of the kth stage. The above objective function subject to:

\[
S_{k,j} = 1 \text{ or } 0
\]

\[
\sum_{k=1}^{N} \sum_{j=1}^{k_n} S_{k,j} = m \quad (3.10)
\]

\[
\sum_{k=g}^{N} \sum_{j=1}^{k_n} S_{k,j} \geq 1,
\]

\[
0 \leq m \leq N
\]

Faults set= \{F_1,F_2,...,F_n\}.

N is the total number of operations/stages in SP-MMP. g is the stage index of the last potential fault. This constraint is to ensure that no potential fault is ignored in the chart allocation strategy.

**Case 1: m is fixed**

In this case the known resources are assigned to the monitoring task. Therefore, the number of charts in an MMP is fixed. When m equals a constant, the above optimization problem becomes a separable integer programming problem. Due to the fact that the allocation decisions on stages are sequentially related, which is a typical dynamic programming problem (White, 1983), the dynamic programming method is used to obtain the optimal solution.

We shall use the notation \( C(S_{k,1},...,S_{k,j},...,S_{k,k_n},k) = \sum_{j=1}^{k_n} S_{k,j} \times E(ATS_{i,k,j}) \), which can be treated as the cost function only in stage k. The state in stage k is the remaining number of charts, c. The optimal cost (sum of expected ATS) from stage k onwards to the end is:

\[
f(c,k,S_{k,1},...,S_{k,k_a},...,S_{N,1},...,S_{N,N_n}) = \min \sum_{m=k}^{N} \sum_{j=1}^{m_n} S_{m,j} \times E(ATS_{i,m,j})
\]

The above function is subjected to the following constraints:

\[
S_{k,j} = 1 \text{ or } 0
\]
With the above definitions, the dynamic programming recursion can be formulated as:

\[ f(c, k, S_{k,1}, \ldots, S_{k,i}, \ldots, S_{N,k}) = \min \{ f(c - \sum_{j=1}^{k} S_{k,j}, k + 1, S_{k+1,1}, \ldots, S_{N,N}) + \sum_{j=1}^{k} S_{k,j} \times E(ATS_{k,j}) \} \]

By running the above approach recursively, the optimal solution of the objective function can be obtained (White, 1983, Warwick and Phelps, 1986).

**Case 2: m is unknown**

This is a more general case that we may face in reality. Manufactures may have little idea of how many resources should be placed in an MMP in order to achieve the best performance. The m will become a variable in the above formulation to determine not only the chart allocation but also the appropriate number of charts in an MMP. However, this leads the problem to the nonlinear integer programming problem, more precisely, a nonseparable integer programming problem that is certainly more difficult to solve compared with the previous linear integer programming problem. The branch-and-bound method based on continuous relaxation, Lagrangian Decomposition method and the Monotone Integer Programming method are commonly used in dealing with such problems. Extensive research has been done in the nonlinear integer programming field (Cooper, 1981, Tawarmalani and Sahinidis, 2002, and Omprakash and Ravindran, 1985). A thorough review can be found in Li and Sun (2006). In our case, we will use Branch-and-Bound method which is commonly used in different nonlinear integer problems (Cabot and Erenguc, 1986, Gupta and Ravindran, 1985, Leyffer, 2001) to solve the problem in the numerical example; the branch-and-bound method using lower bound generated by continuous relaxation can be outlined as follows:

First, an optimal solution, x0, of the continuous relaxation problem of (P) is obtained. If x0 is an integral, then the problem ends with optimal solution-x0. If x0 is not an
integer, we set \( x_i^0 \) to be a component variable of \( x_0 \). We generate two new sub-problems by adding new constraints \( x_i \leq \lfloor x_i^0 \rfloor \) and \( x_i \geq \lceil x_i^0 \rceil + 1 \), respectively. \( \lfloor x_i^0 \rfloor \) is the maximum integer less than or equal to \( x_i^0 \). At each iteration, we choose one of the newly generated sub-problems to solve. If its optimal solution is an integral and its objective value is better than that of the incumbent, then it becomes the new incumbent. The sub-problem is fathomed from further consideration if it satisfies one of the following conditions: 1) an optimal integer solution is obtained in the corresponding continuous relaxation sub-problem; 2) the optimal value of the continuous relaxation is larger than or equal to the upper bound of the current incumbent; 3) the continuous relaxation problem is infeasible. Otherwise the sub-problem is divided again, and the process is repeated until no sub-problem remains to be solved.

We summarize the development of the SP-MMP chart allocation strategy for the output monitoring as follows. First of all, we need to model the SP-MMP by drawing the work flow structure among workstations and stages and depicting the interrelationships with the modified state space model. Then, we can obtain the fault set, which we want to focus on, including the approximate direction and magnitude from historical data. For the faults of interest, we calculate \( \delta \) and \( \Delta \) for all workstations downstream of the faulty workstations (including the faulty workstation itself) based on the state space model. After \( \delta \) and \( \Delta \) are obtained, the out-of-control ARL and also ATS can be calculated for each downstream workstation. With all the information in hand, we are able to determine the chart allocation strategy by introducing the data we have into the formulated optimization problem. After the chart allocation is determined, if there is some new fault of interest, we may add it into the original fault set and run the procedure again to have an updated chart allocation strategy.

3.3. Numerical example

In this section, the multistage automobile hood fit example from Lawless et al. (1999) is used to illustrate how to apply the chart allocation strategy procedure in
real applications. The results indicate that the monitoring performance is at least as good as the common-sense chart allocation and, in many cases, better, by using the proposed strategy.

This is a four-stage process including HANG, PAINT, HARDWARE, and FINESSE operations. Lawless et al. (1999) analyzed the variation in a multistage manufacturing process based on a transmission model. Using appropriate transformation, the model can be rewritten as a state space model as in Xiang and Tsung (2006).

\[
y_i^* = \alpha_i x_i^* + \varepsilon_i, \quad i = 1, \ldots, 4
\]
\[
x_{i+1}^* = \beta_i x_i^* + \varepsilon_i,
\]

where \((\beta_1, \beta_2, \beta_3) = (1.15, 0.98, 1.06), \quad (\hat{\sigma}_{\alpha_1}, \hat{\sigma}_{\alpha_2}, \hat{\sigma}_{\alpha_3}) = (0.13, 0.11, 0.2), \quad \text{and} \quad \alpha_i = 1 \quad \text{for all} \ i\). Without loss of generality, we assume that \(\sigma_{\varepsilon} = 0.1\).

Originally, this is a serial MMP. However, in reality, it is always inevitable that more than one workstation is used in one stage in order to balance the efficiency among different stages. Here, we do some modification to the process. We assume that because of efficiency needs, two workstations should be involved at the PAINT and HARDWARE operations. In this example, we have divergence and multiple convergences, which represent the typical characteristics of an SP-MMP. The whole structure of the process is shown below:

![Figure 3.4 Hood fit process](image-url)
where $p_1=p_2=0.5$.

The process is modeled using the state space model:

**Stage 2:**

\[
\begin{align*}
x_{2,1} &= A_1x_1 + v \\
y_{2,1} &= C_1x_{2,1} + w \\
x_{2,2} &= A_1x_1 + v \\
y_{2,2} &= C_1x_{2,2} + w
\end{align*}
\]

**Stage 3:**

\[
\begin{align*}
\begin{pmatrix} \tilde{x}_{3,1,1} \\ \tilde{x}_{3,1,2} \end{pmatrix} &= \begin{pmatrix} A_2 & 0 \\ 0 & A_2 \end{pmatrix} \begin{pmatrix} x_{2,1} \\ x_{2,2} \end{pmatrix} + \begin{pmatrix} v \\ v \end{pmatrix} \\
\begin{pmatrix} \tilde{y}_{3,1,1} \\ \tilde{y}_{3,1,2} \end{pmatrix} &= \begin{pmatrix} C_3 & 0 \\ 0 & C_3 \end{pmatrix} \begin{pmatrix} \tilde{x}_{3,1,1} \\ \tilde{x}_{3,1,2} \end{pmatrix} + \begin{pmatrix} w \\ w \end{pmatrix} \\
x_{3,2} &= A_2x_{2,2} + v \\
y_{3,2} &= C_3x_{3,2} + w
\end{align*}
\]

**Stage 4:**

\[
\begin{align*}
\begin{pmatrix} \tilde{x}_{4,1} \\ \tilde{x}_{4,2} \end{pmatrix} &= \begin{pmatrix} A_1 & 0 \\ 0 & A_1 \end{pmatrix} \begin{pmatrix} x_{3,1} \\ x_{3,2} \end{pmatrix} + \begin{pmatrix} v \\ v \end{pmatrix} \\
\begin{pmatrix} \tilde{y}_{4,1} \\ \tilde{y}_{4,2} \end{pmatrix} &= \begin{pmatrix} C_4 & 0 \\ 0 & C_4 \end{pmatrix} \begin{pmatrix} \tilde{x}_{4,1} \\ \tilde{x}_{4,2} \end{pmatrix} + \begin{pmatrix} w \\ w \end{pmatrix}
\end{align*}
\]

After we model the process, we can identify the potential fault we want to consider. We assume that there is a potential mean shift fault $= 1 \sigma_i$ at workstation 2 at the PAINT operation. For this particular potential fault, we can use the fault propagation patterns we have investigated in the previous sections for different scenarios in the process to calculate $\delta$ and $\Delta$ for all the downstream workstations of the faulty workstations (including the faulty workstation itself) based on the state space model.

For the output monitoring, corresponding fault propagation patterns and ARLs are shown in table 3.2
Table 3.2 - Fault propagation patterns and corresponding ARL for output monitoring

<table>
<thead>
<tr>
<th>Operations &amp; workstations</th>
<th>HANG, workstation 1</th>
<th>PAINT, workstation 1</th>
<th>PAINT, workstation 2</th>
<th>HARDWARE, workstation 1</th>
<th>HARDWARE, workstation 2</th>
<th>FINESSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \delta ) and ( \Delta )</td>
<td>( \delta = 0 )</td>
<td>( \delta = 0 )</td>
<td>( \delta = 0.793 )</td>
<td>( \delta = 0.324 )</td>
<td>( \delta = 0.65 )</td>
<td>( \delta = 0.353 )</td>
</tr>
<tr>
<td>( \Delta )</td>
<td>( \Delta = 0 )</td>
<td>( \Delta = 0 )</td>
<td>( \Delta = 0 )</td>
<td>( \Delta = 1.05 )</td>
<td>( \Delta = 0 )</td>
<td>( \Delta = 1.021 )</td>
</tr>
<tr>
<td>ARL</td>
<td>200</td>
<td>200</td>
<td>6.53</td>
<td>13.08</td>
<td>15.31</td>
<td>14.41</td>
</tr>
</tbody>
</table>

As shown in the above table, if we do not consider the processing delay in the process, we can conclude that the faulty workstation itself, workstation 2 in the PAINT operation, is the best place to monitor this potential fault. Because the best allocation is the faulty workstation, no matter how long the processing delay is, it will not change the conclusion. However, we should notice the fact that it is possible that the best allocation is not the faulty workstation but some downstream workstation. Examples can be found in Jin and Tsung (2008). Moreover, the magnitude of the original mean shift fault will not change the relative order of \( \delta \) and \( \Delta \) and the chart allocation decision would remain the same if the processing delay is not very long. The underlying reason for this phenomenon is that the \( \delta \) and \( \Delta \) are proportional to the magnitude of the fault, which means that the relative order of the \( \delta \) s of different stages does not change nor does the order of the ARLs.

If the number of potential faults is greater than 1, we use the dynamic programming method developed in Section 3.2 to determine the chart allocation strategy. Suppose that we have the following fault set:

\[
U_{2,1} = 0.13 \\
U_{2,2} = 0.13 \\
U_{3,1} = 0.11 \\
U_{3,2} = 0.11
\]

We further assume that all the potential faults have equal occurrence probability.

**Case 1: \( m \) is fixed**
Without loss of generality, we suppose that we only have the resources to set up two control charts in the process. $\bar{X}$ and $S^2$ charts are applied to monitor the process. The corresponding fault propagation patterns and ARLs are shown in table 3.3 for output monitoring.

Table 3.3-ARL of multiple potential faults for output monitoring

<table>
<thead>
<tr>
<th>Operations&amp; workstations</th>
<th>HANG</th>
<th>PAINT, workstation 1</th>
<th>PAINT, workstation 2</th>
<th>HARDWARE, workstation 1</th>
<th>HARDWARE, workstation 2</th>
<th>FINESSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$U_{2,1}$</td>
<td>200</td>
<td>6.53</td>
<td>200</td>
<td>13.08</td>
<td>200</td>
<td>17.59</td>
</tr>
<tr>
<td>$U_{2,2}$</td>
<td>200</td>
<td>200</td>
<td>6.53</td>
<td>13.08</td>
<td>15.31</td>
<td>14.41</td>
</tr>
<tr>
<td>$U_{3,1}$</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>12.03</td>
<td>200</td>
<td>16.96</td>
</tr>
<tr>
<td>$U_{3,2}$</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>12.03</td>
<td>16.96</td>
</tr>
</tbody>
</table>

To obtain the corresponding ATS for the above, we assume the delay (processing time) at each station is 5 min and the time interval between runs is 1 min. Therefore, the corresponding ATS are in table 3.4:

Table 3.4-ATS of multiple potential faults for output monitoring

<table>
<thead>
<tr>
<th>Operations&amp; workstations</th>
<th>HANG</th>
<th>PAINT, workstation 1</th>
<th>PAINT, workstation 2</th>
<th>HARDWARE, workstation 1</th>
<th>HARDWARE, workstation 2</th>
<th>FINESSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$U_{2,1}$</td>
<td>200</td>
<td>6.53</td>
<td>200</td>
<td>18.08</td>
<td>200</td>
<td>27.59</td>
</tr>
<tr>
<td>$U_{2,2}$</td>
<td>200</td>
<td>200</td>
<td>6.53</td>
<td>18.08</td>
<td>20.31</td>
<td>24.41</td>
</tr>
<tr>
<td>$U_{3,1}$</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>12.03</td>
<td>200</td>
<td>21.96</td>
</tr>
<tr>
<td>$U_{3,2}$</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>12.03</td>
<td>21.96</td>
</tr>
</tbody>
</table>

By introducing the above ATS into the dynamic programming function we
discussed, the chart allocation is determined as the FINESSE operation and one of the workstations in the PAINT operation.

**Case 2: m is unknown**

The ATS tables are the same as in case 1. m becomes a variable. Table 3.5 shows the optimal solution solved by LINGO 8.0.

| Table 3.5-Optimal solution of the chart allocation strategy for the hood fit process |
|-----------------------------------------------|---------------|---------------|---------------|---------------|
| Stage       | Workstation | Optimal Chart Allocation variables | Optimal Chart Allocation variables when m=3 | |
| HANG        | 1           | $S_{1,1} = 0$ | $S_{1,1} = 0$ | |
| PAINT       | 1           | $S_{2,1} = 1$ | $S_{2,1} = 1$ | |
|             | 2           | $S_{2,2} = 1$ | $S_{2,2} = 1$ | |
| HARDWARE    | 1           | $S_{3,1} = 1$ | $S_{3,1} = 0$ | |
|             | 2           | $S_{3,2} = 1$ | $S_{3,2} = 0$ | |
| FINESSE     | 1           | $S_{4,1} = 0$ | $S_{4,1} = 1$ | |

The optimal solution shows that if m=4, allocating the charts to the four potential faulty workstations, workstation 1 and 2 at the PAINT operation, and workstation 1 and 2 at the HARDWARE operation, will achieve the best performance according to the objective function we defined. The optimal value is 11.32. If fewer resources are obtained, say, at most three control charts can be allocated in the process, the optimal solution (12.35) is achieved by allocating charts to workstation 1 and 2 at the PAINT operation and the final stage. In this case, due to the small number of stages, it is easy for us to obtain the optimal solution by enumerating all the feasible solutions; however, when the number of stages becomes large, algorithms from nonlinear integer programming, such as the branch-and-bound method, can be applied to reduce the complexity of the calculation. Details of these algorithms can be found in Li and Sun (2006).
3.4. Summary

In this chapter, we develop a rational chart allocation strategy for serial parallel-multistage processes based on the criterion of the average time to signal. This chart allocation strategy is a general approach and not restricted to any charting schemes, so the choice of charting schemes is not our focus. Besides the conventional charts such as $\bar{X}$ and $S^2$ charts, more advanced charts can certainly be used whenever appropriate. The state space model is modified to describe the three special scenarios in SP-MMP. We study mean shift fault propagation in order to obtain the corresponding ARL and ATS. Moreover, we further formulate the objective function for multiple potential faults with resource constraints. Dynamic programming is used to solve the above optimization problem. We also extend the strategy to determine how many resources should be placed in SP-MMP in terms of the number of control charts.

We use a real-life example from the automobile hood assembly processes to explore some properties of the proposed strategy and to demonstrate step-by-step how the strategy works. Our strategy achieves good performance on the entire fault set with limited resources.

In this chapter, only univariate cases are studied. However, the results can be easily extended to multivariate cases using multivariate SPC charts, such as $T^2$ charts and MEWMA charts. The output monitoring is considered in this paper, nevertheless, the authors think that it is possible to obtain better monitoring performance by using both output and residual monitoring methods together in a multistage process.
Chapter 4
Integrated Monitoring Strategy for MMP

4.1 Introduction

There is quite a lot research on how to improve the monitoring and diagnosis techniques for MMP. In the previous chapters, we discussed the chart allocation strategies for two different types of MMP (S-MMP and SP-MMP) respectively, and the strategies are determined rationally based on the interrelationship among stages to maximize the efficiency of conventional control charts. However, very few works on the whole monitoring strategy for MMP, which includes both the charting types and chart allocations. In this chapter, we propose an integrated monitoring strategy to help people to decide where to place charts and which kind to place in a serial MMP.

Besides the allocation of charts, the charting types, which are determined by the data source of the chart, are also important to MMP monitoring. We divided the chart types into two kinds: 1. Output monitoring, which means the control charts are plotted based on the real outputs of stages; 2. Residual monitoring, which represents the control charts based on the residual of the real outputs and the prediction of outputs based on previous stage’s observations. The former one is simple to apply but hard to provide useful diagnosis information. The later one is more complicated as it requires the prediction based on previous stage’s output, however, it shows useful information about the faulty stage. In this integrated monitoring strategy, we take the interrelationship information among stages into our consideration to make the decision rational. The interrelationship is described using the state space model shown in chapter 2. The fault propagation patterns results in chapter 2 are also used here. Through this chapter, we use X-bar and $S^2$ chart to demonstrate how to apply the strategy to real applications. The strategy is formulated into a nonlinear optimization problem; moreover, it can be modified to a max-min problem which is possible to be linearized and much easier to obtain a good solution compared to the original nonlinear problem. Based on the proposed strategy, we are able to make scientific monitoring decisions to achieve quicker detection for the whole potential
fault set. A hood assembly example is used to demonstrate the applications of the chart allocation strategy. Extensions are also discussed at the end of this chapter.

4.2 Problem formulation

In the integrated monitoring strategy, we consider two main monitoring concerns:

1. The chart allocation
2. The charting type

Jin and Tsung (2008a) and (2008b) have investigated the chart allocation strategies for serial MMP and serial parallel MMP with the consideration of interrelationships among stages. Average time to signal (ATS) are used as the main criterion to evaluate the strategies.

4.2.1 Review of chart allocation strategy

In order to describe the interrelationships, the state space model in chapter 2 is applied together with engineering and physical knowledge. Potential fault patterns can be obtained from historical data and experience. The multiple potential faults form a fault set for the process. After knowing the potential fault patterns, for each fault, the noncentrality parameters which are explained in the previous section are calculated for all the downstream stages based on propagated fault patterns and the state space model.

For potential fault \( U_i \), if output charting type is used, the noncentrality parameter on stage \( n \) is:

\[
\delta_n = \frac{C_n \cdot \prod_{k=n}^{i+1} A_k U_i}{\sqrt{\sum_{i=2}^{N} (C_{n} A_{i} \ldots A_{j})^2 \sigma_1^2 + C_n^2 \sigma_1^2 + \sigma_2^2}}.
\] (4.1)

where \( i \leq n \leq N \). \( N \) is the total number of stages.

If residual chart type is applied, the results are below:

\[
\delta_n = \sqrt{\frac{\sigma_{\text{Y}_{(n)} \text{_res}}^2 \mu_{\text{Y}_{(n)} \text{_res}}^2}{\sigma_{\text{Y}_{(n)} \text{_res}}^2 \mu_{\text{Y}_{(n)} \text{_res}}^2}},
\] (4.2)

where \( \mu_{\text{Y}_{(n)} \text{_res}} = C(n) \prod_{k=n}^{i+1} A(k) U(i) \), and
\[ \sum_{Y(n)_{res}} = \sum_{k=1}^{n} [C(n)A(n)...A(k)]^2 \sigma_1^2 + C^2(n)\sigma_1^2 \]
\[ + \sigma_2^2 + [C(n)A(n)...A(j+1)C^{-1}(j)]^2 \sigma_2^2, \quad (4.3) \]

\( \delta_n \) has an inverse proportion to the out-of-control ARL in the control chart. ARL is a popular way to evaluate the monitoring performance, however, the situation with multi-stages is different, as the processing (delay) time between stages should be considered in the chart evaluation. In this chapter, ATS criterion is also used as the main performance measure as we have done in the previous two chapters. Based on the corresponding ARL of \( \delta_n \), we can calculate the ATS accordingly if the processing delay is known. When there is single fault in the potential fault set, the chart allocation strategy can be determined by choosing the stage that has the minimal ATS.

When there are multiple potential faults in the fault set, the chart allocation strategy is to minimize the sum of expected ATS of every chart. Therefore, the strategy is formulated to an optimization problem if the number of charts is known:

Decision variable:

\( S_j \) - binary variable, 1 means a chart allocated at stage j.

Objective function:

In order to consider the performance of the chart allocation strategy on the whole fault set, we try to minimize the sum of expected ATS of all the charts:

\[ \min \frac{1}{m} \sum_{k=1}^{N} S_k \times E(ATS_{i,k}) \]

where the expected ATS of a chart on stage k is defined as

\[ E(ATS_{i,k}) = EE[ATS_{i,k} \text{ ith fault in the upstream occurs}] = \sum_{i=1}^{j} p_i ATS_{i,k} \]

and \( p_i \) is the probability of ith fault conditional on the faults in the upstream of stage k; \( ATS_{i,k} \) is the ATS of the ith fault on the kth stage.

Subject to:

\[ S_j = 1 \text{ or } 0 \]

\[ \sum_{j=1}^{m} S_j = m \]

\[ m \leq N \]
\[
\sum_{j=g}^{n} S_j \geq 1',
\]

Faults set= \{F_1, F_2, ..., F_n\}

g is the stage index of the last potential fault. This constraint is to ensure that no potential fault is ignored in the chart allocation strategy. We then apply the dynamic programming method to obtain the optimal solution.

### 4.2.2 Integrated monitoring strategy formulation

When practitioners set up a monitoring system for a MMP, besides the chart allocation, the charting type is also one of their main concerns. As we have mentioned in the introduction, the charting types can be divided into two groups: 1. Output monitoring; 2. Residual monitoring.

Due to the fact that both of them have their own advantages and disadvantages, it is difficult to say that we can abandon any of them in MMP monitoring. Therefore, we try to consider both these two charting types and chart allocation together in our integrated monitoring strategy. In the final solution of the integrated monitoring strategy, we try to answer the following two questions for each stage: 1. should we place a control chart on this stage? 2. If yes, should we use output monitoring chart or residual monitoring chart?

As we can see, the main difficulty comes from the second question. For this question, not only need we decide the charting type, but also we need to determine the residual base if residual monitoring is applied. The residual base of a stage can be any upstream stage of its own. Due to this fact, we have \(n+1\) options for stage \(n\) (\(n-1\) different residual based residual charts, one output monitoring chart, and no chart option). This causes the formulation of the problem more complex than just considering chart allocation alone. Here, we modify the formulation of chart allocation strategy (Jin and Tsung, 2008) to match the requirements of the integrated monitoring strategy:

First, we identify the possible decision space of the problem:
Table 4.1. The decision space of integrated monitoring strategy

<table>
<thead>
<tr>
<th>Stage &amp; Monitoring method</th>
<th>1</th>
<th>2</th>
<th>…</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>No chart</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Output monitoring</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Residual based on i-1 stage</td>
<td>NA</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>…</td>
<td>NA</td>
<td>NA</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Residual based on N-1 stage</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>√</td>
</tr>
</tbody>
</table>

As we can see from the above table, the number of possible decisions for stage is a variable itself, and it changes alone with the stage index. Therefore, it is hard for us to describe the decision at one stage by using one decision variable as what we did in the chart allocation problem. As we can see from this decision space, the fixed charting methods we used in the previous chapters are special cases in this decision space. Thus using the integrated monitoring strategy can obtain equivalent or better results than only adopting chart allocation strategy in a MMP. The following decision variables are re-defined for the integrated monitoring strategy.

Decision variable:

\[ m, \text{ number of charts in the process, integer variable.} \]

\[ S_{k,j}, \ 1 \leq j \leq k \]

For \( j \) is less than \( k \), \( S_{k,j} \) denotes the residual chart based on stage \( j \)'s outputs. If \( j = k \), \( S_{k,k} \) means output monitoring chart. \( S_{k,j} \) are binary variables, \( 1 \) means a chart allocated at stage \( k \). Each possible decision has a corresponding ATS for the fault set, and the objective is also to minimize the sum of expected ATS of all the charts. Therefore, the objective function is modified as:

\[
\min \frac{1}{m} \sum_{k=1}^{N} \sum_{j=1}^{k} S_{k,j} \times E(ATS_{i,k,j})
\]

where the expected ATS of decision \( S_{k,j} \) is defined as

\[
E(ATS_{i,k,j}) = EE[ATS_{i,k,j} | ith \text{ fault in the upstream occurs}] = \sum_{j=1}^{N} \overline{p}_{j} ATS_{i,k,j}
\]
u is the total number of faults in the upstream and $\bar{p}_i$ is the probability of ith fault conditional on the faults in the upstream of stage k; $ATS_{i,k,j}$ is the ATS of the ith fault if decision $S_{k,j}$ is adopted.

Subject to:

$$S_{k,j} = 1 \text{ or } 0$$

$$\sum_{j=1}^{k} S_{k,j} = 1 \quad \text{for } 1 \leq k \leq N$$

$$\sum_{k=1}^{N} \sum_{j=1}^{k} S_{k,j} = m$$

$$m \leq N$$

$$\sum_{k=g}^{N} \sum_{j=1}^{k} S_{k,j} \geq 1$$

Faults set = $\{F_1, F_2, \ldots, F_n\}$

The second constraint is to ensure that only one chart is placed on each stage. The final constraint is to make sure that all the potential faults are under monitoring.

It is noticed that the above optimization problem is a nonlinear integer programming problem, more precisely, a non separable integer programming problem. Branch-and-Bound method based on continuous relaxation, Lagrangian Decomposition method and Monotone Integer Programming method are commonly used in dealing with such problems (Li and Sun, 2006). In this chapter, we use Branch-and-Bound method to solve the optimization problem in the following sections.

**4.3 Integrated monitoring strategy max-min problem formulation**

In this section, we introduce a new approach to formulate the integrated monitoring strategy problem. It is formulated and solved by using linear integer programming method to minimize the maximum out of control ATS of faults. It is logical to choose the monitoring strategy which minimizes the maximum ATS due to the fact that the
bottleneck of monitoring strategy is determined by the fault with maximum ATS. Therefore, decreasing the maximum ATS can increase the monitoring performance efficiently.

4.3.1 Max-min optimization review
The max-min design method is applied and implemented in various areas, such as network, economics, and reliability system design. However, such method has not been implemented for the integrated monitoring strategy optimization, while the applications of this method are successful in other fields.

In network design area, Misra and Banerjee (2002) investigated a power-aware routing algorithm by using the max-min formulation in order to increase the operational life time of multi-hop wireless networks. With the max-min formulation, they are able to find the path which has the largest packet capacity at a node with the smallest residual packet transmission capacity. Badr (1990) developed an automated network design tool which helps people to choose the capacities to minimize the maximum time delay in a distributed network based on the max-min formulation.

Misra (1991) applied the min-max approach together with a direct search technique to solve multiple criteria reliability design problem. Marquez, et. al (2004) formulated the redundancy allocation problem for serial-parallel system into a max-min approach. With the max-min approach, the problem is transfer to a linear program which is easier to solve by available commercial software.

4.3.2 Max-min approach for the integrated monitoring strategy
The max-min approach for the integrated monitoring strategy is to maximize the monitoring efficiency for the fault which is the hardest to detect. The underlying rationale is simple that if the monitoring performance on the fault which has the longest time to signal can be improved, the monitoring performance on the whole fault set will be improved.

The most important reason why we use max-min approach is that we can obtain an equivalent linear formulation through transformations and linear integer programming tools can be used, unlike the nonlinear integer programming in section
4.2. Although this method does not solve the optimal integrated monitoring strategy directly, it can be used as a heuristic method to find good solutions to the problem. Moreover, readily available commercial software can be used to obtain the solutions as the problem is transformed to a linear integer programming. Due to the advantages of simplicity, reduced computational effort, and very good results, we believe that the max-min approach is helpful in solving the integrated monitoring strategy problem.

In solving a max-min problem, we may obtain more than one optimal solutions, which means we have numerous alternative optimal solutions. All these solutions have equivalent effect as they all have the same optimal value of the objective function; however, it is not the case if we consider the real situation. For example, if we have two optimal max-min solution for a process: strategy 1: the largest fault ATS=10, and the second largest fault ATS=8; strategy 2: the largest fault ATS=10, and the second largest fault ATS=6. Both of them have the same value of objective function-10, however, we always choose the second strategy due to the fact that it has the minimum maximum ATS except the largest ATS. Therefore, we prioritize the alternative optimal solutions based on the ATS of other faults. This approach leads us to compare all the other ATSs of solutions in order. Such approach is ended when only one optimal solution is available.

We first formulate the monitoring strategy into a general max-min form in problem (P1):

Problem (P1):

\[
\min_{\text{monitoring decision}} \left( \max_{S_j} ATS_i \right)
\]

ATS(i) denotes the time to detect the ith fault in the whole monitoring strategy. \( S_k = (S_{k,1}, S_{k,2}, \ldots, S_{k,k}) \) is the decision vector at kth stage. \( S_{k,1} = 1 \) means that residual chart based on stage 1 will be used. \( S_{k,k} \) indicates output monitoring chart.

Subject to:

\[
\sum_{j=1}^{k} S_{k,j} \leq 1, \quad \text{for } 1 \leq k \leq N
\]

\[
\sum_{k=1}^{N} \sum_{j=1}^{k} S_{k,j} \leq m
\]
\[
\sum_{i=1}^{l} S_{j,i} - S_{k,j} \geq 0
\]

\[S_{k,j} \leq 1\]

\[S_{k,j} \in \mathbb{Z}^+\]

Faults set= \{\text{F}_1, \text{F}_2, ..., \text{F}_n\}

The third constraint is added to ensure that the residual base has a control chart. \(ATS\) means the time interval from the fault appears to when the fault is detected by any downstream charts in the process. For each downstream chart at stage \(k\), there is a out-of-control \(ATS\) for this particular fault:

\[ATS_{k,i,j} = t^* \ ARL_{k,i,j} + delay = \frac{t}{1 - \beta_{k,i,j}} + delay\].

Due to the existence of delay, we can create an artificial \(\overline{ARL}_{k,i,j}\) and \(\overline{\beta}_{k,i,j}\) which are defined as:

\[ATS_{k,i,j} = t \times \overline{ARL}_{k,i,j} = \frac{t}{1 - \overline{\beta}_{k,i,j}}\]

Therefore, the artificial ARL of the chart at stage \(k\) is:

\[\overline{ARL}_{k,i} = \frac{1}{1 - \prod_{j=1}^{k} \overline{\beta}_{k,i,j} S_{k,j}}\]

And the overall artificial ARL for the \(i\)th fault is:

\[\frac{1}{1 - \prod_{k=1}^{N} \prod_{j=1}^{k} \overline{\beta}_{k,i,j} S_{k,j}}\]

The overall ATS for the fault is:

\[ATS_{i} = \frac{t}{1 - \prod_{k=1}^{N} \prod_{j=1}^{k} \overline{\beta}_{k,i,j} S_{k,j}}\]

With the above equation, the Problem (P1) can be further transformed to Problem (P2).

Problem (P2):

\[
\min \limits_{\text{monitoring decision}} \left( \max_i \left( \frac{t}{1 - \prod_{k=1}^{N} \prod_{j=1}^{k} \overline{\beta}_{k,i,j} S_{k,j}} \right) \right)
\]

Subject to:
Now, the integrated monitoring strategy has been formulated to a standard max-min problem. Nevertheless, as we can see in P3, this is a nonlinear integer programming problem. This makes the problem hard to solve. In the following paragraphs, we illustrate how to transform P3 to an equivalent linear integer programming problem.

Problem (P3):

\[
\min_{\text{monitoring decision}} z
\]

Subject to:

\[
\sum_{j=1}^{k} S_{k,j} \leq 1, \text{ for } 1 \leq k \leq N
\]

\[
\sum_{k=1}^{N} \sum_{j=1}^{k} S_{k,j} \leq m
\]

\[
\sum_{j=1}^{j} S_{j,d} - S_{k,j} \geq 0
\]

\[
S_{k,j} \leq 1
\]

\[
S_{k,j} \in \mathbb{Z}^+
\]

Faults set = \{\text{ faults }\} \{F_1, F_2, ..., F_n\}

\[
\frac{t}{1 - \prod_{k=1}^{N} \prod_{j=1}^{k} \bar{P}_{k,j} S_{k,j}} \leq z
\]

\[
\frac{t}{1 - \prod_{k=1}^{N} \prod_{j=1}^{k} \bar{P}_{k,j} S_{k,j}} \leq z
\]

\[
\frac{t}{1 - \prod_{k=1}^{N} \prod_{j=1}^{k} \bar{P}_{k,j} S_{k,j}} \leq z
\]
\[ S_{k,j} \in Z^+ \]

Faults set = \{ \{ F_1, F_2, \ldots, F_n \} \}

\[
\frac{t}{1 - \prod_{k=1}^{N} \prod_{j=1}^{k} \beta_{k,j} S_{k,j}} \leq z
\]

\[
\frac{t}{1 - \prod_{k=1}^{N} \prod_{j=1}^{k-1} \beta_{k,j} S_{k,j}} \leq z
\]

\[
\ldots
\]

\[
\frac{t}{1 - \prod_{k=1}^{N} \prod_{j=1}^{k-n} \beta_{k,j} S_{k,j}} \leq z
\]

First, the “equivalent objective function” is defined as: the one that has the same optimal solution while the objective function is not necessary equal (Ramirez, et. al., 2004). Based on this definition, the transformation of the objective function in P2 is done in the following way:

\[
\min_{\text{monitoring decision}} \left( \max_{i} \left( \frac{t}{1 - \prod_{k=1}^{N} \prod_{j=1}^{k} \beta_{k,j} S_{k,j}} \right) \right)
\]

\[
\min_{\text{monitoring decision}} \left( \max_{i} \left( \frac{1}{1 - \prod_{k=1}^{N} \prod_{j=1}^{k} \beta_{k,j} S_{k,j}} \right) \right)
\]

\[
\max_{\text{monitoring decision}} \left( \min_{i} \left( \left( 1 - \prod_{k=1}^{N} \prod_{j=1}^{k} \beta_{k,j} S_{k,j} \right) \right) \right)
\]

\[
\max_{\text{monitoring decision}} \left( \min_{i} \left( \left( - \prod_{k=1}^{N} \prod_{j=1}^{k} \beta_{k,j} S_{k,j} \right) \right) \right)
\]

\[
\max_{\text{monitoring decision}} \left( \min_{i} \left( \sum_{k=1}^{N} \sum_{j=1}^{k} S_{k,j} \ln \beta_{k,j} \right) \right)
\]

\[
\max_{\text{monitoring decision}} \left( \min_{i} \left( \sum_{k=1}^{N} \sum_{j=1}^{k} S_{k,j} \gamma_{k,j} \right) \right), \quad \gamma_{k,j} = - \ln \beta_{k,j}
\]
In the above transformations, logarithm is used as it remains the maximum/minimum property of the solution. By introducing the linearized equivalent objective function, we obtain the linearized formulation of the problem as shown in Problem (P4):

**Problem (P4):**

$$\max_{\text{monitoring decision}} z$$

Subject to:

$$\sum_{j=1}^{k} S_{k,j} \leq 1, \quad \text{for } 1 \leq k \leq N$$

$$\sum_{k=1}^{N} \sum_{j=1}^{k} S_{k,j} \leq m$$

$$\sum_{j=1}^{k} S_{j,d} - S_{k,j} \geq 0$$

$$S_{k,j} \leq 1$$

$$S_{k,j} \in Z^+$$

Faults set $$\{F_1, F_2, ..., F_n\}$$

$$\sum_{k=1}^{N} \sum_{j=1}^{k} S_{k,j} y_{1,j} \geq z$$

$$\sum_{k=1}^{N} \sum_{j=1}^{k} S_{k,j} y_{2,j} \geq z$$

...$

$$\sum_{k=1}^{N} \sum_{j=1}^{k} S_{k,j} y_{r,j} \geq z$$

The max-min problem is transformed to a linear problem as P4, therefore standard linear integer programming methods, such as branch-and-bound, can be used to solve the problem.

As we have mentioned, there may be multiple alternative optimal solutions for the problem. We need prioritize the alternative solutions based on ATS of other faults in the fault set. Among all the optimal solutions, the solutions with lower maximum ATS for the rest of the faults are preferable. In order to select such solutions, we fixed the fault with minimum maximum ATS and compare the second largest ATS...
between the alternative solutions. The ones with smaller value have higher priority. The comparison continues until only one optimal solution is obtained.

Ties, which mean two or more faults with the same maximum ATS, may happen when we proceed the above approach. We select one fault arbitrarily as the fault with minimum maximum ATS. By doing so, it won’t impact the effectiveness of the optimal solution.

The whole procedure is summarized as following:
1. Formulate the problem as shown in P4
2. Solve P4 by using standard integer programming methods, such as branch-and-bound.
3. If one unique optimal solution is obtained, the integrated monitoring strategy is determined.
4. If multiple alternative optimal solutions are obtained, compute and compare the second largest fault ATS. Select the solutions with smaller ATS.
5. If still multiple solutions are selected, continue to compare their third largest fault ATS. The comparison goes on until only one solution is selected. Then the integrated monitoring strategy is obtained.

4.4 Numerical example

In this section, we apply the integrated monitoring strategy developed on a multistage hood fit example in Lawless et al. (1999) to illustrate how to apply the strategy procedure in real applications. Both the original nonlinear integer programming and modified max-min linear integer programming are calculated. Through comparison of these two problem formulations, the results indicate that the max-min formulation can obtain good solutions, if not the best, with less computation effort, compared with the original nonlinear formulation.

The model of the process can be rewritten as a state space model as in Xiang and Tsung (2006).

\[
y_i = \alpha_i x_i + \epsilon_i, \quad i = 1, \ldots, 4 \\
x_{i+1} = \beta_i x_i + \epsilon_i,
\]
where \((\beta_1, \beta_2, \beta_3) = (1.15, 0.98, 1.06)\), \((\hat{\sigma}_{e_1}, \hat{\sigma}_{e_2}, \hat{\sigma}_{e_3}) = (0.13, 0.11, 0.2)\), and \(\alpha_i = 1\) for all \(i\). Without loss of generality, we assume that \(\sigma_e = 0.1\) and the fault set is \(\{U_2 = \sigma_{e1} = 0.13\), and \(U_3 = \sigma_{e2} = 0.11\}\).

In order to formulate to the form of P4, we first calculate the parameters- \(\overline{\beta}_{k,i,j}\) in P4. The noncentrality parameters for different charting methods at different stages for each fault are calculated according to the results in Jin and Tsung (2008a):

Table 4.2. Noncentrality parameters for \(U_2\)

<table>
<thead>
<tr>
<th>Stage&amp; Monitoring method</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output monitoring</td>
<td>NA</td>
<td>0.65</td>
<td>0.564</td>
<td>0.435</td>
</tr>
<tr>
<td>Residual based on stage</td>
<td>NA</td>
<td>0.65</td>
<td>0.564</td>
<td>0.435</td>
</tr>
<tr>
<td>Residual based on stage</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Residual based on stage</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

Table 4.3. Noncentrality parameters for \(U_3\)

<table>
<thead>
<tr>
<th>Stage&amp; Monitoring method</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output monitoring</td>
<td>NA</td>
<td>NA</td>
<td>0.487</td>
<td>0.376</td>
</tr>
<tr>
<td>Residual based on stage</td>
<td>NA</td>
<td>NA</td>
<td>0.487</td>
<td>0.376</td>
</tr>
<tr>
<td>Residual based on stage</td>
<td>NA</td>
<td>NA</td>
<td>0.618</td>
<td>0.427</td>
</tr>
<tr>
<td>Residual based on stage</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

Based on the above tables, we can further obtain the ARL, ATS and corresponding \(\overline{\beta}_{k,i,j}\). According to the definition of \(\overline{\beta}_{k,i,j}\) in section 4, the \(\overline{\beta}_{k,i,j}\) table is as follows:
Table 4.4. $\bar{p}_{k,i,j}$ for $U_2$ and $U_3$

<table>
<thead>
<tr>
<th>$U_2$</th>
<th>$U_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{p}_{2,1,2} = 0.68$</td>
<td>$\bar{p}_{3,2,3} = 0.83$</td>
</tr>
<tr>
<td>$\bar{p}_{3,1,3} = 0.89$</td>
<td>$\bar{p}_{3,2,4} = 0.93$</td>
</tr>
<tr>
<td>$\bar{p}_{4,1,4} = 0.94$</td>
<td>$\bar{p}_{3,2,1} = 0.83$</td>
</tr>
<tr>
<td>$\bar{p}_{2,1,1} = 0.68$</td>
<td>$\bar{p}_{3,2,2} = 0.7$</td>
</tr>
<tr>
<td>$\bar{p}_{3,1,1} = 0.89$</td>
<td>$\bar{p}_{4,2,1} = 0.93$</td>
</tr>
<tr>
<td>$\bar{p}_{4,1,1} = 0.94$</td>
<td>$\bar{p}_{4,2,2} = 0.92$</td>
</tr>
</tbody>
</table>

For other $\bar{p}_{k,i,j}$ not shown in this table, we set them as 1 due to the fact that their corresponding charting methods are unable to detect the corresponding fault. The solutions are displayed in table 4.5. A commercial software-LINGO is used to solve the max-min formulation. As we can see from the table, the optimal solution of the max-min formulation is the same as the optimal solution of the nonlinear formulation. In this case, max-min formulation can achieve the same optimal solution as the original formulation by using standard integer programming algorithms. But it does not always achieve the same optimal solution as the original one, while max-min approach can lead us to a “good” solution with reduced computation effort. This advantage of max-min formulation would be more apparent when the number of stages within a MMP becomes large.

Table 4.5. Optimal solution of the integrated monitoring strategy

<table>
<thead>
<tr>
<th>Stage</th>
<th>Decision variables of Max-min formulation</th>
<th>Decision variables of Nonlinear formulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$S_{1,1} = 1$</td>
<td>$S_{1,1} = 1$</td>
</tr>
<tr>
<td>2</td>
<td>$S_{2,1} = 1$</td>
<td>$S_{2,1} = 1$</td>
</tr>
<tr>
<td>3</td>
<td>$S_{3,2} = 1$</td>
<td>$S_{3,2} = 1$</td>
</tr>
<tr>
<td>4</td>
<td>$S_{4,1} = 1$</td>
<td>$S_{4,1} = 1$</td>
</tr>
</tbody>
</table>
4.5 Summary

An integrated monitoring strategy, which considers both the chart allocation and charting types together, is discussed in this chapter. The problem is originally formed to a nonlinear integer programming problem which is a modification of chart allocation strategy. However, due to the complexity of nonlinear programming, we introduce the max-min approach to formulate the problem in a new way in which the bottleneck of the monitoring performance is considered. In the new max-min formulation, the problem can be transformed to an integer programming problem. It enables us to apply standard methods, such as branch and bound, to solve the problem. Commercial software is also available to solve the max-min formulation. By comparing the solutions of the max-min formulation and the original nonlinear formulation, we find that the max-min approach is an effective way to obtain good solutions, if not the best.

The two formulations can be applied to different levels of complexity of the optimization problem. For example, if the complexity exceeds certain level, the max-min approach is preferred. There are mainly two ways to measure the complexity or the algorithmic efficiency for these two formulations: 1. Time efficiency- how fast an algorithm runs; 2. Space efficiency which focuses on extra memory space an algorithm requires. For the first kind, there are several methods to measure, such as program running time, number of repetitions of the basic operation, and empirical (or experimental) analysis of an algorithm’s efficiency. However, the complexity analysis is beyond the scope of this research. Further investigation can be done in future research.
Chapter 5

Smith-EWMA Run-to-Run Control Schemes for a Process with Measurement Delay

In the previous chapters, we develop the monitoring strategy including both chart allocation and integrated monitoring methods. Besides the monitoring scheme, the corrective actions are also critical when a signal appears in the monitoring system. In this chapter, we study how to take sequential corrective actions for popular run-to-run processes with measurement delay.

As we mentioned in chapter one, one of the main characteristics of run-to-run processes is that there always exist metrology delay. There are numerous schemes are proposed to improve the controller’s performance based on EWMA control scheme. Nevertheless, the metrology delay will not only hurt the controller’s performance but also decrease the stability region of a controller, which makes the controller experiencing higher risk in stability issues.

In this chapter, both the performance and stability issues are considered to develop a new controller based on the EWMA control scheme. In control theory, numerous control schemes that deal with time delay systems have been proposed. There are many publications focus on time-delay systems, such as Zavarei and Jamshidi (1987), Gorecki et al (1989), and Zhong (2006). To handle the time-delay systems, the Smith predictor (Smith (1959), (1957)), modified Smith predictor (Watanabe and Ito (1981)), and finite spectrum assignment (Manitius and Olbrot (1979), Olbrot (2000), Watanabe et al. (1983), and Watanabe (1986)) have been studied in control theory. The later two are always applied to a system that has an unstable delay or a delay other than the input and output, while the Smith predictor is used in controlling constant input or output delay systems together with PID type controllers (Zhong (2006)). The PID controllers have been well studied in the literature (Ang, et al. (2005), Astrom and Hagylund (1995), (2001), and Astrom et al. (1994)). On the other hand, the popular EWMA run-to-run control schemes, which in fact can be written as a form of a PID type controller, i.e., the integral controller (Del Castillo
have not been considered to combine with the Smith predictor to improve their performances and stability properties under measurement delays in a run-to-run practice.

The chapter is organized as follows. First, we introduce the models and basic properties of EWMA controllers and the Smith predictor scheme. We then propose a Smith-EWMA run-to-run controller, to control measurement delay processes. After that, we investigate the stability properties of the Smith-EWMA run-to-run controller, and make a comparison with the conventional EWMA and RLS type controllers. Finally, a performance analysis is discussed under two popular disturbance models: the first-order integrated moving average IMA(1, 1) model and the first-order autoregressive AR(1) model. The analysis and simulation results show that the Smith-EWMA run-to-run controller has better stability properties and also a more satisfactory performance for a process with serious measurement delay and model uncertainty.

5.1 Process model and Smith-EWMA

5.1.1 The Smith predictor

Because of inherent characteristics and current measurement methods, measurement delay in the feedback loop always occurs in semiconductor manufacturing. A long delay destabilizes the control loop and good responses are difficult to obtain. Control theory has some effective methods to deal with such time delays. The most commonly used method is the Smith predictor. The idea of the Smith predictor (also called the dead-time compensator (DTC)) was proposed by Smith (1957). A control system using the Smith predictor is shown in Fig. 5.1 (Smith, 1957, Morari, 1989; Newell, 1989). In Fig. 5.1, the Smith predictor takes the form of two models of the process: one is M, which includes the dead time, and other one is $M^*$, which has no time-delay.
That is,

\[ M = M^* z^{-d} \]  \hspace{1cm} (5.1)

where \( d \) means the dead time or time delay. The block (M-M*) is called the predictor. Without the predictor, the feedback signal, \( e \), is equal to the output, \( y \)

Without the predictor:

\[ y = M x. \]  \hspace{1cm} (5.2)

With the predictor:

\[ e = M^* x. \]  \hspace{1cm} (5.3)

Thus, the block (M-M*) predicts the effect of the manipulated variable, \( x \), on the output, \( y \), and modifies the feedback signal, \( e \), accordingly (Morari, 1989; Smith, 1959, 1957; Bahill, 2002). An equivalent IMC structure of a Smith predictor controlled system is shown in Figure 5.2.
PI and PID controllers are the most popular controllers in applications. In time delay systems, we usually use them in the “controller” block shown in Fig. 5.1 together with the Smith predictor. Another extension of the PI-Smith controllers has been investigated by Hagglund (1992), (1996) and (1993) and called the PPI (Predictive PI controller). This extension is particularly for first-order processes, and the tuning parameters of PI controllers have a fixed relationship with the parameters of the process.

5.1.2. Smith-EWMA model

Since a time delay deteriorates both the stability properties and performance of a PI/PID controlled system, people always introduce the Smith predictor to overcome this drawback. The same problem also occurs in EWMA controllers. In the semiconductor industry, a measurement delay in the process is always inevitable. A natural response for us is to use the Smith predictor again with the EWMA controller. Based on this thinking, we propose the following method: treat the EWMA controller as an Internal Model Controller (Adivikolanu, and Zafiriou (2000)) and introduce the Smith predictor control scheme into the EWMA controlled system.

To describe the procedure of how to construct a new Smith-EWMA controller based on an EWMA controller, we first do some transformation of the IMC structure of the EWMA controller from Fig. 1.3 to Fig. 5.3 and treat the parts in the dotted box as an integrated section called C.
The structure in Fig. 5.3 is the same as in Fig. 5.4; therefore, we transform the EWMA controlled system to the simplest feedback control system with controller C, process P and feedback g as in Fig. 5. By introducing the Smith predictor, we obtain the structure of an EWMA controlled system with the Smith predictor, labeled the Smith-EWMA controller (Fig. 5.5), in which \( \overline{p}(z) \) is the process model without measurement delay-M*, and \( \overline{p}(z) g(z) \) can be treated as the process model with measurement-M compared with Fig. 5.1.

\[
C(z) = \frac{q}{1 - \overline{p} g} \quad (5.4)
\]
The closed-loop transfer function of the system is

\[
\frac{q_p}{1 - 2 \tilde{p} q g + p g q + \tilde{p} q}.
\]  

(5.5)

In order to further investigate the underlying meaning of the Smith-EWMA structure, we will transfer the EWMA controller and Smith-EWMA controller to another form to compare the input’s expressions in time domain.

Fig. 5.6-a is equivalent to EWMA controller in Fig. 1 (Del Castillo, 2002), where

\[
G = \beta
\]

(5.6)

\[
G^* = b
\]

(5.7)

\[
f = \frac{B}{1 - (1 - \lambda)B}
\]

(5.8)

\[
D = B^d
\]

(5.9)

Using the basic transformation rule of block diagrams in control theory (Leigh, 2004), Fig. 5.6-a can be converted to Fig. 5.6-b, which has the same structure as in Fig. 5.3.
Figure 5.6. Development of Smith-EWMA and equivalent transformations

In Fig. 5.6-b, we can see that in traditional EWMA, 
\[ a_t = m_t - n_t \]  \hspace{1cm} (5.10)

and
\[ m_t = b \lambda X_{t-d-1} + (1 - \lambda) m_{t-1} \]  \hspace{1cm} (5.11)
\[ n_t = \beta \lambda X_{t-d-1} + (1 - \lambda) n_{t-1} \]  \hspace{1cm} (5.12)
\[ X_t = \frac{-a_t}{b} = \frac{m_t - n_t}{b} \]  \hspace{1cm} (5.13)

According to how the Smith-EWMA controller is formed (Fig. 5.4 and 5.5), Fig. 5.6-c is created based on the structure in Fig. 6-b. Equivalent block diagram transformation is used in Fig. 6-c to form Fig 6-d.

In Fig. 6-d, the input of C \( (c_i) \) includes several signals:
\[ c_i = m_t + m_t - n_t - g_t \]  \hspace{1cm} (5.14)

and
\[ X_t = \frac{c_t}{b} = \frac{(m_t - n_t) + m_t - g_t}{b} = \frac{-a_t + m_t - g_t}{b} \]  \hspace{1cm} (5.15)

where
\[ g_t = b \lambda X_{t-1} + (1 - \lambda) g_{t-1} \]  \hspace{1cm} (5.16)
By comparing equation (5.13) and (5.15), we can see that there is one more term in the numerator of Smith-EWMA controller’s input- $m_t - g_t$. With this term, we can make use of the most recent d-1 runs’ input information to adjust the system, though we cannot obtain the system output in the recent d-1 runs because of measurement delay. As you can see in equations (5.14)-(5.16), in Smith-EWMA controller, we add in the information of the most recent input $X_{t-1}$ in $g_t$, and use $g_t$ to further adjust $X_t$. The main feature of the Smith controller is that it uses the output of the process model without delay. This feature is also included in the Smith-EWMA controller. The difference between Smith-EWMA and EWMA controllers is that a new information flow (equation (5.16)) without delay is added to the feedback because of the Smith predictor scheme. Therefore, d plays no role in equation (5.16).

Since the Smith predictor scheme has good stability properties and performance with measurement delay, it is reasonable to expect that Smith-EWMA controllers achieve a better stability region and performance with measurement delay compared with traditional EWMA controllers. The analysis of the stability properties of the Smith-EWMA controller will be discussed in detail below, and simulation results about performance of the Smith-EWMA controller will also be presented.

### 5.2 Stability properties

Whenever we design a control system, the stability properties are the most important and fundamental things. We need to determine under what conditions the system can remain stable and when the output will converge to the target value. Ingolfsson and Sachs (1993) established the stability conditions for a delay-free system with EWMA controllers:

$$0 < \frac{\lambda \beta}{b} < 2$$

(5.17)

Under this condition, the system will always be stable with the EWMA controller. Ingolfsson and Sachs (1993) also analyzed some other systems with drifting processes with noise and wondering processes with noise. Further extensions have been made by Adivikolanu and Zafiriou (2000). Stability conditions for Double-EWMA controllers were presented by Del Castillo and Rajagopal (2002).
Adivikolanu and Zafiriou (1997) present a EWMA controller in an IMC structure and use the Jury stability test to define the stability region of the whole system. Their result is the same as Ingolfsson and Sachs’ results. Moreover, they also work out the z transform function of the EWMA controller with an IMC structure:

\[ q(z) = \frac{\lambda z / b}{z - 1 + \lambda} \]

Fig.5.1 shows the equivalent IMC structure for an EWMA controller, where \( p(z) = \beta \), \( \hat{p}(z) = b \), and \( g(z) \) means the measurement delay. If there is no measurement delay, then \( g(z) = z^{-1} \) accounts for run-to-run control. If there is \( d \) unit measurement delay, \( g(z) = z^{-(d+1)} \).

The closed-loop transfer function of the system becomes

\[ \frac{q p}{1 + q g(p - \hat{p})} \quad (5.18) \]

Good and Qin (2004) and Adivikolanu and Zafiriou (1997), (2000) used the Jury stability test to determine the stability condition for a process with one unit of measurement delay:

\[ 0 < \frac{\beta}{b} < \frac{1 + \lambda}{\lambda} \quad (5.19) \]

However, when \( d \) becomes bigger, the analytic solution of the stability region becomes very difficult to obtain by using the Jury test. Adivikolanu and Zafiriou (2000) proposed a method in which they used the critical value of the characteristic function’s roots to obtain the boundary of the stability region.

We describe how to derive the stability region of a Smith-EWMA controller based on Adivikolanu and Zafiriou’s (2000) method. We show the stability region for the Smith-EWMA controller. Then, we compare the stability properties between the EWMA controllers and the Smith-EWMA controllers in the coming content.

**5.2.1 No measurement delay**

The closed loop transfer function of a Smith-EWMA controlled system is:

\[ G(z) = \frac{q p}{1 - 2\hat{p}g + pqg + \hat{p}q}. \]

90
According to the Jury test in control theory, the stability region for Smith-EWMA controllers under no measurement delay is:

\[ 0 < \frac{\beta}{b} < \frac{2(1 + \lambda)}{\lambda}. \]  (5.20)

Compared with the stability region for EWMA controllers, \( 0 < \frac{\beta}{b} < \frac{2}{\lambda} \), the stability region has been enlarged by 2, which means that the Smith-EWMA controller gives us a larger tolerance for mismatch between the real process’s parameters and the model’s parameters (details of the calculations are given in the Appendix).

### 5.2.2 One unit measurement delay

We obtain the stability region for the Smith-EWMA controller with one unit metrology delay:

\[ 0 < \frac{\beta}{b} < \frac{1 + 3\lambda}{\lambda} \]  (5.21)

Compared with the stability region for EWMA controllers (5.19), this result also indicated that the stability property has been improved by introducing the Smith predictor scheme into EWMA controllers (the Appendix presents the detailed calculations).

### 5.2.3 More than one unit measurement delay

Nevertheless, similarly to the problem we face when we analyze the stability region for EWMA controllers, if \( d \) is larger than 1, the analytic solution of the stability region will be difficult to determine. Here, we use a critical value method by setting the root equal to \( z^* = e^{i\theta} = \cos \theta + i \sin \theta \) (5.22). This means that the root is on the unit cycle. Equation (5.22) will be used to determine the boundary of the stability region (Adivikolanu and Zafiriou, 2000).

We introduce \( z^* = e^{i\theta} = \cos \theta + i \sin \theta \) into

\[
D(z) = z^{d+1}(1 + \lambda) + z^d(\lambda - 1) + \frac{\lambda \beta}{b} - 2\lambda = 0:
\]

\[
[\cos(d + 1)\theta + i \sin(d + 1)\theta](1 + \lambda) + (\cos d\theta + i \sin d\theta)(\lambda - 1) + \frac{\lambda \beta}{b} - 2\lambda = 0. \]  (5.23)

and separate it into the real part and imaginary parts:
Real part: $\cos(d + 1)\theta \times (1 + \lambda) + \cos d\theta \times (\lambda - 1) + \frac{\lambda \beta}{b} - 2\lambda = 0. \quad (5.24)$

Imaginary part: $i\sin(d + 1)\theta \times (1 + \lambda) + i\sin d\theta \times (\lambda - 1) = 0. \quad (5.25)$

We first solve for the $\theta$s, which can satisfy the imaginary part of the equation, and then introduce all the $\theta$s into the real part and calculate the values of $\frac{\beta}{b}$. With different $\theta$s, the values of $\frac{\beta}{b}$ may be different. We choose the least positive as the critical value of $\frac{\beta}{b}$, and this is the boundary of the stability region.

Fig. 5.7 shows the stability region comparison between the Smith-EWMA controller and EWMA controller under different measurement delays.

![Figure 5.7. Stability comparison between Smith-EWMA and EWMA controllers](image)

The y-axis is $\Delta = \frac{\beta}{b}$, which represents the mismatch between the real process parameters and the model parameter. The x-axis is $\lambda$, the discount factor of the Smith-EWMA controller.
The solid line indicates the stability region of the Smith-EWMA controllers, and the dotted line indicates the stability region of the EWMA controller. The area below the line is the stability region.

From the above results, it indicates that the stability region is enlarged by using the Smith-EWMA controller. Now we consider the improvement in the stability region for some certain $\lambda$ s. Fig. 5.8 shows the increase of $\frac{\beta}{b}$ along with the increase of delay under some certain $\lambda$ s.

In the following results, the x-axis denotes $d+1$, where $d$ is the measurement delay, and the y-axis is $\Delta = \frac{\beta}{b}$. The solid line represents the stability region for the Smith-EWMA controller, and the dotted line represents the stability region for the EWMA controller. The stability region, $\Delta$, is enlarged by nearly 2 regardless of the values of $\lambda$ and $d$.

Figure 5.8. The stability comparison between two controllers with a specific $\lambda = 0.2$, 0.3 and 0.4.
Fig. 5.9 shows the improvement of delta, i.e., the enlargement of the stability boundary. From this figure, we can see that when $\lambda=0.2$, the improvement of the stability region (delta) would fluctuate a lot along with the increase of delay. However, if $\lambda$ is bigger, say, 0.4, the improvement of delta is more consistent even if the delay increases.

### 5.3 Performance comparison

In this section, we conduct simulation to evaluate the performance of the Smith-EWMA controller by comparing it with the EWMA controller and the RLS-random walk drift (RLS-RWD) controller (Wang et al., 2005), which is a type of run-to-run controller specially dealing with metrology delay. The RLS-RWD controller uses the recursive least square algorithm (RLS) to estimate the process parameter, $a_t$. In order to make the simulation results more practical, two popular disturbance models in a run-to-run process are considered in the simulation: the IMA (1,1) model and the AR(1) model. Mean square error (MSE) is applied to evaluate the performance of simulations. The simulations were carried out in MATLAB. MSE estimates are obtained based on 2000 replications of simulations:
where \( MSE_i \) denotes the MSE in the i-th replication of simulations.

### 5.3.1 Under IMA (1,1) disturbance

An IMA(1, 1) model with certain parameters produces disturbances that can closely mimic the behavior of many industrial processes including the run-to-run process (Box et al. 1983, Box and Luceno, 1994). Thus, we start from an IMA(1,1) process disturbance, \( z_t \), in our control model:

\[
z_t = z_{t-1} + a_t - \theta a_{t-1}
\]

(5.26)

where

\[
a_t \sim N(0, \sigma^2)
\]

\[
\theta \in [0, 1]
\]

and series of \( a_t \) are iid.

In a practical situation, the model parameter, \( b \), always differs from the true value, \( \beta \). Therefore in the simulation, we assume that \( \beta = 2 \), and \( b=1.5 \) or 1, with a target value of 0 in the model. As a smaller value of \( \theta \) produce disturbances with a greater degree of instability, and industrial time series usually shows a degree of instability as \( \theta \) between 0.8 and 0.6 (Box and Luceno, 1994), we may assume \( \theta =0.7 \) in the simulation.

Fig. 5.10 and 5.11 show the performance comparison among these controllers under \( \sigma =0.2 \). From Fig. 5.10, we can see that the Smith-EWMA controller achieve a smaller MSE compared with the EWMA controller, and the improvement becomes greater when the delay is larger. Furthermore, when the process model mismatch is moderate (\( b=1.5 \)), the Smith-EWMA controller outperforms the others if a large lambda is chosen. However, the RLS-RWD controller has a better performance than the Smith-EWMA controller when people choose a small tuning parameter. Another observation from this figure is that the performance of the Smith-EWMA controller is more consistent under various tuning parameter values than the other controllers. This feature may ensure a robust performance against inappropriate parameter turning.
Figure 5.11 presents the performance comparison when the process model mismatch is relatively large ($b=1$). Under such situation, the Smith-EWMA controller achieves better performances than the other two controllers even when a small tuning parameter is used ($\lambda=0.2$), and the advantage is more apparent along with the increase of the delay. Moreover, the performance consistency over various tuning parameter values still holds. Another feature of the Smith-EWMA controller is that in most cases its performance does not deteriorate much when the delay becomes larger. We have also conducted a sensitive analysis for $\sigma$ and concluded that the Smith-EWMA controller is not sensitive to the noise level, as we observed very similar trends and patterns to that in Fig. 5.10 and 5.11 for various $\sigma$ values (complete simulation data are available from the authors upon request).

Figure 5.10. Performance comparison with metrology delay when $b=1.5$ (solid line for Smith-EWMA, dash-dotted line for EWMA, dotted line for RLS-RWD)
5.3.2 Under AR (1) disturbance

In some run-to-run processes, the disturbance may be better described by a popular AR(1) model (see Eq (5.27)) instead. In this section we present the simulation results of the newly developed controllers under AR (1) disturbances with different parameters.

\[ z_{t+1} = \phi z_t + a_t \]  

(5.27)

where \( a_t \sim N(0, \sigma^2) \), \( \sigma = 0.2 \)

\[ \phi \in [-1,1] \]

In our simulation, we investigate the AR (1) models with different parameters, \( \phi = 0.5, 0.7 \) and 0.9. Fig. 5.12 and 5.13 show the performance comparisons when \( \phi = 0.9 \). In Fig. 5.12 that presents the moderate process model mismatch (b=1.5), we can see that when \( \lambda = 0.2 \), the RLS-RWD controller has smaller MSE than the Smith EWMA when delay is not large. However, as the delay increases, the Smith-EWMA controller outperforms the others. The advantage of the Smith-EWMA is more obvious when people choose a bigger value of \( \lambda \).
The performance comparisons with large process model mismatch are shown in Fig. 5.13. When $b=1$, the Smith-EWMA controller has an overwhelming advantage for different metrology delays and tuning parameters. The performance deterioration of the Smith-EWMA controller is also less than that of the RLS-RWD and EWMA controllers in that the Smith-EWMA has the flattest slope. We have also obtained the similar trend from the simulation of the AR(1) processes with $\phi = 0.5$ and 0.7.

Figure 5.12. Performance comparison with metrology delay when $b=1.5$ (solid line for Smith-EWMA, dash-dotted line for EWMA, dotted line for RLS-RWD)
Based on the simulation data of the conditions shown in the paper (both IMA and AR noises with different parameters), we calculate standard errors of MSE estimates and find out that they are within 3% of MSE estimates. For example, in the IMA noise case, when delay=5, b=1 and lambda=0.2, MSE estimate=0.0877 and its standard error=0.0005. Therefore, we conclude that the simulation runs are adequate. Based on the simulation results under different IMA (1,1) and AR (1) disturbances, we conclude that the Smith-EWMA controller can achieve a better performance when the process model mismatch or the delay is large. For moderate mismatch and delay cases, choosing an RLS-RWD controller with a small tuning parameter may be a better decision. Another nice feature of the Smith-EWMA controller is that its performances are more consistent under various tuning parameter values. This feature may ensure a robust performance against inappropriate parameter turning, especially when the engineers have limited knowledge about the process and disturbance.
5.4 Summary

In this paper, we proposed a Smith-EWMA run-to-run controller based on the Smith predictor structure and the EWMA run-to-run controller to deal with the measurement delay in a process. The stability properties have been discussed in comparison with traditional EWMA controllers. The stability properties of EWMA controllers are worse because of the lag in the measurement part. The results show that the Smith-EWMA run-to-run controller expands the stability region whether there is a measurement delay or not. In the performance comparison, we compare the Smith-EWMA run-to-run controller with both the EWMA controller and the RLS-RWD controller. The results show that under IMA and AR (1) disturbances, the Smith-EWMA run-to-run controller yields a more satisfactory result when the measurement delay and model mismatch are serious. Moreover, the Smith-EWMA run-to-run controller’s performance is robust to the tuning parameter of the controller, which ensures a reasonable good performance under model uncertainty. All these nice features make the Smith-EWMA run-to-run controller suitable for run-to-run processes with measurement delay, especially in semiconductor manufacturing in which the measurement delay is often inevitable.
Chapter 6
Conclusions

6.1 Summary
The quality control for multistage manufacturing processes becomes increasingly important recently due to the fact that production processes become more complex and always require multi-operation to finish a product. In practice, conventional SPC techniques, such as X-bar and R control charts, are applied directly to monitor a multistage process. However, nearly all the conventional SPC methods are developed for single stage process. When they are used in multistage processes, the performance is not very satisfactory because the interrelationships among stages are not considered. Another obstacle of applying conventional SPC in multistage process monitoring is that it is not always possible to have enough resources to place a control chart on each stage, especially when the number of stage is large. Therefore, how to rationally allocate limited resources (control chart) becomes a challenging task. Furthermore, in order to have a complete quality control scheme, charting methods and corrective control actions are also discussed in this thesis.

This research first proposes a chart allocation strategy for serial multistage processes. The interrelationships among stages are modeled with the state space model. Such information is considered in developing the fault propagation patterns which is presented in the form of noncentrality parameters along the process. Based on the fault propagation patterns, the corresponding ARL and ATS on each stage are obtained. Decisions are made to achieve minimum ATS along the process. Through the numerical example, we find that it is possible to obtain better performance by allocating a chart on the downstream stage of the faulty stage. This fact is deviated from the common sense of allocating chart on the faulty stage. Moreover, the strategy is extended to more realistic cases in which multiple potential faults are considered. The impact of uncertainty in the state space model is also studied to test if the decision is reliable or not.

In reality, SP-MMP are very common and more complex than S-MMP. We further develop the chart allocation strategy for SP-MMP. However, due the specialty of
SP-MMP structure, modifications on the state space model are made. Three special scenarios are modeled and analyzed. Fault propagation patterns are obtained through the process models. The special fault propagation in SP-MMP is found: an upstream mean shift fault may cause the shifts in both mean and variance in the downstream stages, if there is convergence in the middle. The strategy is formulated into an optimization problem in which we aim to minimize the sum of expected ATS of all the charts. Standard optimization tools, such as dynamic programming and branch and bound are used to achieve the optimal solutions.

Besides the allocation of charts, the charting methods we adopt in the process monitoring is also a key factor of the performance. We divided the charting methods into two kinds: 1. Output monitoring; 2. Residual monitoring. An integrated monitoring strategy is developed to optimize the monitoring performance by considering both the chart allocation and charting methods. The strategy is originally formulated to a nonlinear integer programming problem which is difficult to solve when we have a large number of stages. Therefore, the max-min approach is applied in formulating the problem in a new way. The max-min formulation can be transformed to a linear integer programming problem in which standard integer programming tools can be applied.

In the development of the chart allocation and integrated monitoring strategies, ATS was used as the main criterion. However, the strategies are not restricted by this criterion. By modifying the objective functions of the strategies, they can also be easily extended to cases that the loss due to fault or the cost of quality is treated as the main concern when parameters of the cost model can be obtained.

Another problem associated with the multistage processes quality control is how to take sequential corrective actions if there is a signal in the control charts. We study the run-to-run processes which are common in semiconductor manufacturing, and find out the performance of its control scheme would be worse due to the measurement delay within the process. A new Smith-EWMA controller is developed by introducing Smith predictor into the well known EWMA controller. The stability properties of the Smith-EWMA controlled are investigated and compared with traditional EWMA controllers. Smith-EWMA shows a larger stability region under
different parameters of the controller. Through the simulation results, we find that when the measurement delays are large, the Smith-EWMA controller shows bigger advantages in terms of performance (MSE). Moreover, the newly proposed controller is more robust to the tuning parameters compared with other control schemes.

This thesis has proposed quality control for multistage manufacturing processes from the allocation of chart, charting methods, to the sequential corrective control schemes. The successful implementation of these methods has important practical implications.

6.2 Future research areas

Besides the SPC problems in multistage manufacturing processes, service quality improvement has become a necessity in many industries, and many researchers also have noticed this trend. A lot of research pointed out that statistical process control has great possibilities in service industries, such as health care and education, and was already proven to be useful in the health care industry. The adoption of SPC into service operations provides a huge opportunity for service quality improvement. Most applications of SPC charts into multistage service processes assume the service quality characteristics in every stage are independent. Even if some researchers (Sulek, et al. 2005) have considered the inherent correlation information by using techniques such as cause selecting chart (Zhang (1985, 1989), and Hawkins (1993)), they still have not conquered the two main problems mentioned here. Moreover, due to the nature of service quality characteristics, we may face additional technical challenges:

1. Define quality characteristics in each operation.
2. Many of the quality characteristics in service processes may be not easy to quantify.
3. It is not easy to figure out the inherent correlation among operations.

Many researchers have investigated the first two challenges. One of the most popular quality definitions was proposed by Parasuraman, Zeuthaml, and Berry (1985), who defined service quality as the extent to which service meets or exceeds customers’
expectations. Parasuraman, et al. (1988) identified five service dimensions from their survey across industries and developed the service quality measurement scale called SERVQUAL, which is the most widely used measurement scale of service quality. TOPSIS (Hwang and Lin, 1987) and Loss Function (Ross, 1988) are also alternatives for perceptual quality measurement.

The study of monitoring strategy for multistage service processes will start from modeling issue. The research on service quality definition and measurement can be used to identify what to measure (quality characteristics) and how to measure in a service process. Statistical tools, such as design of experiments and regression analysis, will be applied to study the inherent correlations among stages. Finally, I plan to incorporate the above information together with traditional SPC charts to optimize the monitoring performance under resource constraints.

Finally, the quality control for multistage processes is also open for further enrichment and this work is far from complete. We believe that more methods will bring improvement to the topic and benefit practitioners.
Appendix A. Noncentrality parameters of output monitoring for multivariate cases

Consider the state space model first:

\[ X(i) = A(i)X(i-1) + U(i) + V(i) \]

\[ Y(i) = C(i)X(i) + W(i) \]  \hspace{1cm} (A1)

Suppose we need to focus on the fault at stage \( i \), and it may not be optimal to monitoring this fault at stage \( i \). Based on the state space model, we can easily derive the following relationship

\[ Y(n) = \sum_{j=1}^{n-1} C(n) \prod_{k=n}^{j+1} A(k)V(j) + C(n)V(n) + W(n) + C(n) \prod_{k=n}^{i+1} A(k)U(i) \]  \hspace{1cm} (A2)

The covariance matrix of \( Y(n) \):

\[
\Sigma_{Y(n)} = \begin{pmatrix}
\Sigma_{v1} & 0 & 0 & 0 \\
0 & \ldots & 0 & 0 \\
0 & 0 & \Sigma_{wn} & 0 \\
0 & 0 & 0 & \Sigma_{wn}
\end{pmatrix}
\]

\[
\begin{pmatrix}
C(n)A(n)...A(2) & C(n)A(n)...A(3) & \ldots & C(n)A(n) & C(n)
\end{pmatrix}
\]

\[ \begin{pmatrix}
1
\end{pmatrix}^T
\]

(A3)

\[ \Sigma_{Y(n)} = \sum_{i=2}^{n} C(n)A(n)...A(i)\Sigma_{v(i-1)}A(i)^T...A(n)^T C(n)^T + C(n)\Sigma_{wn} C(n)^T + \Sigma_{wn} \]  \hspace{1cm} (A4)

where \( \Sigma_1 \) and \( \Sigma_2 \) are the covariance of \( V(j) \) and \( W(j) \), for all \( j \).

The mean of \( Y(n) \):

\[ \mu_{Y(n)} = C(n) \prod_{k=n}^{i+1} A(k)U(i) \]  \hspace{1cm} (A5)
Appendix B. Noncentrality parameters of residual monitoring for multivariate cases

We base on stage-j to obtain the residuals of stage-n, in order to monitor the fault in stage-i. Where j<i<=n.

Object: To find an appropriate stage, n, in which we can detect the fault as soon as possible.

\[ Y(n) = \sum_{k=j+1}^{n} C(n) \prod_{l=n}^{k} A(k)V(l) + C(n)V(n) + W(n) + C(n) \prod_{k=n}^{i} A(k)U(i) + C(n) \prod_{k=n}^{j} A(k)C^{-1}(j)[Y(j) - W(j)] \]

(A6)

and the predicted value of Y(n):

\[ \hat{Y}(n) = C(n) \prod_{k=n}^{i+1} A(k)C^{-1}(j)Y(j) \]

(A7)

Then, the residual of Y(n)

\[ Y(n)_{\text{res}} = Y(n) - \hat{Y}(n) \]

\[ = \sum_{k=j+1}^{n} C(n) \prod_{l=n}^{k} A(l)V(l) + C(n)V(n) + W(n) + C(n) \prod_{k=n}^{i} A(k)U(i) - C(n) \prod_{k=n}^{j} A(k)C^{-1}(j)W(j) \]

(A8)

\[ Y(i)_{\text{res}} \] follows a MN(\( C(n) \prod_{k=n}^{i+1} A(k)U(i) \), \( \Sigma \))

Where,

\[ \Sigma = \sum_{k=j+1}^{n} C(n)A(n)\ldots A(k)C_{s(k-1)}\ldots A(1)C(n)^T + C(n)C_{m}C(n)^T \]

\[ + \sum_{m=1}^{\infty} C(n)A(n)\ldots A(j+1)C_{s(j)}C_{s(j+1)}\ldots A(1)C(n)^T \]

(A9)

And also,

\[ \mu_{Y(n)_{\text{res}}} = C(n) \prod_{k=n}^{i+1} A(k)U(i) \]

(A10)
Appendix C. Parameters of univariate example for Chart allocation

Four operations are involved in this hood assemble example: 1. Install the hood; 2. Paint the hood; 3. Install other hardware; 4. Finesse the hood. The four operations are denoted as HANG, PAINT, HARDWARE, and FINESSE in the following fig:

According to the results in Lawless, MacKay and Robinson (1999), model parameters are estimated based on the original data and given by $\alpha = (0.429, 0.375, 1.25)$ and $\beta = (1.15, 0.98, 1.06)$. Using appropriate transformation, the state space model can be written as

$$y_i^* = \alpha_i x_i^* + \epsilon_i, \quad i = 1, \ldots, 4$$

$$x_{i+1}^* = \beta x_i^* + e_i,$$

where $(\beta_1, \beta_2, \beta_3) = (1.15, 0.98, 1.06)$, $(\hat{\sigma}_\epsilon, \hat{\sigma}_{\epsilon_2}, \hat{\sigma}_{\epsilon_3}) = (0.13, 0.11, 0.2)$, and $\alpha_i = 1$ for all $i$. Without loss of generality, we assume that $\sigma_\epsilon = 0.1$. 
Appendix D. Parameters of multivariate example for Chart allocation

\[ A_1 = 12 \times 12 \]  Identity matrix

\[
\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.0007 & 1 & 0 & -0.0007 & -0.3497 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -0.3497 & 0 & 0 & 0.3497 & -325.17 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.0007 & 0 & 0 & -0.0007 & 0.6503 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{bmatrix}
\]

\[ A_2 = 12 \times 12 \]  Identity matrix

\[
\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.0005 & 1 & 0 & -0.0005 & -0.2392 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -0.5550 & 0 & 0 & -0.4450 & -222.49 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.0005 & 0 & 0 & -0.0005 & -0.2392 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1 & 0 & 0 & 1 & -0.0005 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -0.2392 & 0 & 0 & 0.2392 & -380.38 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.0005 & 0 & 0 & -0.0005 & 0.7608 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{bmatrix}
\]

\[ A_3 = 12 \times 12 \]  Identity matrix

\[
\begin{bmatrix}
1 & 0 & -550 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & -100 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -750 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & -80 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & -200 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & -100 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{bmatrix}
\]

\[ A_4 = 12 \times 12 \]  Identity matrix

\[
\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{bmatrix}
\]
Appendix E. Impact of parameter uncertainty for univariate cases in chart allocation

E.1 Output monitoring

\[ EX(i) \in [\bar{A}_i^- EX(i-1) + U(i), \bar{A}_i^+ EX(i-1) + U(i)] \]  \hspace{2cm} (A11)

\[ \Rightarrow EY(i) \in [\bar{C}_i^- \bar{A}_i^- EX(i-1) + \bar{C}_i^- U(i), \bar{C}_i^- \bar{A}_i^+ EX(i-1) + \bar{C}_i^- U(i)] \]  \hspace{2cm} (A12)

where

\[ \bar{C}_i^- = \bar{C}_i - \Delta_2 \]
\[ \bar{A}_i^- = \bar{A}_i - \Delta_1 \]
\[ \bar{A}_i^+ = \bar{A}_i + \Delta_1 \]

The noncentrality parameter for \( Y(i) \) is

\[ \sqrt{\mu_{Y(i)}^* \Sigma_{Y(i)}^{-1} \mu_{Y(i)}} \].  \hspace{2cm} (A13)

Since all the parameters are positive constants,

\[ a_1 b_1 \leq \sqrt{\mu_{Y(i)}^* \Sigma_{Y(i)}^{-1} \mu_{Y(i)}} \leq a_2 b_2 , \hspace{2cm} (A14) \]

where

\[ a_1 = (\bar{C} - \Delta_2)(\bar{A} - \Delta_1) EX(i-1) + (\bar{C} - \Delta_2) U(i) \]  \hspace{2cm} (A15),
\[ a_2 = (\bar{C} - \Delta_2)(\bar{A} + \Delta_1) EX(i-1) + (\bar{C} - \Delta_2) U(i) \]  \hspace{2cm} (A16)

\[ b_1 = \frac{1}{(\bar{C} + \Delta_2)(\bar{A} + \Delta_1) \Sigma_{Y(i)}^{-1} (\bar{A} + \Delta_1)'(\bar{C} + \Delta_2) + (\bar{C} + \Delta_2) \Sigma_{Y(i)}(\bar{C} + \Delta_2)' + \Sigma_{Y(i)}} \]  \hspace{2cm} (A17)
\[ b_2 = \frac{1}{(\bar{C} - \Delta_2)(\bar{A} - \Delta_1) \Sigma_{Y(i)}^{-1} (\bar{A} - \Delta_1)'(\bar{C} - \Delta_2) + (\bar{C} - \Delta_2) \Sigma_{Y(i)}(\bar{C} - \Delta_2)' + \Sigma_{Y(i)}} \].  \hspace{2cm} (A18)

E.2 Residual monitoring

\[ \Rightarrow EY(\text{res}) \in [(\bar{C}_n - \Delta_2) \prod_{k=1}^{n} (\bar{A}_k - \Delta_k) U(i), (\bar{C}_n + \Delta_2) \prod_{k=1}^{n} (\bar{A}_k + \Delta_k) U(i)] \]  \hspace{2cm} (A19)

And,

\[ \text{Var}(Y(\text{res})) \in [\sum_{k=2}^{n} (C(n) - \Delta_2)^2 (A(n) - \Delta_1)^2 \cdots (A(k) - \Delta_1)^2 \Sigma_{k-1} + (C(n) - \Delta_2)^2 \Sigma_n + \Sigma_w \]
\[ + (C(n) - \Delta_2)^2 (A(n) - \Delta_1)^2 \cdots (A(j+1) - \Delta_1)^2 (C(j) + \Delta_2)^2 \Sigma_n, \]
\[ \sum_{k=2}^{n} (C(n) + \Delta_2)^2 (A(n) + \Delta_1)^2 \cdots (A(k) + \Delta_1)^2 \Sigma_{k-1} + (C(n) + \Delta_2)^2 \Sigma_n + \Sigma_w \]
\[ + (C(n) + \Delta_2)^2 (A(n) + \Delta_1)^2 \cdots (A(j+1) + \Delta_1)^2 (C(j) - \Delta_2)^2 \Sigma_n] \]  \hspace{2cm} (A20)
We can obtain the noncentrality parameter bounds from

\[ a_3 b_3 \leq \sqrt{\mu' \sum_{Y(i)}^{-1} \mu_Y(i)} \leq a_4 b_4, \quad (A21) \]

where

\[ a_3 = (\bar{C}_n - \Delta_2) \prod_{k=1}^{n} (\bar{A}_k - \Delta_i \mathcal{U}(i)) \quad (A22) \]

\[ a_4 = (\bar{C}_n + \Delta_2) \prod_{k=1}^{n} (\bar{A}_k + \Delta_i \mathcal{U}(i)) \quad (A23) \]

\[ b_3 = \frac{1}{\sum_{k=1}^{n} J_k \times \sum_{k-1} + (C(n) + \Delta_2)^2 \sum_{n} + \sum_{w} + K \times \sum_{w}} \]

\[ b_4 = \frac{1}{\sum_{k=1}^{n} O \times \sum_{k-1} + (C(n) - \Delta_2)^2 \sum_{n} + \sum_{w} + P \times \sum_{w}} \]

where

\[ J_k = (C(n) + \Delta_2)^2 (A(n) + \Delta_i)^2 \ldots (A(k) + \Delta_i)^2 \]

\[ K = (C(n) + \Delta_2)^2 (A(n) + \Delta_i)^2 \ldots (A(k+1) + \Delta_i)^2 (C(j) - \Delta_2)^2 \]

\[ O_k = (C(n) - \Delta_2)^2 (A(n) - \Delta_i)^2 \ldots (A(k) - \Delta_i)^2 \]

\[ P = (C(n) - \Delta_2)^2 (A(n) - \Delta_i)^2 \ldots (A(j+1) - \Delta_i)^2 (C(j) + \Delta_2)^2. \]

Appendix F. Stability analysis for the Smith-EWMA controller with no measurement delay

With the transfer function, we can obtain the characteristic function of the system.

\[ D(z) = z^{d+\lambda}(1 + \lambda) + z^{d}(\lambda - 1) + \frac{\lambda \beta}{b} - 2\lambda \quad (A24) \]

If all the roots of \( D(z) = 0 \) fall in the unit cycle in the \( z \) plane, we claim that the system is stable. Instead of figuring out all the roots, we can also use the Jury test to see if the system is stable with certain parameter values.

When there is no measurement delay in the system (\( d=0 \)),

\[ D(z) = z(1 + \lambda) + \frac{\lambda \beta}{b} - \lambda - 1 \quad (A25) \]
The single root is \( z^* = 1 - \frac{b \lambda \beta}{1 + \lambda} \). Based on the stable test rule we have discussed above, if the system is stable, then

\[
|z^*| < 1 \quad \Rightarrow \quad \left| 1 - \frac{b \lambda \beta}{1 + \lambda} \right| < 1
\]

And

\[
0 < \frac{\beta}{b} < \frac{2(1 + \lambda)}{\lambda}
\]

**Appendix G. Stability analysis for the Smith-EWMA controller with one unit of measurement delay**

The characteristic function becomes:

\[
D(z) = z^2 (1 + \lambda) + z(\lambda - 1) + \frac{\lambda \beta}{b} - 2\lambda
\]  \hspace{1cm} (A26)

We use the Jury test to determine the stability region.

First, we do some modification for \( D(z) \).

\[
D(z) = z^2 + \frac{\lambda - 1}{\lambda + 1} z + \frac{\lambda \beta}{b} - 2\lambda
\]

This modification will not affect the roots of \( D(z) = 0 \).

The roots of \( D(z) = 0 \) are:

\[
z_{1,2}^* = \frac{-(\lambda - 1)(\lambda + 1) \pm \sqrt{(\lambda - 1)^2 / (\lambda + 1)^2 - 4(\lambda \beta / b - 2\lambda)}}{2}
\]

Based on the Jury test rule, if the system is stable, then

\[
|z_{1,2}^*| < 1
\]

\[
\Rightarrow \left| \frac{-(\lambda - 1)(\lambda + 1) \pm \sqrt{(\lambda - 1)^2 / (\lambda + 1)^2 - 4(\lambda \beta / b - 2\lambda)}}{2} \right| < 1
\]

\[
\Rightarrow 0 < \frac{\beta}{b} < \frac{1 + 3\lambda}{\lambda}
\]
Bibliography


Vita

Ming Jin was born in Wuhan, Hubei province, P. R. China. He started his study in the Department of Control Science and Engineering at Huazhong University of Science and Technology from 1999. He ranked top 5% in his department and received the Outstanding Student Award each year. In 2003, he was awarded the degree of Bachelor in Automation Control.

After that, he continued his postgraduate study at the Hong Kong University of Science and Technology (HKUST). Ming Jin enrolled in the Doctor of Philosophy program in the Department of Industrial Engineering and Engineering Management in the School of Engineering at HKUST in 2004.

While doing research in HKUST, Ming Jin was involved in several six sigma and supply chain projects related to manufacturing and services with several Fortune 500 companies. He is an American Society for Quality (ASQ) Certified Six Sigma Black Belt.