Abstract

Many real-world systems experience deterioration with usage and age, which often leads to low product quality, high production cost, and low system availability. Most previous maintenance and reliability models in the literature do not incorporate condition monitoring information for decision making, which often results in poor failure prediction for partially observable deteriorating systems. For that reason, the development of fault prediction and control schemes using condition-based maintenance techniques has received considerable attention in recent years.

This research presents a new framework for predicting failures of a partially observable deteriorating system using Bayesian control techniques. A time series model is fitted to a vector observation process representing partial information about the system state. Residuals are then calculated using the fitted model, which are indicative of system deterioration. The deterioration process is modeled as a 3-state continuous-time homogeneous Markov process. States 0 and 1 are not observable, representing healthy (good) and unhealthy (warning) system operational conditions, respectively. Only the failure state 2 is assumed to be observable. Preventive maintenance can be carried out at any sampling epoch, and corrective maintenance is carried out upon system failure. The form of the optimal control policy that maximizes the long-run
expected average availability per unit time has been investigated. It has been proved that a control limit policy is optimal for decision making. The model parameters have been estimated using the Expectation Maximization (EM) algorithm. The optimal Bayesian fault prediction and control scheme, considering long-run average availability maximization along with a practical statistical constraint, has been proposed and compared with the age-based replacement policy. The optimal control limit and sampling interval are calculated in the semi-Markov decision process (SMDP) framework. Another Bayesian fault prediction and control scheme has been developed based on the average run length (ARL) criterion. Comparisons with traditional control charts are provided. Formulae for the mean residual life and the distribution function of system residual life have been derived in explicit forms as functions of a posterior probability statistic. The advantage of the Bayesian model over the well-known 2-parameter Weibull model in system residual life prediction is shown. The methodologies are illustrated using simulated data, real data obtained from the spectrometric analysis of oil samples collected from transmission units of heavy hauler trucks in the mining industry, and vibration data from a planetary gearbox machinery application.
Acknowledgments

I would like to take this opportunity to thank all the people who have helped, encouraged and guided me along the way of completing my Ph. D. study at the University of Toronto.

First of all, I feel extremely grateful to my respectful supervisor, Professor Viliam Makis. Under his guidance in research and generous financial support, I have had a wonderful life in Toronto and successfully finished my academic study. I deeply appreciate his strict training in all aspects of research and teaching activities, which are definitely life-long treasures for me. His diligence, preciseness and dedication have always impressed me, and will continue to inspire me throughout my future career. I also want to thank my Ph. D. committee members, Professor Roy H. Kwon and Professor Chi-Guhn Lee, for their valuable suggestions and comments in my research progress.

Secondly, I would like to thank all the graduate students and research staff in my laboratory, Michael J. Kim, Catherine Hancharek, Dr. Ming Yang, Jing Yu, Zillur Rahman, Dr. Zhijian Yin, Lawrence Yip, Jue Wang, Chen Lin and Dr. Jian Liu, for their help in research and company in life. I could have never achieved this success without them.

Thirdly, I would like to thank the administrative staff in our department, Professor Markus Bussmann, Professor Hani Naguib, Brenda Fung, Donna Liu and Nicole Treston. Their generous help in my course of study and teaching assistants’ allocation really benefited me a lot. Several other staff members have also helped me in many ways. Special thanks to Joe Baptista, Geoffrey Chow, Sheila Baker, Sandra Chotilal, Joanna D. Gouveia, Fahmeda Bacchus, Oscar D. Rio and Teresa L. Ho.

The financial support from the Natural Science and Engineering Research Council of Canada (NSERC), Cetaris, Syncrude Canada Ltd., the Mathematics of Information Technology and Complex Systems (MITACS) and the University of Toronto Fellowship are also very much appreciated.

My deepest thanks go to my parents for their unconditional supports in my pursuit of study, and last but not least, this thesis is dedicated to my beloved wife for her love, support, understanding and encouragement.
# Table of Contents

Acknowledgments ....................................................................................................................................... IV

Table of Contents .................................................................................................................................... V

List of Tables ........................................................................................................................................ VIII

List of Figures .......................................................................................................................................... IX

List of Appendices .................................................................................................................................. XI

Chapter 1 Introduction and Thesis Outline ..........................................................................................1

1.1 Introduction and background ...........................................................................................................1

1.2 Research objectives and contributions ............................................................................................6

1.3 Overview of the thesis ......................................................................................................................7

Chapter 2 System Modeling and Parameter Estimation .........................................................................9

2.1 Introduction .....................................................................................................................................9

2.2 Vector auto-regressive modeling and calculation of residuals ......................................................13

2.3 Hidden Markov modeling and parameter estimation ......................................................................17

2.3.1 Formula for the likelihood function ............................................................................................19

2.3.2 Formula for the pseudo log-likelihood function .......................................................................22

2.3.3 Maximization of the pseudo log-likelihood function ..............................................................25

2.3.4 EM algorithm estimates for spectrometric oil data .................................................................27

2.4 Hidden semi-Markov modeling and parameter estimation ............................................................28

2.4.1 Formula for the likelihood function ............................................................................................29

2.4.2 Formula for the pseudo log-likelihood function .......................................................................31

2.4.3 Maximization of the pseudo log-likelihood function ..............................................................34

2.4.4 Numerical example .....................................................................................................................36

Chapter 3 The Structure of the Optimal Bayesian Control Policy ......................................................39
3.1 Model formulation .............................................................................................................39
3.2 Derivation of the optimality equation .............................................................................41
3.3 Structure of the optimal control policy ..........................................................................48
3.4 Computation of the optimal control policy ....................................................................52

Chapter 4 Bayesian Fault Prediction and Control Scheme ..................................................54

4.1 Bayesian control chart design for availability maximization ........................................54
  4.1.1 Problem formulation ..............................................................................................54
  4.1.2 The Bayesian control chart approach for availability maximization .....................55
  4.1.3 Computational algorithm in the SMDP framework ...............................................57
  4.1.4 Numerical example ...............................................................................................64
    4.1.4.1 Simulated data ..............................................................................................64
    4.1.4.2 Spectrometric oil data ..............................................................................71
  4.2 Bayesian control chart design using ARL criterion ....................................................73
    4.2.1 Model formulation .............................................................................................73
    4.2.2 Numerical example ............................................................................................74
      4.2.2.1 Simulated data ..........................................................................................74
      4.2.2.2 Vibration data .........................................................................................78
  4.3 Residual life estimation ...............................................................................................84
    4.3.1 Problem description and derivation of estimation formulae ..................................84
    4.3.2 Numerical example ............................................................................................86

Chapter 5 Conclusions and Future Research .....................................................................92

5.1 Conclusions .................................................................................................................92
5.2 Future research ............................................................................................................94

Bibliography ......................................................................................................................96

Appendix 1 ..........................................................................................................................103
List of Tables

Table 2.1 List of All Twenty Metal Elements..........................................................11
Table 2.2 List of Ten Wear Metal Elements............................................................11
Table 2.3 \( p \)-values for the Residual Independence and Normality Tests.............16
Table 2.4 EM Algorithm Estimates for Spectrometric Oil Data. ..............................28
Table 2.5 Optimal Parameter Estimates using the EM Algorithm.............................37
Table 4.1 Results for the Optimal Bayesian Control Chart.......................................65
Table 4.2 Optimal Values of Average Availability and Control Limit with Varying \( T_i \).........66
Table 4.3 Results for the Optimal Bayesian Control Chart with Statistical Constraint.......69
Table 4.4 Results for the Bayesian Control Policy in Spectrometric Oil Application.........72
Table 4.5 Control Limits for Different Control Charts...............................................76
Table 4.6 Fault Prediction Capability Comparison....................................................77
Table 4.7 Teeth and Speed Information for Syncrude Gearbox...............................78
Table 4.8 Indicator Analysis after DWT Applied to TSA Signals...............................79
Table 4.9 \( p \)-values of the Independence and Normality Tests for Variance and RMS Signals...82
Table 4.10 Parameter Estimates using the EM Algorithm............................................87
List of Figures

Figure 2.1 Retention of Wear Debris in Lubricating Oil.................................................10
Figure 2.2 SOAP Measurements of Copper and Iron.......................................................12
Figure 2.3 Scatter Plot for Healthy and Unhealthy Residuals........................................16
Figure 3.1 Illustration of the Optimal Control Limit Policy.............................................52
Figure 4.1 The Bayesian Control Approach.................................................................56
Figure 4.2 Optimal Control Limit & Maximum Average Availability versus $T_0$ ..........66
Figure 4.3 Optimal Control Limit & Maximum Average Availability versus $T_1$ ............67
Figure 4.4 Optimal Control Limit & Maximum Average Availability versus $T_2$ ............67
Figure 4.5 Maximum Average Availability versus Statistical Constraint.......................69
Figure 4.6 Optimal Control Limit versus Statistical Constraint....................................70
Figure 4.7 Optimal Sampling Interval versus Statistical Constraint................................70
Figure 4.8 Process Map of the Bayesian Control Chart for Spectrometric Oil Application....71
Figure 4.9 The Bayesian Control Chart for Spectrometric Oil Failure History ..........72
Figure 4.10 The Bayesian Control Chart for Simulated Failure History using ARL Criterion.....75
Figure 4.11 The Bayesian Control Chart for Simulated Suspension History using ARL Criterion.................................................................76
Figure 4.12 Transmission Diagram of the Gearbox from Syncrude Ltd........................78
Figure 4.13 Variance Signal after DWT Applied to TSA Signals.................................80
Figure 4.14 Kurtosis Signal after DWT Applied to TSA Signals.................................80
Figure 4.15 RMS Signal after DWT Applied to TSA Signals.......................................81
Figure 4.16 Crest Factor Signal after DWT Applied to TSA Signals.............................81
Figure 4.17 The Bayesian Control Chart for Variance and RMS Signals......................83
Figure 4.18 Simulated Residual Observation Process for Failure History.......................88
Figure 4.19 State Process of Simulated Residual Observation History.........................89
Figure 4.20 Simulated Residual Observation Process for Suspension History..................90
Figure 4.21 The Probability Density Function of Residual Life for Suspension History......91
List of Appendices

Appendix 1 Proof of Lemma 2.3 ........................................................................................................103
Appendix 2 Proof of Theorem 2.3 ..................................................................................................105
Appendix 3 Proof of Theorem 2.4 ..................................................................................................108
Appendix 4 Derivation of the Probability Transition Matrix in (4.2) using Laplace Transform110
Appendix 5 Theorem 3.1 of Provost and Rudiuk (1996). ............................................................112
Chapter 1 Introduction and Thesis Outline

1.1 Introduction and background

Many real-world systems experience deterioration with usage and age, which often leads to low product quality, high production cost, and low system availability. To avoid costly system failure, preventive maintenance (PM) is often carried out while the system is still operational. Traditional preventive maintenance models include the age replacement model (Barlow and Hunter 1960, Yeung et al. 2008), the block replacement model (Barlow and Proschan 1975, Sheu and Griffith 2002), and the periodic maintenance model (Barlow et al. 1963, Ait Kadi et al. 2006). Although such models are simple and easy to implement in practice, they do not account for condition monitoring information, which often leads to poor failure prediction (Christer et al. 1997, Jardine et al. 2006).

A modern approach that accounts for condition monitoring information in decision making is known as condition-based maintenance (CBM), which has recently attracted a lot of attention in the literature (Wang 2000, Grall et al. 2002, Makis and Jiang 2003, Kim et al. 2010, Jiang et al. 2011). Under CBM policy, preventive maintenance is triggered after identifying a symptom of impending failure with the aid of condition monitoring techniques. CBM models have been successfully applied to a variety of real-world problems, including furnace erosion prediction using a state space model (Christer et al. 1997), optimal preventive replacement using a proportional hazards model (Makis and Jardine 1992), and gearbox state modeling using a hidden Markov model (HMM) (Bunks et al. 2000). For completely observable systems, where the system state can be observed or identified completely, CBM models have been developed under a variety of different mathematical frameworks. Wang (2000) developed a CBM model based on a random coefficient growth model where the coefficients of the model are assumed to follow given distributions. The author derived the optimal monitoring interval and critical level for observation readings. Hosseini, Kerr and Randall (2000) developed a CBM model using generalized stochastic Petri Nets and obtained an optimal inspection policy.

In many CBM applications, the true state of the system is usually unobservable and can only be inferred from an observation process which is stochastically related to the hidden state process.
(Gupta and Lawsirirat 2006, Makis and Jiang 2003). For example, in mining operations, the oil that lubricates transmission units in heavy hauler trucks is sampled at regular intervals. Spectrometric analysis is carried out which provides the levels in ppm of different metals that come from the wear of the transmission units. Although the true state of the transmission unit is generally unobservable, the metal levels give partial information about the state of the operational unit. In industries that use rotating machinery (e.g. planetary gearboxes), multidimensional vibration data obtained from sensors gives partial information about the state of the operational unit (see e.g. Dalpiaz 2000, Wang and McFadden 1996). This type of process can be modeled using the hidden Markov model, which has been first successfully used in the research area of speech recognition (Rabiner 1990), and later in the areas of system diagnostics and fault detection in CBM applications (Ge et al. 2004, Lin and Makis 2003, Makis and Jiang 2003, Wang 2006, Xu and Ge 2004).

HMM is an extension of the Markov model where the observations are probabilistic functions of the hidden states. The actual sequence of states is not observable and can only be inferred from the observation process. These states are linked by possible transitions, each with an associated probability. The state transition is only dependent on the current state and not on the past states. An overview of the hidden Markov model as well as its associated estimation procedures and inference techniques are given by Rabiner (1990). In this research, it is assumed that the deterioration of a partially observable system evolves according to a 3-state continuous-time homogeneous Markov process, including two unobservable operational states and an observable failure state. The system is subject to condition monitoring at equidistant, discrete sampling epochs. The vector observation process is stochastically related to the system state. Model parameters are estimated using the Expectation-Maximization (EM) algorithm. An excellent overview of the EM algorithm is given by McLachlan and Krishnan (2008).

Although some papers have considered estimation problems for partially observable systems in the hidden Markov model framework, few papers have considered the inclusion of failure information, which is present in almost every maintenance application. Lin and Makis (2003) considered an interesting maintenance model with finite-valued observations and failure information, similar to the model considered in this research. Their objective was to derive a general recursive filter, which is important mainly for on-line re-estimation. The authors were
able to express the parameter updates in each iteration of the EM algorithm in terms of the recursive filter. However, such an approach has been found to be quite computationally intensive and difficult to implement when working with real data sets. We will show that for our model in this research, the parameter updates in each iteration of the EM algorithm have explicit formulae. This implies that each iteration of the EM algorithm can be performed with a single computation, which leads to an extremely fast and simple estimation procedure. This computational advantage is particularly attractive for practical applications.

Once the parameters of the HMM are estimated, a multivariate Bayesian fault prediction and control scheme can be constructed (Makis 2008, Yin and Makis 2011). Generally, two types of objective functions are commonly used in maintenance optimization over an infinite horizon. The first is the minimization of the long-run expected average cost per unit time (Jiang et al. 2001, Makis and Jiang 2003), and the second is the maximization of the long-run expected average availability per unit time (Barlow and Hunter 1960, Biswas and Sarkar 2000, Sarkar and Chaudhuri 1999). Although cost models are far more common in the maintenance literature, in many real applications, availability models are preferable since cost parameters are more difficult to estimate than availability parameters (i.e. time durations of maintenance actions). In addition, since availability maximization aims to increase the proportion of time that a system remains operational, it is especially desirable when the consequences of system downtime are serious, e.g. in public transportation systems, tracking radar systems, nuclear power systems, etc.

The concept of availability was first introduced into the stochastic optimization literature by Barlow and Hunter (1960) who analyzed the traditional age-based replacement policy for availability maximization. Some research has been done focusing on the mathematical analysis of availability models recently. However, very few structural results are known about the form of the optimal control policy maximizing system availability. For example, Cui and Xie (2001) studied the average availability for periodically inspected systems using a random walk model. They considered unobservable failures and derived an expression for the average system availability. Similar results were obtained by Kiessler et al. (2002), who also investigated systems with unobservable failures and deterioration governed by a Markov chain. The limiting average availability was computed and expressed in a closed form. Wang and Pham (2006) investigated the optimal maintenance policy for series systems subject to imperfect repair. The
authors showed that the long-run average availability for the imperfect repair case is significantly smaller than for the perfect repair case. Several other quantities such as the mean time between system failures, mean time between system repairs and asymptotic fractional system downtime were also derived. Kharoufeh et al. (2006) analyzed a system with unobservable failures in which system wear was modeled by a continuous-time Markov chain with additional damage induced by a Poisson shock process. The authors explicitly derived the system’s limiting average availability and other quantities such as the system’s lifetime distribution and mean time to failure. In many practical applications, the state of an operational system is often unobservable and can only be partially observed using on-line condition monitoring techniques (see e.g. Jin et al. 2005, Maillart 2006, Makis 2009, Makis and Jiang 2003, Ohnishi et al. 1986, Wu and Makis 2008). Lin and Makis (2004a, 2004b) studied filtering and parameter estimation problems for a partially observable failing system using the condition monitoring information. Their model has some similarities with the model in this research, namely, the state process evolves as a continuous-time homogeneous Markov process with unobservable working states and observable failures. However, to our knowledge, no research has been done on availability maximization under partial observations utilizing condition monitoring information for optimal decision making. In this research, we considered the availability maximization problem under partial observations. The problem was formulated as an optimal stopping problem with partial information. Under standard assumptions, we proved that the optimal control policy which maximizes the long-run expected average availability per unit time has a simple control limit structure, which leads to computational feasibility and easy implementation in practice.

To compute the optimal control policy for maintenance problems, several general computational approaches have been developed in the literature, such as the value-iteration algorithm (Chen and Trivedi 2005), the policy-iteration algorithm (Kim and Makis 2009), goal programming (Bertolini and Bevilacqua 2006), and linear programming (Taylor 1996). Although such general computational methods can provide optimal solution for a large class of maintenance problems, they are typically computationally demanding and difficult to implement in practice. Alternative approaches have been proposed to alleviate such computational difficulties by considering a particular policy structure. The control limit policy has been found to be quite effective in maintenance applications (Tagaras 1988, Rahim and Banerjee 1993, Ben-Daya and Rahim 2000, Cassady et al. 2000, Wu and Makis 2008, Yeung et al. 2008, Jiang et al. 2011, Jiang et al. 2011).
Makis (2008, 2009) has considered a quality control cost model with two unobservable in-control and out-of-control states, similar to the model considered in this research. The author proved that the optimal control policy, which monitors the posterior probability that system is in the out-of-control state, is in fact a control limit policy. Our model differs from the model in Makis (2008, 2009) since we are considering availability maximization and we also incorporate failure information, a key property present in maintenance applications. We designed a Bayesian control chart, which monitors the posterior probability that system is in an unhealthy state, and developed an efficient computational algorithm that determines the optimal values of the control limit and the sampling interval maximizing the long-run expected average availability per unit time. We solved the problem in the semi-Markov decision process (SMDP) framework and showed that the problem is equivalent to solving a system of linear equations, parameterized by the control limit and sampling interval. We also considered a practical statistical constraint, which imposes a lower bound on the probability of true alarms.

Over the last several decades, reliability engineers and statisticians have shown increasing interests in system residual life prediction and its application to real-world problems. For example, Watson and Wells (1961) used the mean residual life (MRL) to study burn-in processes of industrial systems, Elandt and Johnson (1980) used the MRL to study lifetime distributions of biomedical processes, and Morrison (1978) studied the MRL in social science for the life lengths of wars and strikes. Bryson and Siddiqui (1969), Deshpande, Kochar and Singh (1986) and Leemis (1986), showed how condition monitoring data can be used to compute quantities such as the hazard rate function, the cumulative hazard function, the conditional distribution function and the MRL for deteriorating equipments. Recently, researchers have also shown interests in the mathematical properties of the MRL function. For example, Tang, Lu and Chew (1999) characterized the behavior of the MRL for both continuous and discrete lifetime distributions based on the hazard rates. Similar results can also be found in Ghai and Mi (1999). Bradley and Gupta (2003) studied the limiting behavior of the MRL and derived an asymptotic expansion which can be used to obtain a good approximation of the MRL. This asymptotic expansion is valid for a general class of hazard rate distributions. For some processes with finite mean, Cox (1962), Kotz and Shanbhag (1980), Hall and Wellner (1981) showed that the MRL function completely determines the remaining life distribution via an inversion formula. These authors recommended the use of the MRL function as a helpful tool in system model building. The novel
feature of this research is the investigation of system residual life prediction in the HMM framework. The advantage this approach has over other traditional methods is that estimates for various quantities such as the mean residual life and the distribution function of system residual life are dynamically updated as new condition monitoring information becomes available. We show that the update formulae have closed-form expressions. The advantage of this approach compared with the well known 2-parameter Weibull model, in which no condition monitoring data is taken into account in the residual life prediction, is also illustrated.

1.2 Research objectives and contributions

In this research, we have considered a partially observable system subject to deterioration and random failure. The deterioration of the system is assumed to evolve according to a continuous-time, homogeneous Markov process with two unobservable operational states (healthy and unhealthy) and one observable failure state. The system is monitored at equidistant, discrete sampling epochs and multivariate observation vectors are collected, which provide partial information about the hidden condition of the system. Preventive maintenance can be carried out at any sampling epoch, and corrective maintenance is carried out upon system failure. We assume that two types of data histories are available: histories that end with observable system failure and histories that end when the system has been suspended from operation but has not failed.

The following contributions have been made in this research:

- A vector autoregressive (VAR) model is fitted to the healthy portions of each data histories, and residuals of the fitted model, which are indicative of system deterioration, are chosen as the observation process in the HMM framework for Bayesian fault prediction and control scheme design.

- HMM parameters, including parameters of the deterioration process and residual observation process, are estimated using the EM algorithm. We show that both the pseudo log-likelihood function and the parameter updates in each iteration of the EM algorithm have explicit formulae. Extensions, i.e. when the deterioration process evolves according to a hidden semi-Markov process, have also been considered.
The form of the optimal control policy which maximizes the long-run expected average availability per unit time has been investigated. The problem is formulated as an optimal stopping problem with partial information. It is proved that a simple control limit policy is optimal for decision making, which leads to computational benefits and easy implementation in practice.

An optimal Bayesian fault prediction and control scheme considering long-run average availability maximization along with a practical statistical constraint is proposed. The optimal control limit and sampling interval are calculated in the SMDP framework. Another Bayesian fault prediction and control scheme based on the average run length (ARL) criterion is also developed. Formulae for the mean residual life and the distribution function of system residual life are derived in explicit forms as functions of a posterior probability statistic.

The methodologies are illustrated using simulated data, real data obtained from the spectrometric analysis of oil samples collected from transmission units of heavy hauler trucks in the mining industry, and vibration data from a planetary gearbox machinery application.

1.3 Overview of the thesis

The remainder of the thesis is organized as follows.

In Chapter 2, a VAR model is fitted to the observation process data to remove the autocorrelation among observations. Residuals are calculated and chosen as the new observation process in the HMM framework. Using the EM algorithm, parameters for both the state and residual observation processes are estimated. Explicit estimation formulae are provided. One inherent weakness of the HMM methodology is that the state durations have exponential distribution which is sometimes not realistic in applications. To resolve this issue, hidden semi-Markov model (HSMM) has been proposed in the maintenance literature (see e.g. Ferguson 1980, Dong and He 2007). Unlike the HMM, the HSMM does not assume that the sojourn time in each
state is exponentially distributed, and therefore has stronger modeling capabilities. Such extensions including parameter estimations are also shown in this chapter.

In Chapter 3, the model for long-run average availability maximization is developed and the control problem is formulated as an optimal stopping problem under partial information. The value function and dynamic optimality equation are established. Important properties of the value function are derived. We analyze the dynamic optimality equation to show that a simple control limit policy is optimal for the long-run average availability maximization objective. A computational algorithm is provided for finding the optimal control limit and the maximum average availability.

In Chapter 4, the multivariate Bayesian control chart is introduced and formulated for the long-run average availability maximization objective. The posterior probability process is described and its stochastic evolution is characterized. A computational algorithm is developed in the SMDP framework to compute the optimal control limit and sampling interval, which could achieve the maximum long-run average availability. The incorporation of a practical statistical constraint that bounds the probability of true alarms is also considered, and the effect of the statistical constraint on the availability objective function is discussed. We compare the Bayesian model with the well-known age-based replacement model, and show the advantage of our approach. Another Bayesian fault prediction and control scheme is also designed based on the ARL criterion. Comparisons with traditional control charts, such as $\chi^2$ chart, multivariate exponentially moving average (MEWMA) chart, and multivariate cumulative sum (MCUSUM) chart, are provided. Several important quantities, such as the distribution function, reliability function, probability density function and hazard rate function of the residual life, are derived. The system mean residual life is computed. We use these quantities to illustrate the advantage of the Bayesian model over the well-known 2-parameter Weibull model, in which no condition monitoring data is taken into account in system residual life prediction.

In Chapter 5, we summarize the conclusions of this research and discuss future research.
Chapter 2 System Modeling and Parameter Estimation

In this chapter, we fit a VAR model to the observation process data obtained from the spectrometric oil application and calculate the residuals, which are chosen as the observation process in the HMM framework. Using the EM algorithm, parameters for both the state and residual observation processes are estimated. Explicit estimation formulae are provided.

2.1 Introduction

Deterioration of many mechanical equipments results in generation of wear debris in lubricating oil which carries important information about the system unobservable condition. Since it is often very costly to repair or replace a piece of equipment once it has failed, an important topic in maintenance and reliability engineering is to determine how wear debris information should be properly integrated into decision making.

In oil analysis applications, the most widely used technique is the spectrometric oil analysis program (SOAP), which was first commercially employed by U.S. railroad companies in the mid-20th century for detecting diesel engine failures in locomotives (Stachowiak et al. 2004). The analysis is based on principles of atomic physics whereby chemical elements emit or absorb light of distinct wavelengths. Oil samples are heated and radiant energy is converted to a signal which identifies the concentration levels (in $ppm$) of elements present in the oil sample. As a machine degrades, wear debris accumulates in the lubricating oil and the concentrations of certain wear elements increase (see e.g. Stachowiak et al. 2004).
Although SOAP has been utilized in practice for many years, very little work has been done using methods from statistics and operations research to model and analyze oil data for the purpose of optimal decision making. We have found that most of current methods used in oil analysis are generally qualitative in nature and are based on “best practices” that are unique to each type of machinery. Such an approach is sometimes referred to as an Expert System (see e.g. Yan et al. 2005). Readers are referred to Totten (2006) for an extensive overview of current methods and techniques used in the oil analysis area.

In this research, we present a method for predicting failures of a partially observable deteriorating system. We study the real spectrometric diagnostic oil data coming from the samples of transmission units of 240-ton heavy hauler trucks used in the Athabasca Oil Sands of Alberta, Canada. During the operational life of each transmission unit, oil samples are collected every $\Delta = 600$ hours and spectrometric oil analysis is carried out which provides the concentrations in ppm of 20 elements. Two types of oil data histories are considered: histories that end with observable failures and histories that end when the system is suspended from operation but has not failed. The total number of histories recorded is 36, which consist of $N = 13$ failure histories and $M = 23$ suspension histories.

The 20 different metal elements are listed in Table 2.1.
Although 20 different elements are measured at each sampling epoch, only 10 elements come from direct wear of the transmission units. The other 10 elements are known as oil additives, antifreeze additives and contaminants, and they do not provide information about the state of the transmission units (see e.g. Totten 2006). The 10 wear elements are listed below along with their respective means and standard deviations in ppm. These are the grand means and standard deviations computed over the entire lifetimes of all the 36 data histories.

Table 2.2 suggests that copper and iron carry most of the statistical information since the other 8 wear elements have relatively small means and standard deviations. Thus, to avoid over-parameterization, we only use the bivariate data consisting of copper and iron for the remainder of our analysis. A typical oil data history is plotted below. This particular transmission unit failed after the 13th sampling epoch at 8123 operational hours.
Figure 2.2 SOAP Measurements of Copper and Iron

Figure 2.2 shows that from the 1st to 9th sampling epoch the measurements are stable, indicating that the transmission unit is operating under normal conditions. This is regarded as the healthy portion of the data history. Beyond the 9th sampling epoch, the metal levels dramatically increase, which indicates severe wear of the transmission unit. This is regarded as the unhealthy portion of the data history. For the mining application, the transmission oil was changed regularly after every other oil sample was collected. This explains the concentration decreases in between two consecutive sampling epochs in Figure 2.2. Obviously, in applications where oil changes occur frequently at non-regular times, such oil changes should be modeled explicitly since they would affect subsequent parameter estimation and control. However, for this application, the sampling interval Δ is 600 operational hours, which is long enough for wear debris to accumulate in between oil changes if the system experiences severe wear. Therefore, even with oil changes, significant increases in metal concentrations are still detectable.
2.2 Vector auto-regressive modeling and calculation of residuals

As stated before, we first need to fit a VAR model that accounts for the autocorrelation in the data histories, and choose as the observation process in the HMM framework, the residuals of the fitted model. The idea is that the residuals contain most of the significant information regarding the hidden state process. The approach of analyzing the residuals of the fitted model has a solid theoretical backing, and has been successfully applied in a variety of statistical, scientific and engineering applications (see e.g. Yang and Makis 2000, Schneider and Frank 1996, Schoenberg 2003, Wang and Wong 2002).

Since the switching time from healthy data to unhealthy data is unobservable, before fitting a model to the data histories, we have to approximate the healthy and unhealthy portions of the data histories. We now identify the healthy portions of the data histories. Partitioning a non-stationary time series into a finite number of stationary portions is known as time series segmentation. The purpose of segmentation in our application is to achieve stationarity in the healthy portions of the data histories so that the residuals of the fitted model can be calculated. Generally, there is no agreed upon criterion for selecting the optimal segmentation. Thus, a variety of segmentation methods exist in the literature, ranging from very sophisticated algorithms to simple heuristic graphical methods (see e.g. Keogh 1993, Fukuda et al. 2004). In this application, since we only need to segment the data histories into two portions (healthy and unhealthy), for simplicity, we have chosen to segment the data histories via graphical examination. For each of the $N + M = 36$ data histories, the healthy portions of the data histories are denoted as $\{z_1^{l}, \ldots, z_t^{l}\}$, where $l = 1, \ldots, N + M$. We next fit a VAR model using the healthy data histories and calculate the residuals in the healthy state. Using this fitted VAR model, the residuals are calculated also for the unhealthy portion of data histories.

The healthy data histories are assumed to follow a common stationary VAR process (see e.g. Reinsel 1997) given by

$$Z_n - \delta_0 = \sum_{r=1}^{p} \Phi_r (Z_{n-r} - \delta_0) + \epsilon_n, \quad n \in \mathbf{Z}$$

(2.1)
where \( \varepsilon_n \) are i.i.d. \( N_z(0, C) \), the model order \( p \in \mathbb{N} \), the autocorrelation matrices \( \Phi_r \in \mathbb{R}^{2 \times 2} \), and the mean and covariance model parameters \( \delta_0 \in \mathbb{R}^2 \) and \( C \in \mathbb{R}^{2 \times 2} \). All the model parameters are unknown and need to be estimated. We set \( \delta = \delta_0 - \sum_{r=1}^{p} \Phi_r \delta_0 \), and rewrite equation (2.1) in the standard form

\[
Z_n = \delta + \sum_{r=1}^{p} \Phi_r Z_{n-r} + \varepsilon_n, \quad n \in \mathbb{Z}
\]  

(2.2)

so that the observed healthy data histories \( \{z_1^l, \ldots, z_t^l\} \), \( l = 1, \ldots, N + M \), have the regression representation \( \mathbf{W} = \mathbf{V}A + \mathbf{E} \), where

\[
\mathbf{W}' = \begin{pmatrix} z_{1}^{N+M} & \ldots & z_{p+1}^{N+M} & \ldots & z_{1}^{1} & \ldots & z_{p}^{1} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ z_{1}^{N+M} & \ldots & z_{p}^{N+M} & \ldots & z_{1}^{1} & \ldots & z_{p}^{1} \end{pmatrix},
\]

\[
\mathbf{A}' = \left( \delta, \Phi_1, \ldots, \Phi_p \right),
\]

\[
\mathbf{E}' = \begin{pmatrix} \varepsilon_{1}^{N+M} & \ldots & \varepsilon_{p+1}^{N+M} & \ldots & \varepsilon_{1}^{1} & \ldots & \varepsilon_{p}^{1} \end{pmatrix},
\]

\[
\mathbf{V}' = \begin{pmatrix} 1 & 1 & 1 & 1 \\ z_{1}^{N+M} & \ldots & z_{p}^{N+M} & \ldots & z_{1}^{1} & \ldots & z_{p}^{1} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ z_{1}^{N+M} & \ldots & z_{1}^{N+M-p} & \ldots & z_{1}^{1} & \ldots & z_{1}^{1} \end{pmatrix}.
\]

Reinsel (1997) showed that the least squares estimates for \( \mathbf{A} \) and \( \mathbf{C} \) are given by

\[
\hat{\mathbf{A}} = (\mathbf{V}'\mathbf{V})^{-1}\mathbf{V}'\mathbf{W}
\]

\[
\hat{\mathbf{C}} = \frac{(\mathbf{W}-\mathbf{V}\hat{\mathbf{A}})(\mathbf{W}-\mathbf{V}\hat{\mathbf{A}})}{T-(2p+1)}
\]

(2.3)

where \( T = \sum_{l=1}^{N+M} (t_l - p) \) is the total number of available data points. The estimate for the model order \( p \in \mathbb{N} \) is obtained by testing \( H_0 : \Phi_p = \mathbf{0} \) against \( H_a : \Phi_p \neq \mathbf{0} \), using the likelihood ratio statistic given by
\[ M_p = -(T - 2p - 1 - 1/2) \ln \left( \frac{\det(S_p)}{\det(S_{p-1})} \right) \]  

(2.4)

where \( S_p = (W - V\hat{A})(W - V\hat{A})' \) is the residual sum of squares matrix obtained from equation (2.3) when fitting a VAR model of order \( p \in \mathbb{N} \). For large \( T = \sum_{j=1}^{N+M} (t_j - p) \), if \( \Phi_p = 0 \) is true, \( M_p \) converges in distribution to \( \chi^2_4 \). Thus, for the significance level \( \alpha \in (0,1) \), we reject \( H_0 : \Phi_p = 0 \) if \( M_p > \chi^2_{4, \alpha} \).

In this bivariate spectrometric oil analysis, we find that \( M_2 = 221.84 \) and \( M_3 = 10.28 \). From the Chi-square distribution with 4 degrees of freedom and \( \alpha = 0.01 \), we know that \( \chi^2_{4,0.01} = 13.28 \). Since \( M_2 > \chi^2_{4,0.01} \) and \( M_3 < \chi^2_{4,0.01} \), we reject \( H_0 : \Phi_2 = 0 \) and fail to reject \( H_0 : \Phi_3 = 0 \). Thus, we conclude that \( \hat{p} = 2 \) is an adequate model order. Using equation (2.3), the VAR model parameter estimates are given by

\[
\begin{bmatrix}
\hat{\Phi}_1 \\
\hat{\Phi}_2
\end{bmatrix} =
\begin{bmatrix}
0.3825 & -0.0758 \\
-0.0672 & 0.1775
\end{bmatrix},
\hat{\delta} =
\begin{bmatrix}
7.6819 \\
4.0570
\end{bmatrix},
\hat{C} =
\begin{bmatrix}
7.1789 & 2.0260 \\
2.0260 & 3.5725
\end{bmatrix}
\]  

(2.5)

From the parameter estimates given in (2.5), the eigenvalues of \( \hat{\Phi} = \begin{bmatrix} \hat{\Phi}_1 & \hat{\Phi}_2 \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \) are computed as 0.8200, 0.6729, -0.4156 and -0.5173, which are all smaller than one in the absolute value implying the fitted model is stationary.

Using estimates \( \hat{\gamma} = (\hat{\delta}, \hat{\rho}, \hat{\Phi}_1, \hat{\Phi}_2, \hat{C}) \), we define the residual process \((Y_n : n \in \mathbb{N})\) by

\[ Y_n := Z_n - E_{\hat{\gamma}}(Z_n | \hat{Z}_{n-1}) \]  

(2.6)

where \( \hat{Z}_{n-1} = (Z_1, \ldots, Z_{n-1}) \). The residuals are then computed for both the healthy and unhealthy portions of each data history. All the residuals are provided graphically in a 2-dimensional
scatter plot below. The crosses are residuals computed from the healthy data set and the circles are residuals computed from the unhealthy data set.

![Figure 2.3 Scatter Plot for Healthy and Unhealthy Residuals](image)

We statistically test the independence and normality assumptions using the Portmanteau Independence Test (Cromwell et al. 1994) and the Henze-Zirkler Multivariate Normality Test (Henze and Zirkler 1990), respectively, and obtain the following results.

<table>
<thead>
<tr>
<th></th>
<th>Healthy data set</th>
<th>Unhealthy data set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independence (Portmanteau)</td>
<td>0.0675</td>
<td>0.4284</td>
</tr>
<tr>
<td>Normality (Henze-Zirkler)</td>
<td>0.6911</td>
<td>0.5270</td>
</tr>
</tbody>
</table>

Table 2.3 shows that there is no statistical evidence to reject the null hypotheses that the residuals of the fitted model are independent and have multivariate normal distribution (see also the theoretical results provided by Yang and Makis 2000).

In the next section, the residuals of the fitted model defined in equation (2.6) are chosen as the observation process in the HMM framework, and the EM algorithm is used to estimate the parameters for both the state and observation processes of the HMM.
2.3 Hidden Markov modeling and parameter estimation

Now suppose that the state (deterioration) process of the deteriorating system evolves according to a continuous-time homogeneous Markov chain \((X_t: t \geq 0)\) with state space \(S = \{0, 1, 2\}\), where states 0 and 1 are unobservable, representing the healthy and unhealthy operational states respectively, and state 2 represents the observable failure state. The system is always assumed to start in the healthy state, i.e. \(X_0 = 0\), and the instantaneous transition matrix for the state process is given by

\[
\Lambda = (\lambda_{ij})_{i,j \in S} = \begin{pmatrix}
-&(\lambda_{01} + \lambda_{02})&\lambda_{01}&\lambda_{02} \\
0&-\lambda_{12}&\lambda_{12} \\
0&0&0
\end{pmatrix}
\] (2.7)

where \(\lambda_{01}, \lambda_{02}, \lambda_{12} \in (0, +\infty)\) are unknown state parameters.

Let \(\xi = \inf\{t \in \mathbb{R}: X_t = 2\}\) be the observable system failure time. Suppose at equidistant sampling times \(\Delta, 2\Delta, \ldots, \Delta > 0\), residual observations \(Y_1, Y_2, \ldots \in \mathbb{R}^d\) are obtained using equation (2.6), which give partial information about the system hidden state. The \(Y_i\)'s are assumed to be conditionally independent given the state of the system, and for each \(n \in \mathbb{N}\), it is assumed that \(Y_n\) conditional on \(X_{n\Delta} = x, x = 0, 1\), has a \(d\)-variate normal distribution \(N_d(\mu_x, \Sigma_x)\) with density

\[
f_{Y_n | X_{n\Delta}}(y | x) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma_x)}} \exp\left(-\frac{1}{2} (y - \mu_x)\Sigma_x^{-1}(y - \mu_x)\right)
\] (2.8)

where \(\mu_0, \mu_1 \in \mathbb{R}^d\) and \(\Sigma_0, \Sigma_1 \in \mathbb{R}^{d \times d}\) are the unknown residual observation process parameters. Once the system fails, \(P(Y_n = \eta | X_{n\Delta} = 2) = 1\), where \(\eta \in \mathbb{R}^d\) represents a failure signal.

Suppose we have collected \(N\) failure histories, which we denote as \(F_1, \ldots, F_N\). Failure history \(F_i\) is assumed to be of the form \(\tilde{Y}_i = (y_{i1}, \ldots, y_{iT_i})\) and \(\xi_i = t_i\), where \(T_i \Delta < t_i \leq (T_i + 1)\Delta\). The history \(\tilde{Y}_i\) represents the collection of all vector data \(y_{ij} \in \mathbb{R}^d, j \leq T_i\), which was obtained through
condition monitoring until system failure at time $t_j$. Suppose further that we have collected $M$ suspension histories, which we denote as $S_1,\ldots,S_M$. Suspension history $S_j$ is assumed to be of the form $\tilde{Y}_j = (y^j_1,\ldots,y^j_t)$ and $\tilde{y}_j > T_j \Delta$. Let $O = \{F_1,\ldots,F_N,S_1,\ldots,S_M\}$ represent all the observable data and $L(\lambda, \theta | O)$ be the associated likelihood function, where $\lambda = (\lambda_0, \lambda_1, \lambda_2)$ and $\theta = (\mu_0, \mu_1, \Sigma_0, \Sigma_1)$ are the sets of unknown state and residual observation parameters. Next, we use the EM algorithm to estimate $\lambda$ and $\theta$. Details can be found in Kim, Jiang, Makis and Lee (2011).

**E-STEP.** For $n \geq 0$, compute the pseudo log-likelihood function defined by

$$Q(\lambda, \theta | \lambda_n, \theta_n) := E_{\lambda_n, \theta_n} \left( \ln L(\lambda, \theta | C) | O \right)$$

(2.9)

where $\lambda_0$ and $\theta_0$ are some initial values of the unknown parameters, and $C = \{F_1,\ldots,F_N,\overline{S}_1,\ldots,\overline{S}_M\}$ represents the complete data set, in which each failure history $F_i$ and suspension history $S_j$ of the observable data set $O$ have been augmented with the unobservable sample path information of the state process.

**M-STEP.** Choose $\lambda_{n+1}$ and $\theta_{n+1}$ such that

$$(\lambda_{n+1}, \theta_{n+1}) \in \arg \max_{\lambda, \theta} Q(\lambda, \theta | \lambda_n, \theta_n)$$

(2.10)

The E and M steps are repeated until the Euclidean norm $\| (\lambda_{n+1}, \theta_{n+1}) - (\lambda_n, \theta_n) \| < \varepsilon$, for small $\varepsilon > 0$. The estimates $\lambda_{n+1}$ and $\theta_{n+1}$ then approximate the maximizers of $L(\lambda, \theta | O)$. We will show in Theorems 2.1 and 2.2, that $Q(\lambda, \theta | \lambda_n, \theta_n)$ admits the following decomposition

$$Q(\lambda, \theta | \lambda_n, \theta_n) = Q^{\text{state}} (\lambda | \lambda_n, \theta_n) + Q^{\text{obs}} (\theta | \lambda_n, \theta_n)$$

(2.11)
where $Q^{\text{state}}(\lambda | \lambda_n, \theta_n)$ depends only on the state parameter $\lambda$, and $Q^{\text{obs}}(\theta | \lambda_n, \theta_n)$ depends only on the residual observation parameter $\theta$. This implies in particular that the M-step can be carried out separately for the state and residual observation parameters, which considerably simplifies the algorithm and increases the speed of computation.

### 2.3.1 Formula for the likelihood function

Let $\tau_0 = \inf\{t \in \mathbb{R}_+ : X_t > 0\}$ be the unobservable sojourn time of the state process in the healthy state 0. It is clear that there is a one-to-one correspondence between the entire sample path of the system state process and the two random variables $\tau_0$ and $\xi$. The distributional properties of the sojourn time $\tau_0$ and failure time $\xi$ are given by the following lemma.

**Lemma 2.1** For each $t \in \mathbb{R}_+$, the density function of $\xi$ is given by

$$f_\xi(t) = p_{01} \frac{v_0 v_1}{v_0 - v_1} (e^{-\tau_0 t} - e^{-\xi t}) + p_{02} v_0 e^{-\xi t}$$

(2.12)

For all non-negative $s < t$, the conditional density function of $\tau_0$ given $\xi$ is

$$f_{\tau_0|\xi}(s \mid t) = \frac{p_{01} v_1 e^{-\tau_0 t} e^{-(v_0 - v_1)s}}{p_{01} \frac{v_1}{v_0 - v_1} (e^{-\tau_0 t} - e^{-\xi t}) + p_{02} e^{-\xi t}}$$

(2.13)

and for each $t \in \mathbb{R}_+$, the conditional probability $P(\tau_0 = t \mid \xi = t)$ is given by

$$m_{\tau_0|\xi}(t \mid t) = \frac{p_{02} e^{-\tau_0 t}}{p_{01} \frac{v_1}{v_0 - v_1} (e^{-\tau_0 t} - e^{-\xi t}) + p_{02} e^{-\xi t}}$$

(2.14)
where \( \nu_0 = \lambda_{01} + \lambda_{02}, \nu_1 = \lambda_{12}, \) \( p_{01} = \frac{\lambda_{01}}{\lambda_{01} + \lambda_{02}} \) and \( p_{02} = \frac{\lambda_{02}}{\lambda_{01} + \lambda_{02}}. \)

**Proof.** A general proof can be found in Appendix 1.

Before we derive the formula for the likelihood function \( L(\lambda, \theta \mid C) \) in the general case for \( N \) failure histories and \( M \) suspension histories, we first consider the case with a single failure history \( F \), i.e. we have collected data \( \tilde{Y} = (y_1, \ldots, y_T) \) and the system is known to have failed at time \( \tilde{\xi} = t \), where \( T\Delta < t \leq (T + 1)\Delta \). Denote the observable data set \( O = \{F\} \), the complete data set \( C = \{\tilde{F}\} \), and the likelihood function \( L(\lambda, \theta \mid C) \) as \( L_F(\lambda, \theta) \). Since \( \tau_0 \) and \( \tilde{\xi} \) are sufficient for characterizing the sample path of the state process, equations (2.12) – (2.14) imply that the likelihood function \( L_F(\lambda, \theta) \) is given by

\[
L_F(\lambda, \theta) = \begin{cases} 
    g_{\tilde{Y} \mid \tilde{\xi}, \tau_0}(\tilde{Y} \mid t, \tau_0) f_{\tau_0/\xi}(\tau_0 \mid t) f_{\tilde{\xi}}(t), & \text{if } \tau_0 < t \\
    g_{\tilde{Y} \mid \tilde{\xi}, \tau_0}(\tilde{Y} \mid t, t) m_{\tau_0/\xi}(t \mid t) f_{\tilde{\xi}}(t), & \text{if } \tau_0 = t 
\end{cases} 
\]  
(2.15)

where \( g_{\tilde{Y} \mid \tilde{\xi}, \tau_0}(\tilde{Y} \mid t, s) \) is the conditional density function of the residual observation process \( \tilde{Y} = (Y_1, \ldots, Y_T) \) given \( \tilde{\xi} = t \) and \( \tau_0 = s \leq t \), which can be expressed in an explicit form. For any \( s \in ((k - 1)\Delta, k\Delta], k = 1, \ldots, T \), equation (2.8) implies that \( g_{\tilde{Y} \mid \tilde{\xi}, \tau_0}(\tilde{Y} \mid t, s) \) is given by

\[
g_{\tilde{Y} \mid \tilde{\xi}, \tau_0}(\tilde{Y} \mid t, s) = g_{\tilde{Y} \mid \tilde{\xi}, \tau_0}(\tilde{Y} \mid t, k\Delta) 
= \frac{1}{\sqrt{(2\pi)^d \det^{-1}(\Sigma_0) \det^{-1}(\Sigma_1)}} 
\exp \left\{-\frac{1}{2} \sum_{n=1}^{k-1} (y_n - \mu_0)' \Sigma_0^{-1} (y_n - \mu_0) 
- \frac{1}{2} \sum_{n=k}^{T} (y_n - \mu_1)' \Sigma_1^{-1} (y_n - \mu_1) \right\} 
\]  
(2.16)

and for any \( s > T\Delta \), \( g_{\tilde{Y} \mid \tilde{\xi}, \tau_0}(\tilde{Y} \mid t, s) \) is given by
\[
g_{\bar{Y}_{i_{0},t_{0}}} (Y | t, s) = g_{\bar{Y}_{i_{0},t_{0}}} (Y | t, t) = \frac{1}{\sqrt{(2\pi)^d \det (\Sigma_0)}} \exp \left( -\frac{1}{2} \sum_{n=1}^{T} (y_n - \mu_0)' \Sigma_0^{-1} (y_n - \mu_0) \right)
\] (2.17)

We next consider the case where we have observed only a single suspension history \( S \), i.e. we have collected data \( \bar{Y} = (y_1, \ldots, y_T) \) and stopped observing the operating system at time \( T \Delta \).

Denote the observable data set \( O = \{S\} \), the complete data set \( C = \{\bar{S}\} \), and the likelihood function \( L(\lambda, \theta | C) \) as \( L_{\bar{S}} (\lambda, \theta) \).

**Lemma 2.2** For each \( s, t \in \mathbb{R}_+ \), the conditional reliability function of \( \xi \) given \( \tau_0 \) is given by

\[
h(t | s) := P(\xi > t | \tau_0 = s) = \begin{cases} p_0 e^{-\tau_0 (t-s)}, & t \geq s \\ 1, & t < s \end{cases}
\] (2.18)

Furthermore, the density function of the unobservable sojourn time \( \tau_0 \) is given by

\[
f_{\tau_0} (s) = \begin{cases} \nu_0 e^{-\nu_0 s}, & s \geq 0 \\ 0, & s < 0 \end{cases}
\] (2.19)

Then equations (2.18) and (2.19) imply that the likelihood function \( L_{\bar{S}} (\lambda, \theta) \) is given by

\[
L_{\bar{S}} (\lambda, \theta) = g_{\bar{Y}_{i_{0},t_{0}}} (Y | t, \tau_0) h(t | \tau_0) f_{\tau_0} (\tau_0)
\] (2.20)

Thus, for the general case in which we have observed \( N \) independent failure histories \( F_1, \ldots, F_N \) and \( M \) independent suspension histories \( S_1, \ldots, S_M \), the likelihood function is given by
where the likelihood functions for the individual failure history and suspension history are given by equations (2.15) and (2.20), respectively.

### 2.3.2 Formula for the pseudo log-likelihood function

We first analyze the case in which we have observed only a single failure history \( F \) of the form \( \mathbf{Y} = (y_1, \ldots, y_T)' \) and \( \xi = t \), where \( T\Delta < t \leq (T + 1)\Delta \). Thus, for any fixed estimates \( \hat{\lambda}, \hat{\theta} \) of the state and residual observation parameters, we are interested in deriving the formula for the pseudo log-likelihood function \( Q_F(\lambda, \theta | \hat{\lambda}, \hat{\theta}) = E_{\hat{\lambda}, \hat{\theta}}(\ln L_F(\lambda, \theta) | F) \), where \( L_F(\lambda, \theta) \) is given in equation (2.15). To simplify the notation, for the remainder of the analysis we denote vectors \( \lambda = (\lambda_{01}, \lambda_{02}, \lambda_{12})' \) and \( g = (g_{\mathbf{Y} | t, \Delta}, \ldots, g_{\mathbf{Y} | t, T\Delta}, g_{\mathbf{Y} | t, t})' \). Also, for any vector \( v = (v_1, \ldots, v_n)' \), we denote \( \ln v = (\ln v_1, \ldots, \ln v_n)' \). The inner product \( \langle v, w \rangle := v'w \).

**Theorem 2.1** Given a single failure history \( F \), the pseudo log-likelihood function is decomposed as

\[
Q_F(\lambda, \theta | \hat{\lambda}, \hat{\theta}) = Q_F^{\text{state}}(\lambda | \hat{\lambda}, \hat{\theta}) + Q_F^{\text{obs}}(\theta | \hat{\lambda}, \hat{\theta})
\]  

(2.22)

where

\[
Q_F^{\text{state}}(\lambda | \hat{\lambda}, \hat{\theta}) = \langle \hat{\mathbf{a}}, \lambda \rangle + \langle \hat{\mathbf{b}}, \ln \lambda \rangle
\]

and

\[
Q_F^{\text{obs}}(\theta | \hat{\lambda}, \hat{\theta}) = \langle \hat{\mathbf{c}}, \ln g \rangle
\]

(2.23)

and vectors \( \hat{\mathbf{a}} = (\hat{a}_{01}, \hat{a}_{02}, \hat{a}_{12})' \), \( \hat{\mathbf{b}} = (\hat{b}_{01}, \hat{b}_{02}, \hat{b}_{12})' \) and \( \hat{\mathbf{c}} = (\hat{c}_1, \ldots, \hat{c}_r, \hat{c}_i)' \) that depend only on the fixed estimates \( \hat{\lambda} \) and \( \hat{\theta} \) are given by
\[
\hat{a}_{01} = \hat{a}_{02} = -\frac{\hat{p}_{01}|\hat{y}|}{d} \langle \hat{e}_2, \hat{g} \rangle - \frac{\hat{p}_{02}|\hat{y}|}{d} \hat{g}_{\xi, \tau_0}(\hat{y} | t, t)
\]
\[
\hat{a}_{12} = \frac{\hat{p}_{01}|\hat{y}|}{d} \left( \langle \hat{e}_2, \hat{g} \rangle - t \langle \hat{e}_1, \hat{g} \rangle \right)
\]
\[
\hat{b}_{01} = \hat{b}_{12} = \frac{\hat{p}_{01}|\hat{y}|}{d} \langle \hat{e}_1, \hat{g} \rangle
\]
\[
\hat{b}_{02} = \frac{\hat{p}_{02}|\hat{y}|}{d} \hat{g}_{\xi, \tau_0}(\hat{y} | t, t)
\]
\[
\hat{d} = \hat{p}_{01}\hat{u}_1 e^{-i(\hat{\lambda} - \hat{\theta}) t} \langle \hat{e}_1, \hat{g} \rangle + \hat{p}_{02}\hat{u}_2 e^{-i(\hat{\lambda} - \hat{\theta}) t} \hat{g}_{\xi, \tau_0}(\hat{y} | t, t)
\]
\[
\hat{c}_k = \frac{\hat{p}_{01}\hat{u}_1 e^{-i(\hat{\lambda} - \hat{\theta}) t} \hat{e}_k}{d} \hat{g}_{\xi, \tau_0}(\hat{y} | t, k\Delta), \quad k = 1, \ldots, T
\]
\[
\hat{c}_t = \left( \frac{\hat{p}_{01}\hat{u}_1 e^{-i(\hat{\lambda} - \hat{\theta}) t} \hat{e}_t + \hat{p}_{02}\hat{u}_2 e^{-i(\hat{\lambda} - \hat{\theta}) t}}{d} \right) \hat{g}_{\xi, \tau_0}(\hat{y} | t, t)
\]

and vectors \( \hat{e}_1 = (\hat{e}_1^1, \ldots, \hat{e}_1^T, \hat{e}_1^\prime) \) and \( \hat{e}_2 = (\hat{e}_2^1, \ldots, \hat{e}_2^T, \hat{e}_2^\prime) \) are defined by

\[
\hat{e}_1^k = \int_{(k-1)\Delta}^{k\Delta} e^{-(\hat{\lambda}_0 - \hat{\lambda}_1)u} du = \frac{e^{-(\hat{\lambda}_0 - \hat{\lambda}_1)(k-1)\Delta} - e^{-(\hat{\lambda}_0 - \hat{\lambda}_1)k\Delta}}{\hat{\lambda}_0 - \hat{\lambda}_1}
\]
\[
\hat{e}_1^\prime = \int_{(k-1)\Delta}^{k\Delta} e^{-(\hat{\lambda}_0 - \hat{\lambda}_1)u} du = \frac{e^{-(\hat{\lambda}_0 - \hat{\lambda}_1)k\Delta} - e^{-(\hat{\lambda}_0 - \hat{\lambda}_1)u}}{\hat{\lambda}_0 - \hat{\lambda}_1}
\]
\[
\hat{e}_2^k = \int_{(k-1)\Delta}^{k\Delta} u e^{-(\hat{\lambda}_0 - \hat{\lambda}_1)u} du = \frac{\hat{e}_k - k\Delta e^{-(\hat{\lambda}_0 - \hat{\lambda}_1)k\Delta} + (k-1)\Delta e^{-(\hat{\lambda}_0 - \hat{\lambda}_1)(k-1)\Delta}}{\hat{\lambda}_0 - \hat{\lambda}_1}
\]
\[
\hat{e}_2^\prime = \int_{(k-1)\Delta}^{k\Delta} u e^{-(\hat{\lambda}_0 - \hat{\lambda}_1)u} du = \frac{\hat{e}_1^\prime - te^{-(\hat{\lambda}_0 - \hat{\lambda}_1)t} + T\Delta e^{-(\hat{\lambda}_0 - \hat{\lambda}_1)T\Delta}}{\hat{\lambda}_0 - \hat{\lambda}_1}
\]

**Proof.** A general proof can be found in Appendix 2. \(\square\)

We next analyze the case in which we have observed only a single suspension history \( S \) of the form \( \hat{y} = (y_1, \ldots, y_T) \) and \( \xi > T\Delta \). That is, for any fixed estimates \( \hat{\lambda} \) and \( \hat{\theta} \), we are interested in deriving the formula for the pseudo log-likelihood function \( Q_S(\lambda, \theta | \hat{\lambda}, \hat{\theta}) = E_{\hat{\lambda}, \hat{\theta}} \left( \ln L_S(\lambda, \theta) | S \right) \), where \( L_S(\lambda, \theta) \) is given in equation (2.20).
Theorem 2.2 Given a single suspension history \( S \), the pseudo log-likelihood function is decomposed as

\[
Q_S(\lambda, \theta | \hat{\lambda}, \hat{\theta}) = Q_S^{\text{state}}(\lambda | \hat{\lambda}, \hat{\theta}) + Q_S^{\text{obs}}(\theta | \hat{\lambda}, \hat{\theta})
\]  
(2.24)

where

\[
Q_S^{\text{state}}(\lambda | \hat{\lambda}, \hat{\theta}) = \left\{ \hat{\alpha}, \hat{\lambda} \right\} + \hat{\gamma}_1 \ln(\hat{\lambda}_{01}) + \hat{\gamma}_2 \ln(\hat{\lambda}_{01} + \hat{\lambda}_{02})
\]
\[
Q_S^{\text{obs}}(\theta | \hat{\lambda}, \hat{\theta}) = \left\{ \hat{\beta}, \ln g \right\}
\]  
(2.25)

and vectors \( \hat{\alpha} = (\hat{\alpha}_{01}, \hat{\alpha}_{02}, \hat{\alpha}_{12})' \), \( \hat{\beta} = (\hat{\beta}_1, ..., \hat{\beta}_T, \hat{\beta}_I)' \), \( \hat{\gamma}_1 \) and \( \hat{\gamma}_2 \) that depend only on the fixed estimates \( \hat{\lambda} \) and \( \hat{\theta} \) are given by

\[
\hat{\alpha}_{01} = \frac{\hat{\lambda}_{01} e^{-\hat{\eta}_t}}{\hat{\delta}} \left( \langle \hat{e}_2, \hat{g} \rangle - (t + \hat{\nu}_0^{-1}) e^{-\hat{\eta}_t} \hat{g}_{\hat{\eta}, r_0}(Y | t, t) \right)
\]
\[
\hat{\alpha}_{02} = \frac{\hat{\lambda}_{02} e^{-\hat{\eta}_t}}{\hat{\delta}} \left( \langle \hat{e}_2, \hat{g} \rangle - t \langle \hat{e}_1, \hat{g} \rangle \right)
\]
\[
\hat{\alpha}_{12} = \frac{\hat{\lambda}_{02} e^{-\hat{\eta}_t}}{\hat{\delta}} \left( \langle \hat{e}_2, \hat{g} \rangle - t \langle \hat{e}_1, \hat{g} \rangle \right)
\]
\[
\hat{\gamma}_1 = \frac{\hat{\lambda}_{01} e^{-\hat{\eta}_t}}{\hat{\delta}} \langle \hat{e}_1, \hat{g} \rangle
\]
\[
\hat{\gamma}_2 = \frac{e^{-\hat{\eta}_t}}{\hat{\delta}} \hat{g}_{\hat{\eta}, r_0}(Y | t, t)
\]
\[
\hat{\delta} = \hat{\lambda}_{01} e^{-\hat{\eta}_t} \langle \hat{e}_1, \hat{g} \rangle + e^{-\hat{\eta}_t} \hat{g}_{\hat{\eta}, r_0}(Y | t, t)
\]
\[
\hat{\beta}_k = \frac{\hat{\lambda}_{01} e^{-\hat{\eta}_t} \hat{e}_{1k}^T}{\hat{\delta}} \hat{g}_{\hat{\eta}, r_0}(Y | t, k\Delta), \quad k = 1, ..., T
\]
\[
\hat{\beta}_I = \left( \hat{\lambda}_{01} e^{-\hat{\eta}_t} \hat{e}_{1I}^T + e^{-\hat{\eta}_t} \hat{e}_{1I} \right) \hat{g}_{\hat{\eta}, r_0}(Y | t, t)
\]

where vectors \( \hat{e}_1 = (\hat{e}_1^T, ..., \hat{e}_1^T) \) and \( \hat{e}_2 = (\hat{e}_2^T, ..., \hat{e}_2^T) \) are defined in Theorem 2.1.

Proof. A general proof can be found in Appendix 3. \( \square \)
Finally, for the general case in which we have observed \( N \) independent failure histories \( F_1, \ldots, F_N \) and \( M \) independent suspension histories \( S_1, \ldots, S_M \), Theorems 2.1 and 2.2 and equation (2.21) imply that the pseudo log-likelihood function is given by

\[
Q(\lambda, \theta | \hat{\lambda}, \hat{\theta}) = E_{\hat{\lambda}, \hat{\theta}} \left( \ln L(\lambda, \theta | C) | O \right) = \sum_{i=1}^{N} Q_{F_i}(\lambda, \theta | \hat{\lambda}, \hat{\theta}) + \sum_{j=1}^{M} Q_{S_j}(\lambda, \theta | \hat{\lambda}, \hat{\theta})
\]  

(2.26)

Thus, to evaluate the pseudo log-likelihood function for all available histories, it suffices to evaluate the pseudo log-likelihood function for individual failure and suspension histories separately. Equation (2.26) completes the E-step of the EM algorithm. In the next section, we solve the M-step of the EM algorithm and derive explicit parameter update formulae for the maximizers of the pseudo log-likelihood function in equation (2.26).

2.3.3 Maximization of the pseudo log-likelihood function

By Theorems 2.1 and 2.2, the pseudo log-likelihood function can be decomposed as

\[
Q(\lambda, \theta | \hat{\lambda}, \hat{\theta}) = Q^{\text{state}}(\lambda | \hat{\lambda}, \hat{\theta}) + Q^{\text{obs}}(\theta | \hat{\lambda}, \hat{\theta})
\]

(2.27)

where \( Q^{\text{state}}(\lambda | \hat{\lambda}, \hat{\theta}) \) is a function only of the state parameter \( \lambda \) and \( Q^{\text{obs}}(\theta | \hat{\lambda}, \hat{\theta}) \) is a function only of the residual observation parameter \( \theta \). This means that the M-step can be carried out separately for the state and residual observation parameters. Using equation (2.26) and Theorems 2.1 and 2.2, we solve for the stationary point of the state parameter \( \lambda \). Set

\[
\frac{\partial Q(\lambda, \theta | \hat{\lambda}, \hat{\theta})}{\partial \lambda_{01}} = \frac{\partial Q(\lambda, \theta | \hat{\lambda}, \hat{\theta})}{\partial \lambda_{02}} = \frac{\partial Q(\lambda, \theta | \hat{\lambda}, \hat{\theta})}{\partial \lambda_{12}} = 0
\]

(2.28)

After some algebra, it is not difficult to check that there is a unique stationary point \( \lambda^* = (\lambda_{01}^*, \lambda_{02}^*, \lambda_{12}^*) \) of the pseudo log-likelihood function given explicitly by
\[
\lambda_{01}^* = -\frac{\sum_{i=1}^{N} \hat{b}_{01}^i + \sum_{j=1}^{M} \hat{\gamma}_{1}^j + \sum_{j=1}^{M} \hat{\gamma}_{1}^j \left( \sum_{i=1}^{N} \hat{b}_{01}^i + \sum_{j=1}^{M} \hat{\gamma}_{1}^j \right)}{\sum_{i=1}^{N} \hat{\alpha}_{01}^i + \sum_{j=1}^{M} \hat{\alpha}_{01}^j}
\]

\[
\lambda_{02}^* = \lambda_{01}^* \frac{\sum_{i=1}^{N} \hat{b}_{02}^i}{\sum_{i=1}^{N} \hat{b}_{01}^i + \sum_{j=1}^{M} \hat{\gamma}_{1}^j}
\]

\[
\lambda_{12}^* = -\frac{\sum_{i=1}^{N} \hat{b}_{12}^i}{\sum_{i=1}^{N} \hat{\alpha}_{12}^i + \sum_{j=1}^{M} \hat{\alpha}_{12}^j}
\]

where \(\hat{a}^i = (\hat{a}_{01}^i, \hat{a}_{02}^i, \hat{a}_{12}^i)'\), \(\hat{b}^i = (\hat{b}_{01}^i, \hat{b}_{02}^i, \hat{b}_{12}^i)'\), \(\hat{\alpha}^i = (\hat{\alpha}_{01}^i, \hat{\alpha}_{02}^i, \hat{\alpha}_{12}^i)'\), \(\hat{\gamma}_{1}^j\) and \(\hat{\gamma}_{2}^j\) are given in Theorems 2.1 and 2.2. Similarly, we set

\[
\frac{\partial Q(\lambda, \theta | \hat{\lambda}, \hat{\theta})}{\partial \mu_0} = \frac{\partial Q(\lambda, \theta | \hat{\lambda}, \hat{\theta})}{\partial \mu_1} = \mathbf{0}
\]

(2.30)

and

\[
\frac{\partial Q(\lambda, \theta | \hat{\lambda}, \hat{\theta})}{\partial \Sigma_0^{-1}} = \frac{\partial Q(\lambda, \theta | \hat{\lambda}, \hat{\theta})}{\partial \Sigma_1^{-1}} = \mathbf{0}
\]

(2.31)

We find there is a unique stationary point of the residual observation parameter \(\theta' = (\mu_0^*, \mu_1^*, \Sigma_0^*, \Sigma_1^*)\) given by
where vectors

\[
\begin{align*}
n_1^i &= \left(0, \sum_{n=1}^{y_n}, \ldots, \sum_{n=T_i}^{y_n} \right) \\
n_2^i &= \left(\sum_{n=1}^{y_n}, \ldots, y_{T_i}, 0 \right) \\
n_3^i &= \left(0, \sum_{n=1}^{(y_n - \mu_0^*)(y_n - \mu_0^*)'}, \ldots, \sum_{n=T_i}^{(y_n - \mu_0^*)(y_n - \mu_0^*)'} \right) \\
n_4^i &= \left(\sum_{n=1}^{(y_n - \mu_1^*)(y_n - \mu_1^*)'}, \ldots, (y_{T_i} - \mu_i^*)(y_{T_i} - \mu_i^*)', 0 \right) \\
d_1^i &= (0,1,\ldots,T_i)' \\
d_2^i &= (T_i,\ldots,1,0)' 
\end{align*}
\]

This completes the M-step of the EM algorithm.

### 2.3.4 EM algorithm estimates for spectrometric oil data

For the spectrometric oil application, the residuals for all the 36 histories are calculated using equation (2.6). Using the EM algorithm estimates, i.e. equations (2.29) and (2.32), and the Euclidean norm stopping criterion \(\|\hat{\lambda}' - \theta' - (\hat{\lambda}, \hat{\theta})\| \leq 10^{-4}\), we obtain the following parameter estimation results.
Table 2.4 EM Algorithm Estimates for Spectrometric Oil Data

<table>
<thead>
<tr>
<th></th>
<th>Initial values</th>
<th>Update 1</th>
<th>Update 2</th>
<th>Optimal estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\lambda}_{01} )</td>
<td>0.0030</td>
<td>0.0410</td>
<td>0.0302</td>
<td>0.0303</td>
</tr>
<tr>
<td>( \hat{\lambda}_{02} )</td>
<td>0.0000</td>
<td>0.0001</td>
<td>0.0001</td>
<td>0.0001</td>
</tr>
<tr>
<td>( \hat{\lambda}_{12} )</td>
<td>0.1500</td>
<td>0.3510</td>
<td>0.3545</td>
<td>0.3548</td>
</tr>
<tr>
<td>( \hat{\mu}_0 )</td>
<td>1.5 (1.2)</td>
<td>1.1 (1.9)</td>
<td>1.1 (1.9)</td>
<td></td>
</tr>
<tr>
<td>( \hat{\mu}_1 )</td>
<td>11 (4.2)</td>
<td>4.2 (5.5)</td>
<td>4.1 (5.5)</td>
<td></td>
</tr>
<tr>
<td>( \hat{\Sigma}_0 )</td>
<td>11.2 (6.8)</td>
<td>7.2 (1.8)</td>
<td>7.2 (1.9)</td>
<td>7.2 (2.0)</td>
</tr>
<tr>
<td>( \hat{\Sigma}_1 )</td>
<td>6.8 (8.9)</td>
<td>1.8 (3.2)</td>
<td>1.9 (3.7)</td>
<td>2.0 (3.6)</td>
</tr>
<tr>
<td>( Q )</td>
<td>-1.78x10^3</td>
<td>-1.41x10^3</td>
<td>-1.39x10^3</td>
<td>-1.39x10^3</td>
</tr>
<tr>
<td>Computation time (sec)</td>
<td>-----</td>
<td>7.19</td>
<td>9.83</td>
<td>11.95</td>
</tr>
</tbody>
</table>

Table 2.4 shows that iterations of the EM algorithm take on average less than 10 seconds which is extremely fast for offline computations. Furthermore, the estimates converge rapidly in 3 iterations, which is an attractive feature for real applications. All computations were coded in MATLAB (R2008a) on an Intel Core 2 6420, 2.13 GHz PC with 2 GB RAM.

### 2.4 Hidden semi-Markov modeling and parameter estimation

In this section, we consider the parameter estimation problem for the HSMM. For general HSMMs, it is not feasible to derive the explicit update formulae for parameter estimates. We thus consider a simple extension to the 3-state HMM introduced before.

Assume that the deteriorating system’s condition can be categorized into one of three states: the healthy or “good as new” state (state 0), the unhealthy or deteriorated state (state 1), and the failure state (state 2). States 0 and 1 are unobservable, whereas failure of the system is immediately observable. We model the state process as a continuous-time homogeneous semi-Markov chain \( \{X_t : t \geq 0\} \) with state space \( \{0,1,2\} \). The sojourn time in state 0 has a 2-phase Erlang distribution with parameter \( \nu_0 = \lambda_{01} + \lambda_{02} \), and the sojourn time in state 1 has an
exponential distribution with parameter \( \nu_t = \lambda_{12} \), where \( \lambda_{01}, \lambda_{02}, \lambda_{12} \in (0, +\infty) \) are the unknown state parameters. The system is assumed to start in the healthy state, i.e. \( X_0 = 0 \). At equidistant sampling times \( \Delta, 2\Delta, \ldots, \Delta > 0 \), residual observations \( Y_1, Y_2, \ldots \in \mathbb{R}^d \) are obtained using equation (2.6). Suppose we have collected \( N \) failure histories and \( M \) suspension histories as described in Section 2.3. Next, we use the EM algorithm (equations (2.9) and (2.10)) to estimate \( \lambda \) and \( \theta \).

2.4.1 Formula for the likelihood function

For a failure history, the distributional properties of the sojourn time \( \tau_0 \) in the healthy state 0 and failure time \( \xi \) are given by the following lemma.

**Lemma 2.3** For each \( t \in \mathbb{R}_+ \), the density function of \( \xi \) is given by

\[
f_{\xi}(t) = t \nu_0 e^{-\nu_0 t} - \frac{p_{01} \nu_0}{\nu_1 - \nu_0} \left( 1 - \nu_0 t \right) \left( \frac{p_{01} \nu_0}{\nu_1 - \nu_0} \right) (v_0 e^{-\nu_0 t} - v_1 e^{-\nu_1 t}) \tag{2.33}
\]

For all non-negative \( s < t \), the conditional density function of \( \tau_0 \) given \( \xi \) is

\[
f_{\tau_0|\xi}(s | t) = \frac{p_{01} \nu_1 se^{-\nu_1 s} e^{(t_0-t_1)s}}{te^{-\nu_0 t} + \frac{p_{01}}{(\nu_1 - \nu_0)^2} (v_1 e^{-\nu_1 t} - v_0 e^{-\nu_0 t} + v_0 v_1 t e^{-\nu_1 t} - v_0^2 t e^{-\nu_0 t})} \tag{2.34}
\]

and for each \( t \in \mathbb{R}_+ \), the conditional probability \( P(\tau_0 = t | \xi = \xi) \) is given by

\[
m_{\tau_0|\xi}(t | t) = \frac{tp_{02} e^{-\nu_0 t}}{te^{-\nu_0 t} + \frac{p_{02}}{(\nu_1 - \nu_0)^2} (v_1 e^{-\nu_1 t} - v_0 e^{-\nu_0 t} + v_0 v_1 t e^{-\nu_1 t} - v_0^2 t e^{-\nu_0 t})} \tag{2.35}
\]

where \( \nu_0 = \lambda_{01} + \lambda_{02}, \nu_1 = \lambda_{12}, p_{01} = \frac{\lambda_{01}}{\lambda_{01} + \lambda_{02}} \) and \( p_{02} = \frac{\lambda_{02}}{\lambda_{01} + \lambda_{02}} \).
Proof. See Appendix 1.

Similarly as in Section 2.3, the likelihood function \( L_F(\lambda, \theta) \) is given by

\[
L_F(\lambda, \theta) = \begin{cases} g_{Y_i|t_0, \tau_0} (Y | t, \tau_0) f_{\xi|t_0} (\tau_0 | t) f_{\xi}(t), & \text{if } \tau_0 < t \\ g_{Y_i|t, \tau_0} (Y | t, t) m_{\xi|t} (t | t) f_{\xi}(t), & \text{if } \tau_0 = t \end{cases}
\] (2.36)

where \( g_{Y_i|t, \tau_0} (Y | t, s) \) is the conditional density function of the residual observation process \( \tilde{Y} = (Y_1, \ldots, Y_T) \) given \( \xi = t \) and \( \tau_0 = s \leq t \), which are given by equations (2.16) and (2.17).

For a suspension history, we have the following lemma.

Lemma 2.4 For each \( s, t \in \mathbb{R} \), the conditional reliability function of \( \xi \) given \( \tau_0 \) is given by

\[
h(t | s) := P(\xi > t | \tau_0 = s) = \begin{cases} p_0 e^{-\nu_0 (t-s)}, & t \geq s \\ 1, & t < s \end{cases}
\] (2.37)

Furthermore, the density function of the unobservable sojourn time \( \tau_0 \) is given by

\[
f_{\tau_0}(s) = \begin{cases} \nu_0^2 s e^{-\nu_0 s}, & s \geq 0 \\ 0, & s < 0 \end{cases}
\] (2.38)

Then the likelihood function \( L_T(\lambda, \theta) \) is given by

\[
L_T(\lambda, \theta) = g_{Y_i|t_0, \tau_0} (Y | t, \tau_0) h(t | \tau_0) f_{\tau_0}(\tau_0)
\] (2.39)
Thus, for the general case in which we have observed $N$ independent failure histories $F_1,\ldots,F_N$ and $M$ independent suspension histories $S_1,\ldots,S_M$, the likelihood function is given by

$$L(\lambda, \theta \mid C) = \prod_{i=1}^{N} L_{F_i}(\lambda, \theta) \prod_{j=1}^{M} L_{S_j}(\lambda, \theta)$$

(2.40)

where the likelihood functions for the individual failure history and suspension history are given by equations (2.36) and (2.39), respectively.

2.4.2 Formula for the pseudo log-likelihood function

Similarly, we first analyze the case in which we have observed only a single failure history $F$ of the form $T\mathbf{y} = (y_1,\ldots,y_T)'$ and $\xi = t$, where $T\Delta < t \leq (T+1)\Delta$. Thus, for any fixed estimates $\hat{\lambda},\hat{\theta}$ of the state and residual observation parameters, we are interested in deriving the formula for the pseudo log-likelihood function $Q_F(\lambda, \theta \mid \hat{\lambda}, \hat{\theta}) = E_{\hat{\lambda}, \hat{\theta}}(\ln L_F(\lambda, \theta) \mid F)$, where $L_F(\lambda, \theta)$ is given in equation (2.36).

Theorem 2.3 Given a single failure history $F$, the pseudo log-likelihood function is decomposed as

$$Q_F(\lambda, \theta \mid \hat{\lambda}, \hat{\theta}) = Q^{state}_F(\lambda \mid \hat{\lambda}, \hat{\theta}) + Q^{obs}_F(\theta \mid \hat{\lambda}, \hat{\theta})$$

(2.41)

where

$$Q^{state}_F(\lambda \mid \hat{\lambda}, \hat{\theta}) = \langle \hat{\mathbf{a}}, \ln \lambda \rangle + \langle \hat{\mathbf{b}}, \ln \lambda \rangle + \ln(\lambda_{01} + \lambda_{02}) + \hat{d}$$

$$Q^{obs}_F(\theta \mid \hat{\lambda}, \hat{\theta}) = \langle \hat{\mathbf{c}}, \ln \mathbf{g} \rangle$$

(2.42)
and vectors $\hat{\mathbf{a}} = (\hat{a}_{01}, \hat{a}_{02}, \hat{a}_{12})'$, $\hat{\mathbf{b}} = (\hat{b}_{01}, \hat{b}_{02}, \hat{b}_{12})'$, $\hat{\mathbf{c}} = (\hat{c}_1, \ldots, \hat{c}_r, \hat{c}_l)'$ and $\hat{d}_1$ that depend only on the fixed estimates $\hat{\lambda}, \hat{\theta}$ are given by

\[
\begin{align*}
\hat{a}_{01} &= \hat{a}_{02} = -\frac{\hat{p}_{01}\tilde{\lambda}_1 e^{-i\hat{t}_l}}{\hat{d}} \langle \tilde{\mathbf{e}}_1, \tilde{\mathbf{g}} \rangle + t^2 \hat{p}_{02} e^{-i\hat{t}_l} \tilde{\mathbf{g}}_{\tilde{\mathbf{y}}|\tilde{\mathbf{y}}, \hat{t}_l} (\tilde{\mathbf{y}}|t,t) \\
\hat{a}_{12} &= \frac{\hat{p}_{01}\tilde{\lambda}_2 e^{-i\hat{t}_l}}{\hat{d}} (\langle \tilde{\mathbf{e}}_3, \tilde{\mathbf{g}} \rangle - t \langle \tilde{\mathbf{e}}_2, \tilde{\mathbf{g}} \rangle) \\
\hat{b}_{01} &= \hat{b}_{12} = \frac{\hat{p}_{01}\tilde{\lambda}_2 e^{-i\hat{t}_l}}{\hat{d}} \langle \tilde{\mathbf{e}}_2, \tilde{\mathbf{g}} \rangle \\
\hat{b}_{02} &= \frac{\hat{p}_{02} e^{-i\hat{t}_l}}{\hat{d}} \tilde{\mathbf{g}}_{\tilde{\mathbf{y}}|\tilde{\mathbf{y}}, \hat{t}_l} (\tilde{\mathbf{y}}|t,t) \\
\hat{d} &= \frac{\hat{p}_{01}\tilde{\lambda}_1 e^{-i\hat{t}_l} \langle \tilde{\mathbf{e}}_2, \tilde{\mathbf{g}} \rangle + \hat{p}_{02} e^{-i\hat{t}_l} \tilde{\mathbf{g}}_{\tilde{\mathbf{y}}|\tilde{\mathbf{y}}, \hat{t}_l} (\tilde{\mathbf{y}}|t,t)}{\hat{d}} \\
\hat{c}_k &= \frac{\hat{p}_{01}\tilde{\lambda}_1 e^{-i\hat{t}_l} \tilde{e}_k}{\hat{d}} \tilde{\mathbf{g}}_{\tilde{\mathbf{y}}|\tilde{\mathbf{y}}, \hat{t}_l} (\tilde{\mathbf{y}}|t,k\Delta) \\
\hat{c}_l &= \left( \frac{\hat{p}_{01}\tilde{\lambda}_1 e^{-i\hat{t}_l} \tilde{\mathbf{e}}_l' + \hat{p}_{02} e^{-i\hat{t}_l}}{\hat{d}} \right) \tilde{\mathbf{g}}_{\tilde{\mathbf{y}}|\tilde{\mathbf{y}}, \hat{t}_l} (\tilde{\mathbf{y}}|t,t) \\
\hat{d}_1 &= \frac{\hat{p}_{01}\tilde{\lambda}_1 e^{-i\hat{t}_l} \int (s e^{(\hat{t}_l - \hat{t}_l)u} \ln s) \tilde{\mathbf{g}}_{\tilde{\mathbf{y}}|\tilde{\mathbf{y}}, \hat{t}_l} (\tilde{\mathbf{y}}|t,s) ds + (\hat{p}_{02} e^{-i\hat{t}_l} \ln t) \tilde{\mathbf{g}}_{\tilde{\mathbf{y}}|\tilde{\mathbf{y}}, \hat{t}_l} (\tilde{\mathbf{y}}|t,t)}{\hat{d}}
\end{align*}
\]

and vectors $\hat{\mathbf{e}}_1 = (\hat{e}_1^1, \ldots, \hat{e}_1^r, \hat{e}_1^l)'$, $\hat{\mathbf{e}}_2 = (\hat{e}_2^1, \ldots, \hat{e}_2^r, \hat{e}_2^r)'$ are defined in Theorem 2.1, and $\hat{\mathbf{e}}_3 = (\hat{e}_3^1, \ldots, \hat{e}_3^r, \hat{e}_3^l)'$ is defined by

\[
\begin{align*}
\hat{\mathbf{e}}_3^k &= \int_{(k-1)\Delta}^{k\Delta} u^2 e^{-(\hat{t}_l - \hat{t}_l)u} du = \frac{2\hat{e}_3^k - k^2 \Delta^2 e^{-(\hat{t}_l - \hat{t}_l)k\Delta} + (k-1)^2 \Delta^2 e^{-(\hat{t}_l - \hat{t}_l)(k-1)\Delta}}{\hat{\nu}_0 - \hat{\nu}_1} \\
\hat{\mathbf{e}}_3^l &= \int_{T\Delta}^{T\Delta} u^2 e^{-(\hat{t}_l - \hat{t}_l)u} du = \frac{2\hat{e}_3^l - T^2 \Delta e^{-(\hat{t}_l - \hat{t}_l)T\Delta} + T^2 \Delta^2 e^{-(\hat{t}_l - \hat{t}_l)T\Delta}}{\hat{\nu}_0 - \hat{\nu}_1}
\end{align*}
\]

**Proof.** See Appendix 2. \(\square\)
We next analyze the case in which we have observed only a single suspension history $S$ of the form $\bar{Y} = (y_1, \ldots, y_T)$ and $\xi > T\Delta$. That is, for any fixed estimates $\hat{\lambda}$ and $\hat{\theta}$, we are interested in deriving the formula for the pseudo log-likelihood function $Q_S(\lambda, \theta | \hat{\lambda}, \hat{\theta}) = E_{\hat{\lambda}, \hat{\theta}} \left( \ln L_S(\lambda, \theta) | S \right)$, where $L_S(\lambda, \theta)$ is given in equation (2.39).

**Theorem 2.4** Given a single suspension history $S$, the pseudo log-likelihood function is decomposed as

$$Q_S(\lambda, \theta | \hat{\lambda}, \hat{\theta}) = Q^\text{state}_S(\hat{\lambda}, \hat{\theta}) + Q^\text{obs}_S(\theta | \hat{\lambda}, \hat{\theta}) \quad (2.43)$$

where

$$Q^\text{state}_S(\hat{\lambda}, \hat{\theta}) = \langle \hat{\alpha}, \hat{\lambda} \rangle + \hat{\gamma}_1 \ln(\hat{\lambda}_{01}) + \hat{\gamma}_2 \ln(\hat{\lambda}_{01} + \hat{\lambda}_{02}) + \hat{d}_2$$

$$Q^\text{obs}_S(\theta | \hat{\lambda}, \hat{\theta}) = \{ \hat{\beta}, \ln g \} \quad (2.44)$$

for vectors $\hat{\alpha} = (\hat{\alpha}_{01}, \hat{\alpha}_{02}, \hat{\alpha}_{12})$, $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_T, \hat{\beta}_t), \hat{\gamma}_1, \hat{\gamma}_2$ and $\hat{d}_2$ that depend only on the fixed estimates $\hat{\lambda}, \hat{\theta}$ are given by

$$\hat{\alpha}_{01} = \hat{\alpha}_{02} = -\frac{\hat{\lambda}_{01} (\hat{\lambda}_{01} + \hat{\lambda}_{02}) e^{-\epsilon t} \langle \hat{e}_3, \hat{g} \rangle}{\hat{\delta}}$$

$$\hat{\alpha}_{12} = \frac{\hat{\lambda}_{01} (\hat{\lambda}_{01} + \hat{\lambda}_{02}) e^{-\epsilon t} \hat{g}_{y, \hat{\xi}, \hat{\tau}_0} (y | t, t) e^{-\epsilon t} (2t + t^2 (\hat{\lambda}_{01} + \hat{\lambda}_{02}) + 2(\hat{\lambda}_{01} + \hat{\lambda}_{02})^{-1})}{\hat{\delta}}$$

$$\hat{\beta}_t = \frac{\hat{\lambda}_{01} (\hat{\lambda}_{01} + \hat{\lambda}_{02}) e^{-\epsilon t} \hat{g}_{y, \hat{\xi}, \hat{\tau}_0} (y | t, k\Delta)}{\hat{\delta}}$$

$$\hat{\beta}_t = \frac{e^{-\epsilon t} (1 + (\hat{\lambda}_{01} + \hat{\lambda}_{02}) t)}{\hat{\delta}} \hat{g}_{y, \hat{\xi}, \hat{\tau}_0} (y | t, t)$$

$$\hat{\gamma}_1 = \frac{\hat{\lambda}_{01} (\hat{\lambda}_{01} + \hat{\lambda}_{02}) e^{-\epsilon t}}{\hat{\delta}} \langle \hat{e}_2, \hat{g} \rangle$$

$$\hat{\gamma}_2 = \frac{e^{-\epsilon t} (1 + (\hat{\lambda}_{01} + \hat{\lambda}_{02}) t)}{\hat{\delta}} \hat{g}_{y, \hat{\xi}, \hat{\tau}_0} (y | t, t) + 1$$
\[
\tilde{\delta} = \hat{\lambda}_{01}(\hat{\lambda}_{01} + \hat{\lambda}_{02})e^{-\hat{\gamma}_t} \left( \hat{e}_2, \hat{\mathbf{g}} \right) + e^{-\hat{\gamma}_t} \hat{\mathbf{g}}_{\mathbf{y} | \mathbf{y}}(\mathbf{y} | t, t) \left( 1 + (\hat{\lambda}_{01} + \hat{\lambda}_{02})t \right)
\]

\[
\tilde{d}_2 = \frac{\hat{\lambda}_{01}(\hat{\lambda}_{01} + \hat{\lambda}_{02})e^{-\hat{\gamma}_t} \int_{s \leq t} \left( se^{-\hat{\gamma}_t} \ln s \right) \hat{\mathbf{g}}_{\mathbf{y} | \mathbf{y}}(\mathbf{y} | t, s)ds + (\hat{\lambda}_{01} + \hat{\lambda}_{02})^2 \hat{\mathbf{g}}_{\mathbf{y} | \mathbf{y}}(\mathbf{y} | t, t) \int_{s \leq t} se^{-\hat{\gamma}_t} \ln sds}{\tilde{\delta}}
\]

and vectors \( \hat{\mathbf{e}}_1 = (\hat{e}_1^1, \ldots, \hat{e}_1^T, \hat{e}_1')', \hat{\mathbf{e}}_2 = (\hat{e}_2^1, \ldots, \hat{e}_2^T, \hat{e}_2')' \) are defined in Theorem 2.1, and \( \hat{\mathbf{e}}_3 = (\hat{e}_3^1, \ldots, \hat{e}_3^T, \hat{e}_3')' \) is defined in Theorem 2.3.

**Proof.** See Appendix 3. \( \square \)

Finally, for the general case in which we have observed \( N \) independent failure histories \( F_1, \ldots, F_N \) and \( M \) independent suspension histories \( S_1, \ldots, S_M \), Theorems 2.3 and 2.4 and equation (2.40) imply that the pseudo log-likelihood function is given by

\[
Q(\lambda, \theta | \hat{\lambda}, \hat{\theta}) = E_{\hat{\lambda}, \hat{\theta}} \left( \ln L(\lambda, \theta | C) | O \right)
\]

\[
= \sum_{i=1}^{N} Q_{F_i}(\lambda, \theta | \hat{\lambda}, \hat{\theta}) + \sum_{j=1}^{M} Q_{S_j}(\lambda, \theta | \hat{\lambda}, \hat{\theta})
\]

Equation (2.45) completes the E-step of the EM algorithm. In the next section, we solve the M-step of the EM algorithm and derive explicit parameter update formulae for the maximizers of the pseudo log-likelihood function in equation (2.45).

### 2.4.3 Maximization of the pseudo log-likelihood function

Using equation (2.45) and Theorems 2.3 and 2.4, we solve for the stationary point of the state parameter \( \hat{\lambda} \). Set

\[
\frac{\partial Q(\lambda, \theta | \hat{\lambda}, \hat{\theta})}{\partial \lambda_{01}} = \frac{\partial Q(\lambda, \theta | \hat{\lambda}, \hat{\theta})}{\partial \lambda_{02}} = \frac{\partial Q(\lambda, \theta | \hat{\lambda}, \hat{\theta})}{\partial \lambda_{12}} = 0
\]

(2.46)
After some algebra, it is not difficult to check that there is a unique stationary point \( \lambda^* = (\lambda_{01}^*, \lambda_{02}^*, \lambda_{12}^*) \) of the pseudo log-likelihood function given explicitly by

\[
\lambda_{01}^* = -\frac{\sum_{i=1}^{N} \hat{b}_{01}^i + \sum_{j=1}^{M} \hat{\gamma}_1^j + \left( N + \sum_{j=1}^{M} \hat{\gamma}_2^j \right) \left( \sum_{i=1}^{N} \hat{b}_{01}^i + \sum_{j=1}^{M} \hat{\gamma}_1^j \right)}{\sum_{i=1}^{N} \hat{a}_i^0 + \sum_{j=1}^{M} \hat{\alpha}_0^j}
\]

\[
\lambda_{02}^* = \lambda_{01}^* \frac{\sum_{i=1}^{N} \hat{b}_{02}^i}{\sum_{i=1}^{N} \hat{b}_{01}^i + \sum_{j=1}^{M} \hat{\gamma}_1^j}
\]

\[
\lambda_{12}^* = -\frac{\sum_{i=1}^{N} \hat{b}_{12}^i}{\sum_{i=1}^{N} \hat{a}_i^1 + \sum_{j=1}^{M} \hat{\alpha}_1^j}
\]

where \( \hat{a}_i = (\hat{a}_{01}^i, \hat{a}_{02}^i, \hat{a}_{12}^i)' \), \( \hat{b}_i = (\hat{b}_{01}^i, \hat{b}_{02}^i, \hat{b}_{12}^i)' \), \( \hat{\alpha}_i = (\hat{\alpha}_{01}^i, \hat{\alpha}_{02}^i, \hat{\alpha}_{12}^i)' \), \( \hat{\gamma}_1^j \) and \( \hat{\gamma}_2^j \) are given in Theorems 2.3 and 2.4. Similarly, we set

\[
\frac{\partial Q(\lambda, \theta | \hat{\lambda}, \hat{\theta})}{\partial \mu_0} = \frac{\partial Q(\lambda, \theta | \hat{\lambda}, \hat{\theta})}{\partial \mu_1} = 0 \tag{2.48}
\]

and

\[
\frac{\partial Q(\lambda, \theta | \hat{\lambda}, \hat{\theta})}{\partial \Sigma_0^{-1}} = \frac{\partial Q(\lambda, \theta | \hat{\lambda}, \hat{\theta})}{\partial \Sigma_1^{-1}} = 0 \tag{2.49}
\]

We find there is a unique stationary point of the residual observation parameter \( \theta^* = (\mu_0^*, \mu_1^*, \Sigma_0^*, \Sigma_1^*) \) given by
\[
\begin{align*}
\mu_0^* &= \frac{\sum_{i=1}^{N} n_i \cdot c_i + \sum_{j=1}^{M} n_j \cdot \hat{\beta}_j}{\sum_{i=1}^{N} \langle c_i, d_i \rangle + \sum_{j=1}^{M} \langle \hat{\beta}_j, d_j \rangle}, & \Sigma_0^* &= \frac{\sum_{i=1}^{N} n_i \cdot c_i + \sum_{j=1}^{M} n_j \cdot \hat{\beta}_j}{\sum_{i=1}^{N} \langle c_i, d_i \rangle + \sum_{j=1}^{M} \langle \hat{\beta}_j, d_j \rangle} \\
\mu_1^* &= \frac{\sum_{i=1}^{N} n_2 \cdot c_i + \sum_{j=1}^{M} n_j \cdot \hat{\beta}_j}{\sum_{i=1}^{N} \langle c_i, d_i \rangle + \sum_{j=1}^{M} \langle \hat{\beta}_j, d_j \rangle}, & \Sigma_1^* &= \frac{\sum_{i=1}^{N} n_i \cdot c_i + \sum_{j=1}^{M} n_j \cdot \hat{\beta}_j}{\sum_{i=1}^{N} \langle c_i, d_i \rangle + \sum_{j=1}^{M} \langle \hat{\beta}_j, d_j \rangle}
\end{align*}
\] (2.50)

where vectors
\[
\begin{align*}
n_i^1 &= \left(0, \sum_{n_{t_1}} y_n, \ldots, \sum_{n_{T_1}} y_n \right) \\
n_i^2 &= \left(\sum_{n_{t_1}} y_n, \ldots, y_{T_1}, 0 \right) \\
n_i^3 &= \left(0, \sum_{n_{t_1}} \left( y_n - \mu_{t_1} \right) \left( y_n - \mu_{t_1}^* \right), \ldots, \sum_{n_{T_1}} \left( y_n - \mu_{0} \right) \left( y_n - \mu_{0}^* \right) \right) \\
n_i^4 &= \left(\sum_{n_{t_1}} \left( y_n - \mu_{t_1} \right) \left( y_n - \mu_{t_1}^* \right), \ldots, \left( y_{T_1} - \mu_{t_1} \right) \left( y_{T_1} - \mu_{t_1}^* \right), 0 \right) \\
d_i^1 &= (0,1,\ldots,T_1)^\prime \\
d_i^2 &= (T_1,\ldots,1,0)^\prime
\end{align*}
\]

This completes the M-step of the EM algorithm.

2.4.4 Numerical example

In this section, we present a numerical example to illustrate the estimation procedure. We assume that the system deterioration follows a continuous-time homogeneous semi-Markov chain \( (X_t : t \geq 0) \) with state space \{0,1,2\}. The sojourn time in state 0 has a 2-phase Erlang distribution with parameter \( \nu_0 = \lambda_{01} + \lambda_{02} \), and the sojourn time in state 1 has an exponential distribution with parameter \( \nu_1 = \lambda_{12} \). The state parameters are given below.
The residual observation process is obtained through condition monitoring and is assumed to follow $N_2(\mu, \Sigma)$ when the system is in the healthy state and follow $N_2(\mu_1, \Sigma_1)$ when the system is in the unhealthy state, where

$$
\begin{align*}
\mu &= \begin{pmatrix} 1.0 \\ 1.2 \end{pmatrix}, \\
\mu_1 &= \begin{pmatrix} 4.5 \\ 5.0 \end{pmatrix}, \\
\Sigma &= \begin{pmatrix} 0.6 & 0.5 \\ 0.5 & 1.0 \end{pmatrix}, \\
\Sigma_1 &= \begin{pmatrix} 1.2 & 1.0 \\ 1.0 & 1.5 \end{pmatrix}.
\end{align*}
$$

Using these parameters, 60 data histories consisting of $M = 30$ suspension histories and $N = 30$ failure histories are generated. We run our estimation procedure on the generated data with a stopping criterion of $\| (\hat{\lambda}^*, \hat{\theta}^*) - (\hat{\lambda}, \hat{\theta}) \| \leq 10^{-4}$. The following parameter estimates are obtained.

### Table 2.5 Optimal Parameter Estimates using the EM Algorithm

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Initial Values</th>
<th>Update 1</th>
<th>Update 2</th>
<th>Final Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\lambda}_{01}$</td>
<td>0.30</td>
<td>0.21</td>
<td>0.17</td>
<td>0.16</td>
</tr>
<tr>
<td>$\hat{\lambda}_{02}$</td>
<td>0.10</td>
<td>0.07</td>
<td>0.06</td>
<td>0.05</td>
</tr>
<tr>
<td>$\hat{\lambda}_{12}$</td>
<td>0.50</td>
<td>0.43</td>
<td>0.39</td>
<td>0.39</td>
</tr>
<tr>
<td>$\hat{\mu}_0$</td>
<td>0.50</td>
<td>0.80</td>
<td>1.10</td>
<td>1.00</td>
</tr>
<tr>
<td>$\hat{\mu}_1$</td>
<td>4.00</td>
<td>4.20</td>
<td>4.40</td>
<td>4.60</td>
</tr>
<tr>
<td>$\hat{\Sigma}_0$</td>
<td>0.80</td>
<td>0.70</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>$\hat{\Sigma}_1$</td>
<td>1.00</td>
<td>1.00</td>
<td>1.10</td>
<td>1.10</td>
</tr>
<tr>
<td>$Q$</td>
<td>-842.7</td>
<td>-761.4</td>
<td>-714.3</td>
<td>-702.9</td>
</tr>
<tr>
<td>Computation time (sec)</td>
<td>-----</td>
<td>7.5</td>
<td>15.3</td>
<td>21.8</td>
</tr>
</tbody>
</table>
As in Table 2.4, Table 2.5 shows that iterations of the EM algorithm take on average less than 10 seconds to find the optimal parameter estimates. The estimates converge rapidly in 3 iterations, which is an attractive feature for practical applications. All computations were coded in MATLAB (R2008a) on an Intel Corel 2 6420, 2.13 GHz PC with 2 GB RAM.
Chapter 3 The Structure of the Optimal Bayesian Control Policy

3.1 Model formulation

We continue to consider the partially observable deteriorating system introduced in Chapter 2, which can be modeled in the HMM framework composed of the hidden state and residual observation processes. The system deterioration is characterized by a continuous-time homogeneous Markov chain \((X_t: t \geq 0)\) with state space \(S = \{0, 1, 2\}\). States 0 and 1 are unobservable, representing the healthy and unhealthy operational states respectively, and state 2 corresponds to the observable failure state. The instantaneous transition rates \(\lambda_{ij}, i, j \in S\), are defined by

\[
\lambda_{ij} = \lim_{h \to 0^+} \frac{P(X_h = j | X_0 = i)}{h} < +\infty, \quad i \neq j \in S
\]

To model monotonic system deterioration, we assume that the state process is non-decreasing with probability 1, i.e. \(\lambda_{ij} = 0\) for all \(j < i\). In particular, this implies that without corrective maintenance the failure state is absorbing. We also assume that the system is more likely to fail in the unhealthy state 1 than in the healthy state 0, i.e. \(\lambda_{02} < \lambda_{12}\). Upon system failure, corrective maintenance is carried out, which brings the system to the healthy state.

While the system is in operational state \(i \in \{0, 1\}\), residual observations \(Y_1, Y_2, \ldots \in \mathbb{R}^d\) are obtained periodically using equation (2.6) through condition monitoring. At sampling epoch \(n\Delta\), given that the system state \(X_{n\Delta} = i\), we assume that the \(d\) - dimensional vector \(Y_n\) is a continuous random vector with state-dependent conditional density function \(f(y | i)\).

At each sampling epoch, a decision is made either to allow the system to run without intervention or to stop the process and initiate a full system inspection. Upon full system inspection, if the system is found to be in state \(i = 0\), it is left operational, and if the system is
found to be in state $i = 1$, preventive maintenance is carried out. Full system inspection takes $T_0 > 0$ time units, and inspection plus preventive maintenance take $T_1 > 0$ time units. Upon system failure, corrective maintenance is carried out which takes $T_2 > 0$ time units. We assume that $T_0 < T_1 < T_2$.

The objective is to determine the optimal control policy that maximizes the long-run expected average availability per unit time. The problem can be formulated as a discrete-time optimal stopping problem with partial information, discussed as follows.

Let $F = (F_n : n \in \mathbb{N})$ be the natural filtration generated by the observable information at each sampling epoch

$$F_n = \sigma(Y_1, Y_2, \ldots, Y_n, I_{\{\xi > n\Delta\}})$$  \hspace{1cm} (3.2)

where $I_{\{\xi > n\Delta\}}$ is the indicator random variable defined as

$$I_{\{\xi > n\Delta\}} = \begin{cases} 1, & \text{if } \xi > n\Delta \\ 0, & \text{if } \xi \leq n\Delta \end{cases}$$  \hspace{1cm} (3.3)

The sigma-algebra $F_n$ defined in equation (3.2) represents the information collected until sampling epoch $n\Delta$. From renewal theory, the long-run expected average availability per unit time is calculated for any control policy as the expected system uptime incurred in one cycle divided by the expected cycle length, where a cycle is completed when the system is brought back to the healthy state. Then, for the average availability maximization criterion, the objective is to find an $F-$stopping time $\tau^*$, if it exists, maximizing the long-run expected average availability per unit time given by

$$\frac{E_{\tau^*}(U_T)}{E_{\tau^*}(CL)}$$  \hspace{1cm} (3.4)

where $U_T$ is the system uptime if full system inspection is scheduled at time $\tau\Delta$, $CL$ is the cycle length if full system inspection is scheduled at time $\tau\Delta$, and $E_{\tau^*}$ is the conditional expectation
given $\Pi_0 = P(X_0 = 1)$. We assume that the system always starts in the healthy state at the beginning of each cycle, i.e. $\Pi_0 = 0$. Since system uptime $UT_\tau$ is taken as the minimum time between full system inspection time $\tau\Delta$ and system failure time $\xi$, it follows that

$$UT_\tau = \tau\Delta \wedge \xi \quad (3.5)$$

where $\wedge$ is the min operator, i.e.

$$a \wedge b := \min\{a, b\}$$

Similarly, since the cycle length $CL_\tau$ is the sum of system uptime and system downtime, it follows that

$$CL_\tau = UT_\tau + \sum_{i \in S} T_i \{X_{\alpha,\xi} = i\} \quad (3.6)$$

where the second term on the right-hand side is the system downtime.

The optimal $F$-stopping time $\tau^*$ represents the first sampling epoch at which full system inspection should take place. It is important to realize however that since we are also considering mandatory corrective maintenance upon system failure, the random variable $\tau^* \Delta \wedge \xi$ represents the optimal time at which full system inspection or corrective maintenance should be carried out.

In the next section we derive the dynamic optimality equation, which will be analyzed to characterize the structure of the optimal control policy.

### 3.2 Derivation of the optimality equation

We first apply the $\lambda$-maximization technique to transform the stopping problem (3.4) to a parameterized stopping problem (with parameter $\lambda$) with an additive objective function. Define $\lambda > 0$,
\[ V^\lambda (\Pi_0) = \sup_{\tau} E_{\Pi_0} (Z^\lambda_{\tau}) \]  
\text{(3.7)}

where the supremum is taken over all \( F \) – stopping times \( \tau \) and

\[ Z^\lambda_{\tau} = U T_{\tau} - \lambda \cdot C L_{\tau} \]
\[ = (1 - \lambda)(\tau \Delta \wedge \xi) - \lambda \sum_{i \in S} T_i I_{\{X_{n+1} = i\}} \]  
\text{(3.8)}

Aven and Bergman (1986) showed that \( \lambda^* \) determined by the equation

\[ \lambda^* = \sup \{ \lambda > 0 : V^\lambda (\Pi_0) > 0 \} \]  
\text{(3.9)}

is the optimal expected average availability for the original stopping problem (3.4), and the \( F \) – stopping time \( \tau^* \) that maximizes the right-hand side of equation (3.7) for \( \lambda = \lambda^* \) determines the optimal stopping time. To simplify the notation, for the remainder of the analysis we suppress the dependence on \( \lambda \) and write, for example \( V^\lambda \) and \( Z^\lambda_{\tau} \) instead of \( V^\lambda (\Pi_0) \) and \( Z^\lambda_{\tau} (\Pi_0) \).

It is worthwhile noting that the optimal control problem discussed in this research can be formulated as an optimal stopping problem under partial observations, which is a special type of partially observable Markov decision process (POMDP), where only two actions at each sampling epoch are allowed, i.e. continue operating the system or stop the system for inspection. The general theory of POMDPs can be found in Smallwood and Sondik (1973), Sondik (1978). It is well known from the theory of POMDP that the conditional distribution of the system state \( X_{n+1} \) given \( F_n \), represents sufficient information for decision making at the \( n \)th decision epoch (see e.g. Bertsekas and Shreve 1978). However, since failure information is observable, the posterior probability that system is in the unhealthy state

\[ \Pi_n = P(X_{n+1} = 1 | F_n) \]  
\text{(3.10)}

is sufficient for decision making at the \( n \)th decision epoch.

Using Bayes’ Theorem, the evolution of the posterior probability \( \Pi_n \) is given by
\[ \Pi_n = P(X_{n\Delta} = 1 \mid F_n) \]
\[ = \frac{f(Y_n \mid 1)(P_{01}(\Delta)(1 - \Pi_{(n-1)}) + P_{11}(\Delta)\Pi_{(n-1)})}{f(Y_n \mid 0)P_{00}(\Delta)(1 - \Pi_{(n-1)}) + f(Y_n \mid 1)(P_{01}(\Delta)(1 - \Pi_{(n-1)}) + P_{11}(\Delta)\Pi_{(n-1)})} \]

(3.11)

where transition probabilities \( P_{ij}(t) = P(X_t = j \mid X_0 = i) \) for each \( i, j \in S \). The transition probabilities \( P_{ij}(t) \) can be solved explicitly by solving the Kolmogorov backward differential equations (see e.g. Tijms 1994) given by

\[
\begin{align*}
P_{00}'(t) &= -(\lambda_{01} + \lambda_{02})P_{00}(t) \\
P_{01}'(t) &= \lambda_{11}P_{11}(t) - (\lambda_{01} + \lambda_{02})P_{01}(t) \\
P_{11}'(t) &= -\lambda_{12}P_{11}(t)
\end{align*}
\]

(3.12)

Using the Laplace transform technique, we obtain the transition probabilities \( P_{ij}(t) \) as below

\[
P(t) = \begin{pmatrix} P_{ij}(t) \end{pmatrix}_{i,j \in S} = \begin{pmatrix} e^{-\lambda_{01}t} & \frac{\lambda_{01}(e^{-\lambda_{11}t} - e^{-(\lambda_{01} + \lambda_{02})t})}{\lambda_{01} + \lambda_{02} - \lambda_{12}} & 1 - e^{-(\lambda_{01} + \lambda_{02})t} - \frac{\lambda_{01}(e^{-\lambda_{11}t} - e^{-(\lambda_{01} + \lambda_{02})t})}{\lambda_{01} + \lambda_{02} - \lambda_{12}} \\
0 & e^{-\lambda_{12}t} & 1 - e^{-\lambda_{12}t} \\
0 & 0 & 1 \end{pmatrix}
\]

(3.13)

Next, we derive the dynamic optimality equation for the value function of the \( m \) - stage optimal stopping problem. For any non-random constant \( \Pi \in [0,1] \) and non-negative integer \( m \), define the \( m \) - stage value function \( V_m(\Pi) = \sup_{\tau \leq m} E_{\Pi}(Z_{\tau}) \), where \( E_{\Pi} \) is the conditional expectation given \( \Pi_0 = \Pi \).

From equation (3.8), we note that \( Z_{\tau} \) has the following representation
\[ Z_r = (1 - \lambda)(\tau \Delta + \xi) - \lambda \sum_{i \in S} T_i I_{(X_{n+i} \geq i)} \]
\[ = (1 - \lambda) \int_{s=0}^{\Delta} I_{(\xi \geq s)} ds - \lambda \sum_{i \in S} T_i I_{(X_{n+i} \geq i)} \]  
\[ = (1 - \lambda) \sum_{n=0}^{\tau-1} \left( \int_{s=n\Delta}^{(n+1)\Delta} I_{(\xi \geq s)} ds \right) - \lambda \sum_{i \in S} T_i I_{(X_{n+i} \geq i)} \]  

(3.14)

Then, by the dynamic programming algorithm (see e.g. Bertsekas and Shreve 1978, Davis 1993), for any \( \Pi \in [0,1] \), the \( m \)-stage value functions \( \{V_m(\Pi)\} \) satisfy

\[
V_{m+1}(\Pi) = \max \left\{ E_{\Pi_1} \left( Z_0 \right), \sup_{1 \leq r \leq m+1} E_{\Pi_r} \left( Z_r \right) \right\}
\]
\[
= \max \left\{ E_{\Pi_1} \left( Z_0 \right), (1 - \lambda)E_{\Pi_1} \left( \int_{s=0}^{\Delta} I_{(\xi \geq s)} ds \right) - \lambda T_2 E_{\Pi_1} \left( I_{(\xi \leq \Delta)} \right) + E_{\Pi_1} \left( I_{(\xi > \Delta)} V_m(\Pi) \right) \right\}
\]

(3.15)

with the boundary condition

\[
V_0(\Pi) = E_{\Pi_1} \left( Z_0 \right)
\]

(3.16)

where the first term in the brackets on the right-hand side of equation (3.15) is the expected ‘cost’ if the process is immediately stopped for full system inspection, and the second term is the expected ‘cost’ to run the system without intervention for one period, and then continue optimally with \( m \) periods left. Thus, using equation (3.8) and the probability transition matrix (3.13), equation (3.15) simplifies to

\[
V_{m+1}(\Pi) = \max \left\{ -\lambda(T_1 - T_0)\Pi - \lambda T_0, \right. \\
\left. -\lambda T_1 (P_{12}(\Delta) - P_{22}(\Delta)) - (1 - \lambda) \int_0^{\Delta} \left( P_{12}(s) - P_{22}(s) \right) ds \right\} \Pi \\
+ \left( -\lambda T_2 P_{22}(\Delta) + (1 - \lambda) \int_0^{\Delta} (1 - P_{22}(s)) ds \right) \\
+ \int_{R^d} V_m(\Pi_1(y_1, \Pi)) g(y_1 | \Pi) dy_1
\]

(3.17)

\[
V_0(\Pi) = -\lambda(T_1 - T_0)\Pi - \lambda T_0
\]
where
\[
g(y_1 | \Pi) = f(y_1 | 0)(1 - \Pi)P_{00}(\Delta) + f(y_1 | 1)((1 - \Pi)P_{01}(\Delta) + \Pi P_{11}(\Delta))
\]
\[
\Pi_1(y_1, \Pi) = P(X_\Delta = 1 | y_1, \Pi)
\] (3.18)

If we group terms as
\[
\begin{align*}
\theta_1 &= -\lambda T_2 (P_{12}(\Delta) - P_{02}(\Delta)) - (1 - \lambda) \int_0^\Delta (P_{12}(s) - P_{02}(s)) ds \\
\theta_2 &= -\lambda T_2 P_{02}(\Delta) + (1 - \lambda) \int_0^\Delta (1 - P_{02}(s)) ds
\end{align*}
\] (3.19)
equation (3.17) simplifies further to
\[
V_{m+1}(\Pi) = \max \left\{ -\lambda(T_1 - T_0)\Pi - \lambda T_0, \theta_1 \Pi + \theta_2 + \int_{R^d} V_m(\Pi_1(y_1, \Pi)) g(y_1 | \Pi) dy_1 \right\}
\] (3.20)

The next two lemmas show important properties of the \(m\)–stage value function \(V_m(\Pi)\).

**Lemma 3.1** For each \(m \geq 0\), \(V_m(\Pi)\) is a convex function of \(\Pi\).

**Proof.** We prove this property by mathematical induction.

Since \(V_0(\Pi) = -\lambda(T_1 - T_0)\Pi - \lambda T_0\) is linear in \(\Pi\), it is therefore convex in \(\Pi\). Assume now, that for some \(m \geq 0\), \(V_m(\Pi)\) is convex in \(\Pi\). We want to show that \(V_{m+1}(\Pi)\) is also convex in \(\Pi\). By equation (3.20), since the ‘max’ operator preserves convexity and \(\theta_1 \Pi + \theta_2\) is linear in \(\Pi\), it suffices to show that \(\int_{R^d} V_m(\Pi_1(y_1, \Pi)) g(y_1 | \Pi) dy_1\) is convex in \(\Pi\). For any \(0 \leq \Gamma_1 \leq \Gamma_2 \leq 1\) and \(\alpha_i \in [0, 1]\), define \(\Pi = \alpha_1 \Gamma_1 + (1 - \alpha_1) \Gamma_2\). Then equations (3.11), (3.18) and (3.19) imply
\[ \Pi_1(y_1, \Pi) = \alpha_2 \Pi_1(y_1, \Gamma_1) + (1 - \alpha_2) \Pi_1(y_1, \Gamma_2) \]  
(3.21)

where

\[ \alpha_2 = \frac{\alpha_1 g(y_1 | \Gamma_1)}{\alpha_1 g(y_1 | \Gamma_1) + (1 - \alpha_1) g(y_1 | \Gamma_2)} \in [0, 1] \]  
(3.22)

Then convexity of \( V_m(\Pi) \) implies

\[ \int_{\mathbb{R}^d} V_m(\Pi_1(y_1, \Pi)) g(y_1 | \Pi) dy_1 \]
\[ \leq \int_{\mathbb{R}^d} \left[ \alpha_2 V_m(\Pi_1(y_1, \Gamma_1)) + (1 - \alpha_2) V_m(\Pi_1(y_1, \Gamma_2)) \right] g(y_1 | \Pi) dy_1 \]
\[ = \alpha_1 \int_{\mathbb{R}^d} V_m(\Pi_1(y_1, \Gamma_1)) g(y_1 | \Gamma_1) dy_1 + (1 - \alpha_1) \int_{\mathbb{R}^d} V_m(\Pi_1(y_1, \Gamma_2)) g(y_1 | \Gamma_2) dy_1 \]

so that \( V_{m+1}(\Pi) \) is convex in \( \Pi \), which completes the proof. \[ \square \]

**Lemma 3.2** The value functions \( \{V_m(\Pi)\} \) are uniformly bounded

\[ V_m(\Pi) \leq \frac{|\theta_2|}{P_{02}(\Delta)} \]  
(3.24)

**Proof.** We prove this property by mathematical induction.

By equation (3.20), \( V_0(\Pi) \leq 0 \), which implies that \( V_0(\Pi) \leq \frac{|\theta_2|}{P_{02}(\Delta)} \). Assume now, that for some \( m \geq 0 \), \( V_m(\Pi) \leq \frac{|\theta_2|}{P_{02}(\Delta)} \). The model assumption \( \lambda_{02} < \lambda_{12} \) implies that for any \( \Delta \geq 0 \), \( P_{02}(\Delta) \leq P_{12}(\Delta) \), which implies that

\[ \theta_1 = -\lambda T_2 (P_{12}(\Delta) - P_{02}(\Delta)) - (1 - \lambda) \int_0^\Delta \left( P_{12}(s) - P_{02}(s) \right) ds \leq 0 \]  
(3.25)
and

\[
\int_{\mathbb{R}^d} g(y_1 | \Pi) dy_1
\]

\[
= \int_{\mathbb{R}^d} (1 - \Pi)P_{00}(\Delta)f(y_1 | 0)dy_1 + \int_{\mathbb{R}^d} ((1 - \Pi)P_{01}(\Delta) + \Pi P_{11}(\Delta)) f(y_1 | 1)dy_1
\]

\[
= (1 - \Pi)P_{00}(\Delta) + (1 - \Pi)P_{01}(\Delta) + \Pi P_{11}(\Delta)
\]

\[
= (1 - \Pi)(1 - P_{02}(\Delta)) + \Pi (1 - P_{12}(\Delta))
\]

\[
\leq \max \{1 - P_{02}(\Delta), 1 - P_{12}(\Delta)\}
\]

\[
= 1 - P_{02}(\Delta)
\]

(3.26)

Using the two inequalities above, and the assumption \(V_m(\Pi) \leq \frac{\theta_1}{P_{02}(\Delta)}\), it follows that

\[
V_{m+1}(\Pi) = \max \left\{-\lambda(T_1 - T_0)\Pi - \lambda T_0, \right. \\
\left. \theta_1 \Pi + \theta_2 + \int_{\mathbb{R}^d} V_m \left( \Pi_1(y_1, \Pi) \right) g(y_1 | \Pi) dy_1 \right\}
\]

\[
\leq \left| \theta_2 + \int_{\mathbb{R}^d} V_m \left( \Pi_1(y_1, \Pi) \right) g(y_1 | \Pi) dy_1 \right|
\]

\[
\leq |\theta_2| + \left| \int_{\mathbb{R}^d} V_m \left( \Pi_1(y_1, \Pi) \right) g(y_1 | \Pi) dy_1 \right|
\]

\[
\leq |\theta_2| + \frac{|\theta_2|}{P_{02}(\Delta)} \left| \int_{\mathbb{R}^d} g(y_1 | \Pi) dy_1 \right|
\]

\[
\leq |\theta_2| + \frac{|\theta_2|}{P_{02}(\Delta)} (1 - P_{02}(\Delta))
\]

\[
= \frac{|\theta_2|}{P_{02}(\Delta)}
\]

(3.27)

which completes the inductive step. Thus, for each \(m \geq 0\), \(V_m(\Pi) \leq \frac{\theta_1}{P_{02}(\Delta)}\).

□

Because \(V_m(\Pi)\) is the value function for the \(m\)–stage stopping problem, we have \(V_{m+1}(\Pi) \geq V_m(\Pi)\) for each \(m \geq 0\). Then, from Lemma 3.2, the value function
\[
\lim_{m \to \infty} V_m(\alpha) = V(\alpha)
\]
eexists, and from Lemma 3.1, \(V(\alpha)\) is a convex function of \(\alpha\) satisfying the following dynamic optimality equation

\[
V(\alpha) = \max \left\{ \begin{array}{l}
-\lambda(T_1 - T_0)\alpha - \lambda T_0, \\
\theta_1 \alpha + \theta_2 + \int_{\mathbb{R}^d} V(\Pi(y_1, \alpha)) g(y_1 \mid \alpha) dy_1
\end{array} \right\} 
\tag{3.28}
\]

In the next section, the dynamic optimality equation (3.28) is analyzed to determine the structure of the optimal control policy.

### 3.3 Structure of the optimal control policy

We now prove a result, which is important in showing that the optimal control policy has a control limit structure.

**Theorem 3.1** Under the model assumptions made in Section 3.1, the control policy that never stops the process for preventive maintenance, i.e. \(\tau = +\infty\), is not optimal.

**Proof.** Consider the age-based control policy that carries out preventive maintenance after \(n\) periods. From renewal theory, the long-run expected average availability per unit time for this policy is given by

\[
A(n) = \frac{E[\Delta \wedge \xi]}{E[\Delta \wedge \xi] + \sum_{i \in S} T_i P_0(n\Delta)} 
\tag{3.29}
\]

Thus, to prove the result, it suffices to show \(\arg \max_n A(n) < +\infty\).

Consider a new set of time parameters \(\bar{T}_0, \bar{T}_1, \bar{T}_2\), where \(\bar{T}_0 = \bar{T}_1 = T_1\) and \(\bar{T}_2 = T_2\). Then, if preventive maintenance is scheduled after \(n\) periods, the expected average availability using time parameters \(\bar{T}_0, \bar{T}_1, \bar{T}_2\) is given by
\[
B(n) = \frac{E[n\Delta \wedge \xi]}{E[n\Delta \wedge \xi] + \sum_{i \in S} T_i P_{0i}(n\Delta)}
\]

\[
= \frac{E[n\Delta \wedge \xi]}{E[n\Delta \wedge \xi] + T_1 P_{00}(n\Delta) + T_1 P_{01}(n\Delta) + T_2 P_{02}(n\Delta)}
\]

(3.30)

We now argue that

\[
n^*_A \equiv \arg \max_n A(n) \leq \arg \max_n B(n) \equiv n^*_B
\]

(3.31)

By definition of \(n^*_A\) and \(n^*_B\), for any \(n_1, n_2\),

\[
A(n_1) \leq A(n^*_A)
\]

\[
B(n_2) \leq B(n^*_B)
\]

(3.32)

In particular, for \(n_1 = n^*_B\) and \(n_2 = n^*_A\), after some straightforward algebra, the above two inequalities imply

\[
\frac{P_{00}(n^*_B \Delta)}{E[n^*_B \Delta \wedge \xi]} \leq \frac{P_{00}(n^*_A \Delta)}{E[n^*_A \Delta \wedge \xi]}
\]

(3.33)

Since \(P_{00}(n\Delta)\) is a decreasing function of \(n\) and \(E[n\Delta \wedge \xi]\) is an increasing function of \(n\), the ratio \(\frac{P_{00}(n\Delta)}{E[n\Delta \wedge \xi]}\) is a decreasing function of \(n\). Thus, the above inequality implies that \(n^*_A \leq n^*_B\).

Since we have assumed that \(\lambda_{02} < \lambda_{12}\), the failure rate function of \(\xi\) is therefore an increasing function. Thus, under the assumption that preventive maintenance time \(T_1\) is smaller than corrective maintenance time \(T_2\), Barlow and Hunter (1960) have shown that there exists a positive real value \(\tilde{t} < +\infty\), such that \(\tilde{t}\) is the unique maximizer of \(B(t)\), where

\[
B(t) = \frac{E[t \wedge \xi]}{E[t \wedge \xi] + T_1 P_{00}(t) + T_1 P_{01}(t) + T_2 P_{02}(t)}
\]
For our problem $\arg \max_n B(n)$ is required to be integer-valued. However, since $t^*$ is a unique maximizer, the function $B(t)$ is decreasing for $t > t^*$, and it follows that $\arg \max_n B(n) \leq \lfloor t^* \rfloor < +\infty$.

Thus, we have shown that

$$\arg \max_n A(n) \leq \arg \max_n B(n) \leq \lfloor t^* \rfloor < +\infty \quad (3.34)$$

which completes the proof.

We are now ready to state and prove the main result of this chapter.

**Theorem 3.2** The optimal control policy has a control limit structure. In particular, there exists a control limit $\Pi^* \in (0,1)$, such that the optimal control policy is determined by the following procedure. At sampling epoch $n\Delta$,

1. If $\Pi_n \geq \Pi^*$, initiate a full system inspection, and, if the system is found to be in the unhealthy state 1, perform preventive maintenance. Otherwise, if $\Pi_n < \Pi^*$, run the system until the next sampling epoch time $(n+1)\Delta$.

2. Corrective maintenance is carried out immediately upon system failure.

**Proof.** By Lemma 3.1, we know that the value function $V(\Pi)$ is a convex function of $\Pi$. Thus, the dynamic optimality equation (3.28) implies that the optimal control region, i.e. the set of $\Pi$ for which stopping the process is optimal, is a convex subset of $[0,1]$. Since it is not optimal to stop the process when $\Pi = 0$, to prove that a control limit policy is optimal, it suffices to show that when $\Pi = 1$, it is optimal to stop the process and initiate full system inspection. To prove this, we use mathematical induction. For $m = 1$, using equation (3.20),
\[ V_i(\Pi) = \max \left\{ -\lambda(T_i - T_0)\Pi - \lambda T_0, \right. \\
\left. \theta_1 \Pi + \int_{\mathbb{R}^d} V_0(\Pi(y_1, \Pi)) g(y_1 | \Pi) dy_1 \right\} \]

\[ = \max \left\{ -\lambda(T_i - T_0)\Pi - \lambda T_0, \right. \\
\left. \theta_1 + \lambda T_0 P_01(\Delta) + \lambda T_0 P_{00}(\Delta) \right\} \Pi \]

\[ = \max \{ V_{\text{stop}}^i(\Pi), V_{\text{run}}^i(\Pi) \} \]

(3.35)

We assume now that \( V_i(1) > V_{\text{stop}}^i(1) \) and get a contradiction. Since it is not optimal to carry out preventive maintenance in the healthy state, i.e. \( V_i(0) > V_{\text{stop}}^i(0) \), linearity of \( V_{\text{stop}}^i(\Pi) \) and \( V_{\text{run}}^i(\Pi) \) implies that \( V_i(\Pi) > V_{\text{stop}}^i(\Pi) \) for all \( \Pi \in [0,1] \). Since \( V_m(\Pi) \leq V_{m+1}(\Pi) \) for all \( m \), it follows that the limit \( V(\Pi) = \lim_{m \to \infty} V_m(\Pi) > V_{\text{stop}}^i(\Pi) \), and the policy that never stops the process, i.e. \( \tau = +\infty \), is optimal. This is a direct contradiction of Theorem 3.1. Thus, it follows that \( V_i(1) = V_{\text{stop}}^i(1) \), and by equation (3.35), we have the following inequality

\[ -\lambda(T_i - T_0) - \lambda T_0 \geq \theta_1 + \theta_2 - \lambda T_1 P_{11}(\Delta) \]

(3.36)

Suppose now that for some \( m \geq 1, V_m(1) = V_{\text{stop}}^m(1) \). Using the above inequality, it follows that

\[ V_{m+1}(1) = \max \left\{ -\lambda(T_i - T_0) - \lambda T_0, \right. \\
\left. \theta_1 + \theta_2 + \int_{\mathbb{R}^d} V_m(\Pi(y_1, 1)) g(y_1 | 1) dy_1 \right\} \]

\[ = \max \left\{ -\lambda(T_i - T_0) - \lambda T_0, \right. \\
\left. \theta_1 + \theta_2 + \int_{\mathbb{R}^d} V_1(1) g(y_1 | 1) dy_1 \right\} \]

\[ = \max \{ -\lambda(T_i - T_0) - \lambda T_0, \theta_1 + \theta_2 - \lambda T_1 P_{11}(\Delta) \} \]

(3.37)

which completes the inductive step. Therefore, the limit \( V(1) = \lim_{m \to \infty} V_m(1) = V_{\text{stop}}^i(1) \), which completes the proof. □
An illustration of the control limit policy is shown in Figure 3.1.

![Figure 3.1 Illustration of the Optimal Control Limit Policy](image)

The solid line denotes the value function $V(\Pi)$. Once the posterior probability $\Pi$ that system is in the unhealthy state exceeds the control limit $\Pi^*$, the optimal action is to stop the process and initiate a full system inspection.

### 3.4 Computation of the optimal control policy

In this section, we present an algorithm to compute the optimal expected long-run average availability and the optimal control limit. The computational algorithm is based on the $\lambda$–maximization technique (Aven and Bergman 1986), and the monotone convergence of the $m$–stage value functions $V_m(\Pi)$ to $V(\Pi)$. Aven and Bergman (1986) proved that for any $\lambda > 0$, if $\lambda > \lambda^*$ then the value function $V^\lambda(\Pi_0) < 0$, if $\lambda < \lambda^*$ then the value function $V^\lambda(\Pi_0) > 0$, and if $\lambda = \lambda^*$ then the value function $V^\lambda(\Pi_0) = 0$, where $\lambda^*$ is the optimal
average availability and \( \Pi_0 = 0 \). The authors also proved that the mapping \( \lambda \mapsto V^\lambda (\Pi_0) \) is continuous.

**The Algorithm**

**Step 1.** Choose \( \varepsilon > 0 \), the lower limit \( \lambda_{\text{lower}} = 0 \) and the upper limit \( \lambda_{\text{upper}} = 1 \).

**Step 2.** Let \( \lambda = (\lambda_{\text{lower}} + \lambda_{\text{upper}}) / 2 \), \( V_0(\Pi) = -\lambda (T_1 - T_0) \Pi - \lambda T_0 \), where \( 0 \leq \Pi \leq 1 \), and \( m = 1 \).

**Step 3.** Calculate \( V_m(\Pi) \), using the dynamic equation (3.20).

**Step 4.** If \( \sup \{ V_m(\Pi) - V_{m-1}(\Pi) \} \geq \varepsilon \), put \( m = m + 1 \), go to Step 3. Otherwise go to Step 5.

**Step 5.** Put the initial posterior probability \( \Pi_0 = 0 \).

If \( V_m(\Pi_0) \geq \varepsilon \), put \( \lambda_{\text{lower}} = \lambda \) and go to Step 2.

If \( V_m(\Pi_0) \leq -\varepsilon \), put \( \lambda_{\text{upper}} = \lambda \) and go to Step 2.

If \( -\varepsilon < V_m(\Pi_0) < \varepsilon \), put \( \lambda^* = \lambda \) and stop.

The optimal average availability is given by \( \lambda^* \), and the optimal control limit \( \Pi^* \) is the smallest value of \( \Pi \) such that \( V_m(\Pi) = -\lambda (T_1 - T_0) \Pi - \lambda T_0 \).

**Remark.** Since we are considering the long-run average availability maximization criterion, the optimal average availability \( \lambda^* \in [0,1] \). Thus, in Step 1 of the algorithm, the initial lower and upper bounds can be naturally chosen as \( \lambda_{\text{lower}} = 0 \) and \( \lambda_{\text{upper}} = 1 \). \( \varepsilon \) is a very small positive value, which is chosen in advance, such as \( 10^{-6} \).
Chapter 4 Bayesian Fault Prediction and Control Scheme

In this chapter, the optimal Bayesian fault prediction and control scheme considering the long-run average availability maximization objective along with a practical statistical constraint, has been proposed for the partially observable deteriorating system. The optimal control limit and sampling interval are calculated in the SMDP framework. Another Bayesian fault prediction and control scheme has been developed based on the ARL criterion. Formulae for the mean residual life and the distribution function of system residual life have been derived in explicit forms as functions of the posterior probability statistic. The advantages of the Bayesian model over previous well-known models are shown.

4.1 Bayesian control chart design for availability maximization

4.1.1 Problem formulation

We continue to assume that the deterioration process of a deteriorating system is described by a continuous-time homogeneous Markov chain \( (X_t : t \geq 0) \), with state space \( S = \{0,1,2\} \). States 0 and 1 are unobservable, representing the healthy and unhealthy operational states respectively, and state 2 corresponds to the observable failure state. The residual observation process \( (Y_n : n \in \mathbb{N}) \) is obtained using equation (2.6) through condition monitoring. When the system is in state 0, the residual observation \( Y_n | X_{n\Delta} = 0 \sim N_d(\mu_0, \Sigma_0) \), and when the system is in state 1, the residual observation \( Y_n | X_{n\Delta} = 1 \sim N_d(\mu_1, \Sigma_1) \), where \( \mu_0, \mu_1 \in \mathbb{R}^d, \Sigma_0, \Sigma_1 \in \mathbb{R}^{d \times d} \) are known model parameters. The conditional density function of the residual observation \( Y_n \) given \( X_{n\Delta} = x \), \( x = 0,1 \), is given in equation (2.8). When the system fails, a failure signal \( \eta \in \mathbb{R}^d \) is generated and failure replacement is carried out immediately.
Assume the system always starts in the healthy state, i.e., \( X_0 = 0 \), and the unobservable state process transits from state \( i \) to state \( j \) with rate \( \lambda_{ij} \), \( i, j \in S \). Thus, for a system that undergoes a replacement, the instantaneous transition rate matrix of the system is given by

\[
\Lambda = (\lambda_{ij})_{i,j \in S} = \begin{pmatrix}
-(\lambda_{01} + \lambda_{02}) & \lambda_{01} & \lambda_{02} \\
\lambda_{10} & -(\lambda_{10} + \lambda_{12}) & \lambda_{12} \\
0 & 0 & 0
\end{pmatrix}
\]  

(4.1)

Let \( P_{ij}(t) = P(X_t = j \mid X_0 = i) \) be the transition probabilities of the state process. The transition probability matrix \( P(t) = (P_{ij}(t))_{i,j \in S} \), provided below, is obtained by solving the Kolmogorov backward differential equations. Details can be found in Appendix 4.

\[
P(t) = (P_{ij}(t))_{i,j \in S} = \begin{pmatrix}
e^{-\beta t} + \frac{\lambda_{00} + \lambda_{02} - \alpha}{\beta - \alpha} (e^{-\beta t} - e^{-\alpha t}) & \frac{\lambda_{01} + \lambda_{02} - \alpha}{\beta - \alpha} (e^{-\beta t} - e^{-\alpha t}) & 1 - e^{-\beta t} - \frac{\lambda_{00} + \lambda_{02} + \lambda_{01} - \alpha}{\beta - \alpha} (e^{-\beta t} - e^{-\alpha t}) \\
\frac{\lambda_{00} + \lambda_{02} - \alpha}{\beta - \alpha} (e^{-\alpha t} - e^{-\beta t}) & e^{-\beta t} + \frac{\lambda_{00} + \lambda_{02} - \alpha}{\beta - \alpha} (e^{-\beta t} - e^{-\alpha t}) & 1 - e^{-\beta t} - \frac{\lambda_{00} + \lambda_{02} + \lambda_{01} - \alpha}{\beta - \alpha} (e^{-\beta t} - e^{-\alpha t}) \\
0 & 0 & 1
\end{pmatrix}
\]  

(4.2)

where

\[
\alpha = \frac{\lambda_{01} + \lambda_{02} + \lambda_{10} + \lambda_{12} - \kappa^{1/2}}{2}
\]

\[
\beta = \frac{\lambda_{01} + \lambda_{02} + \lambda_{10} + \lambda_{12} + \kappa^{1/2}}{2}
\]

\[
\kappa = \lambda_{01}^2 + \lambda_{02}^2 + \lambda_{10}^2 + \lambda_{12}^2 + 2\lambda_{01}\lambda_{02} + 2\lambda_{10}\lambda_{12} + 2\lambda_{01}\lambda_{02} + 2\lambda_{01}\lambda_{10} - 2\lambda_{02}\lambda_{10} - 2\lambda_{01}\lambda_{12} - 2\lambda_{02}\lambda_{12}
\]  

(4.3)

4.1.2 The Bayesian control chart approach for availability maximization

In this section, we describe the Bayesian control chart approach used in this research for availability maximization. The Bayesian control chart has been well studied in the control literature for partially observable systems (see e.g. Calabrese 1995, Makis 2008, 2009, Nenes and Tagaras 2007). For systems characterized by two unobservable operational states and one
observable failure state, we have already proved in Chapter 3 that the Bayesian control approach is the optimal strategy for the availability maximization objective.

The Bayesian control chart monitors the posterior probability $\Pi_n$ that the system is in the unhealthy state, which is defined in equation (3.10). The general approach is illustrated schematically in Figure 4.1. At sampling epoch $n\Delta$, if $\Pi_n$ exceeds the control limit, a system inspection is performed to check whether the system is in the healthy state or in the unhealthy state. If the system is found to be in state 0, it can be treated as a new system because the sojourn time in state 0 is exponentially distributed, which has the memoryless property. The system is left operational without further repairs or replacement. Otherwise preventive maintenance is triggered. Failures can occur at any time in a system cycle. If a failure occurs, failure replacement is carried out immediately. We assume that after an inspection, repair, or replacement, a new system cycle begins for both the state process and the residual observation process, independent of the previous cycle.

![Figure 4.1 The Bayesian Control Approach](image)

Let $T_0, T_1, T_2 > 0$ denote the time durations of system inspection, system inspection plus preventive maintenance, and failure replacement, respectively. Naturally, we have $T_0 < T_1 < T_2$. Let $\bar{T}$ denote the first time at which the posterior probability that system is in the unhealthy state exceeds the control limit on the Bayesian control chart. The system cycle length is given by

$$CL = I_{\{T_0 < \xi, X_0 = 0\}} \cdot (\bar{T} + T_0) + I_{\{T_0 < \xi, X_0 = 1\}} \cdot (\bar{T} + T_1) + I_{\{T_0 < \xi\}} \cdot (\xi + T_2)$$

(4.4)

and the system uptime in one cycle is given by
\[ UT = \min \{ \tilde{T}, \xi \} \] (4.5)

The objective is to find the optimal values of the sampling interval \( \Delta^* > 0 \) and the control limit \( \Pi^* \in (0,1) \) such that the long-run expected average availability per unit time is maximized. From renewal theory, the availability maximization problem is equivalent to finding optimal values of \( \Delta^* \) and \( \Pi^* \), such that

\[ g(\Pi^*, \Delta^*) = \sup_{\Pi \in (0,1), \Delta > 0} g(\Pi, \Delta) \] (4.6)

where

\[ g(\Pi, \Delta) = \frac{E_{\Pi, \Delta}(UT)}{E_{\Pi, \Delta}(CL)} \] (4.7)

In the next section, we develop an efficient computational algorithm in the SMDP framework to determine the optimal values of the sampling interval \( \Delta^* \) and control limit \( \Pi^* \in [0,1] \) that maximize the long-run expected average availability per unit time.

### 4.1.3 Computational algorithm in the SMDP framework

Typically, computing the long-run average availability in the SMDP framework requires discretization of \([0,1]\), the state space of the posterior probability process \( (\Pi_n : n \in \mathbb{N}) \). References that study discretization under a more general setting than the model considered in this research include Chow and Tsitsiklis (1988), Chow and Tsitsiklis (1991), Guihenneuc-Jouyaux and Robert (1998).

For fixed sampling interval \( \Delta \) and control limit \( \Pi \), we first define the state space of the SMDP. Suppose at sampling time \( n\Delta \) the system has not failed, i.e. \( \xi > n\Delta \), and we obtain the posterior probability \( \Pi_n \). Then for a fixed large \( L \), the SMDP is defined to be in state \( 1 \leq l \leq L \), if the current value \( \Pi_n \) of the posterior probability lies in the interval \( \left[ \frac{l-1}{L}, \frac{l}{L} \right] \). We denote the
set $K_1 = \{1, \ldots, L\}$. We have found in practice that when $L \geq 30$, the partition leads to a high degree of precision, so $L$ does not need to be chosen very large. This makes the computational algorithm presented below extremely fast, which is an attractive feature for practical applications. If the posterior probability $\Pi_n$ is above the control limit, i.e. $\Pi_n \geq \overline{\Pi}$, and upon full system inspection the system is found to be in state 0, i.e. $X_{nh} = 0$, the SMDP is defined to be in state $L + 1$. Similarly, if the posterior probability is above the control limit, i.e. $\Pi_n \geq \overline{\Pi}$, and upon full system inspection the system is found to be in state 1, i.e. $X_{nh} = 1$, the SMDP is defined to be in state $L + 2$. We denote the set $K_2 = \{L + 1, L + 2\}$. Lastly, the SMDP is defined to be in state $L + 3$ upon observable system failure. We denote the set $K_3 = \{L + 3\}$. Thus, the state space for the SMDP is given by $K = K_1 \cup K_2 \cup K_3$.

For the availability maximization problem, the SMDP is determined by the following quantities (Tijms 1994):

$p_{lm} =$ the probability that at the next decision epoch the SMDP will be in state $m \in K$ given the current state is $l \in K$.

$\tau_i =$ the mean sojourn time until the next decision epoch given the current state is $l \in K$.

$c_i =$ the mean cost incurred until the next decision epoch given the current state is $l \in K$.

From the theory of SMDP, for given sampling interval $\Delta$ and control limit $\overline{\Pi}$, the long-run expected average availability per unit time $g(\overline{\Pi}, \Delta) = \frac{E_{\Pi,\Delta}(UT)}{E_{\Pi,\Delta}(CL)}$ can be obtained by solving the following linear equations (e.g. Tijms 1994)

$$
u_k = g(\overline{\Pi}, \Delta) \cdot \tau_k - c_k + \sum_{l \in K} p_{kl} \cdot \nu_l, \text{ for each } k \in K \quad (4.8)$$

$$
u_s = 0, \text{ for some } s \in K \quad (4.9)$$
where the quantities \( n_k \) are known in the SMDP literature as the ‘relative values’ for the control limit policy (parameterized by \( \Delta \) and \( \Pi \)) when starting in state \( k \). See, for example, Tijms (1994) for a discussion of this terminology. The optimal control limit and sampling interval

\[
(\Pi', \Delta') = \text{arg sup}_{\Pi \in (0,1)} g(\Pi, \Delta)
\]

which maximize the long-run average availability \( g(\Pi, \Delta) \) can be obtained by solving the system of equations (4.8) and (4.9) considering different combinations of \( \Pi \) and \( \Delta \).

The remainder of the mathematical analysis in this section is devoted to deriving closed-form expressions for \( p_{lm}, \tau, \) and \( c_l, l, m \in K \).

From equation (3.11), we know that \( \Pi_n \) can be expressed recursively as

\[
\Pi_n = P(X_{n\Delta} = 1 | Y_1, \ldots, Y_n, \xi > n\Delta) = \frac{f(Y_n | 1)(P_{01}(\Delta)(1 - \Pi_{(n-1)}) + P_{11}(\Delta)\Pi_{(n-1)})}{f(Y_n | 0)P_{00}(\Delta)(1 - \Pi_{(n-1)}) + f(Y_n | 1)(P_{01}(\Delta)(1 - \Pi_{(n-1)}) + P_{11}(\Delta)\Pi_{(n-1)})}
\]  

(4.11)

where \( P_{ij}(\Delta) \) are given in equation (4.2).

Under the assumption \( \Sigma_0 \neq \Sigma_1 \), which is common in maintenance applications, we have

\[
\frac{f_{Y_n \mid X_{nm}}(y | 0)}{f_{Y_n \mid X_{nm}}(y | 1)} = \frac{\left((2\pi)^d |\Sigma_0| \right)^{-1/2} \exp \left(-\frac{1}{2} (y - \mu_0)' \Sigma_0^{-1} (y - \mu_0) \right)}{\left((2\pi)^d |\Sigma_1| \right)^{-1/2} \exp \left(-\frac{1}{2} (y - \mu_1)' \Sigma_1^{-1} (y - \mu_1) \right)}
\]

(4.12)

\[
= \left(|\Sigma_1| \cdot |\Sigma_0| \right)^{1/2} \exp \left(\frac{1}{2} ((Y_n - B)' A (Y_n - B) + C) \right)
\]

where constants \( A, B \) and \( C \) are given by
\[ A = \Sigma_1^{-1} - \Sigma_0^{-1} \]
\[ B = (\Sigma_1^{-1} - \Sigma_0^{-1})^{-1}(\Sigma_1^{-1} \mu_1 - \Sigma_0^{-1} \mu_0) \]
\[ C = (\mu_1' \Sigma_1^{-1} \mu_1 - \mu_0' \Sigma_0^{-1} \mu_0) - B(\Sigma_1^{-1} \mu_1 - \Sigma_0^{-1} \mu_0) \] (4.13)

Define

\[ V_n = (Y_n - B)'A(Y_n - B) \] (4.14)

so that equation (4.11) simplifies to

\[ \Pi_n = \frac{c_{\Pi(n-i)}^1}{(|\Sigma_1| \cdot |\Sigma_0|)^{1/2} \cdot \exp \left( \frac{1}{2} (V_n + C) \right) \cdot c_{\Pi(n-i)}^0 + c_{\Pi(n-i)}^1} \] (4.15)

where

\[ c_{\Pi(n-i)}^0 = P_{00}(\Delta) \cdot (1 - \Pi_{(n-1)}) + P_{01}(\Delta) \cdot \Pi_{(n-1)} \]
\[ c_{\Pi(n-i)}^1 = P_{01}(\Delta) \cdot (1 - \Pi_{(n-1)}) + P_{11}(\Delta) \cdot \Pi_{(n-1)} \] (4.16)

and transition probabilities \( P_j(\Delta) \) are given in equation (4.2). At sampling epoch \( n\Delta \), for any \( t \geq 0 \), we define the conditional reliability function given by

\[ R(t \mid \Pi_n) = P(\xi > n\Delta + t \mid \xi > n\Delta, Y_1, \ldots, Y_n, \Pi_n) \]
\[ = P(X_{n\Delta+t} \neq 2 \mid \xi > n\Delta, Y_1, \ldots, Y_n, \Pi_n) \]
\[ = \sum_{j=0,1} \left( (1 - \Pi_n)P_{0j}(t) + \Pi_nP_{1j}(t) \right) \]
\[ = (1 - \Pi_n)(1 - P_{02}(t)) + \Pi_n(1 - P_{12}(t)) \] (4.17)

Then, for each \( m \in K_1 \), the SMDP transition probabilities \( p_{lm} \) can be computed as follows

\[ p_{lm} = P(\theta_1 \leq \Pi_n < \theta_2 \mid \xi > (n-1)\Delta, Y_1, \ldots, Y_{(n-1)}, \Pi_{(n-1)}) \]
\[ = P(\theta_1 \leq \Pi_n < \theta_2 \mid \xi > n\Delta, Y_1, \ldots, Y_{(n-1)}, \Pi_{(n-1)}) \cdot P(\xi > n\Delta \mid \xi > (n-1)\Delta, Y_1, \ldots, Y_{(n-1)}, \Pi_{(n-1)}) \]
\[ = P(\theta_1 \leq \Pi_n < \theta_2 \mid \xi > n\Delta, Y_1, \ldots, Y_{(n-1)}, \Pi_{(n-1)}) \cdot R(\Delta \mid \Pi_{(n-1)}) \] (4.18)
where \( \theta_1 = \frac{m-1}{L} \pi \) and \( \theta_2 = \frac{m}{L} \pi \). The second term in equation (4.18) can be computed using the condition reliability function defined in equation (4.17). Since the SMDP at time \((n-1)\Delta\) is in state \( l \in K_1 \), for large \( L \) the length of the interval \( \left[ \frac{l-1}{L} \pi, \frac{l}{L} \pi \right] \) is small, and therefore \( \Pi_{(n-1)} \) can be approximated by the mid-point of the interval given by \( \frac{l-0.5}{L} \pi \). The first term in equation (4.18) can be computed generally as follows

\[
P(\theta_1 \leq \Pi_n < \theta_2 \mid \xi > n\Delta, Y_1, \ldots, Y_{n-1}, \Pi_{(n-1)}) = P\left( 2 \ln \left( \frac{(1-\theta_1)c^1_0}{\theta_2c^0_{\Pi_{(n-1)}}} \left( |\Sigma_0| \cdot |\Sigma_1| \right)^{1/2} - C < V_n \leq 2 \ln \left( \frac{(1-\theta_1)c^1_0}{\theta_2c^0_{\Pi_{(n-1)}}} \left( |\Sigma_0| \cdot |\Sigma_1| \right)^{1/2} - C \mid X_{n\Delta} = 0 \right) \right) \right.
\]

\[
+P\left( 2 \ln \left( \frac{(1-\theta_2)c^1_0}{\theta_2c^0_{\Pi_{(n-1)}}} \left( |\Sigma_0| \cdot |\Sigma_1| \right)^{1/2} - C < V_n \leq 2 \ln \left( \frac{(1-\theta_2)c^1_0}{\theta_2c^0_{\Pi_{(n-1)}}} \left( |\Sigma_0| \cdot |\Sigma_1| \right)^{1/2} - C \mid X_{n\Delta} = 1 \right) \right) \right)
\]

\[
= T_n(\theta_1, \theta_2 \mid \xi > n\Delta, \Pi_{(n-1)}), \left( \frac{c^0_{\Pi_{(n-1)}}}{c^1_{\Pi_{(n-1)}} + c^0_{\Pi_{(n-1)}}} \right) + T_n(\theta_1, \theta_2 \mid \xi > n\Delta, \Pi_{(n-1)}), \left( \frac{c^1_{\Pi_{(n-1)}}}{c^1_{\Pi_{(n-1)}} + c^0_{\Pi_{(n-1)}}} \right)
\]
\[ p_{t, L+3} = P(\xi \leq n\Delta \mid \xi > (n-1)\Delta, Y_1, \ldots, Y_{(n-1)}, \Pi_{(n-1)}) \]
\[ = 1 - P(\xi > (n-1)\Delta, Y_1, \ldots, Y_{(n-1)}, \Pi_{(n-1)}) \]
\[ = 1 - R(\Delta \mid \Pi_{(n-1)}) \]  
(4.20)

\[ p_{t, L+1} = P(\theta_1 \leq \Pi_n < \theta_2, X_{\alpha\Delta} = 0 \mid \xi > (n-1)\Delta, Y_1, \ldots, Y_{(n-1)}, \Pi_{(n-1)}) \]
\[ = P\left\{ 2\ln\left(\frac{(1-\theta_2)c^1_{\Pi_{(n+1)}}}{\theta_2c^0_{\Pi_{(n+1)}}} \left(\|\Sigma_0\| \cdot \|\Sigma_1\|^{-1}\right)^{1/2}\right) - C < V_n \leq 2\ln\left(\frac{(1-\theta_2)c^1_{\Pi_{(n+1)}}}{\theta_2c^0_{\Pi_{(n+1)}}} \left(\|\Sigma_0\| \cdot \|\Sigma_1\|^{-1}\right)^{1/2}\right) - C \mid X_{\alpha\Delta} = 0 \right\} \]  
(4.21)

\[ p_{t, L+2} = P(\theta_1 \leq \Pi_n < \theta_2, X_{\alpha\Delta} = 1 \mid \xi > (n-1)\Delta, Y_1, \ldots, Y_{(n-1)}, \Pi_{(n-1)}) \]
\[ = P\left\{ 2\ln\left(\frac{(1-\theta_2)c^1_{\Pi_{(n+1)}}}{\theta_2c^0_{\Pi_{(n+1)}}} \left(\|\Sigma_0\| \cdot \|\Sigma_1\|^{-1}\right)^{1/2}\right) - C < V_n \leq 2\ln\left(\frac{(1-\theta_2)c^1_{\Pi_{(n+1)}}}{\theta_2c^0_{\Pi_{(n+1)}}} \left(\|\Sigma_0\| \cdot \|\Sigma_1\|^{-1}\right)^{1/2}\right) - C \mid X_{\alpha\Delta} = 1 \right\} \]  
(4.22)

where \( \theta_1 = \Pi, \theta_2 = 1 \) and \( V_n \) is defined in equation (4.14).

Finally, to impose the maintenance actions implied by the Bayesian control chart, we have
\[ p_{L+1, 1} = p_{L+2, 1} = p_{L+3, 1} = 1 \]  
(4.23)

Using the conditional reliability function in equation (4.17), the mean sojourn time for state \( l \in K \) is derived as
\[ \tau_l = \int_0^\Delta R(t \mid \Pi_{(n-1)}) dt \]
\[ = \frac{1}{\beta} (1-e^{-\beta \Delta}) - (1-\Pi_{(n-1)}) \cdot \frac{\lambda_{10} + \lambda_{12} + \lambda_{01} - \alpha}{\beta - \alpha} \cdot \left( \frac{e^{-\alpha \Delta}}{\alpha} - \frac{1}{\alpha} - \frac{e^{-\beta \Delta}}{\beta} + \frac{1}{\beta} \right) \]
\[ - \Pi_{(n-1)} \cdot \frac{\lambda_{01} + \lambda_{02} + \lambda_{10} - \alpha}{\beta - \alpha} \cdot \left( \frac{e^{-\alpha \Delta}}{\alpha} - \frac{1}{\alpha} - \frac{e^{-\beta \Delta}}{\beta} + \frac{1}{\beta} \right) \]  
(4.24)
where $\alpha, \beta$ are defined in equation (4.3) and $\Pi_{(n-1)}$ is approximated by $\frac{l-0.5}{N} \Pi$.

The remaining mean sojourn times are given by

$$\tau_{L+1} = T_0, \tau_{L+2} = T_1, \tau_{L+3} = T_2$$

(4.25)

For the availability maximization problem, the mean ‘costs’ of the SMDP are simply the system uptimes given by

$$c_i = \tau_i, \quad l \in K_1$$

$$c_{L+1} = c_{L+2} = c_{L+3} = 0$$

(4.26)

In many real CBM applications, additional statistical constraints are typically imposed to improve the performance of the control chart. We consider the constraint that the probability of a true alarm (upon system inspection the system is found to be in the unhealthy state 1) is greater than $\gamma \in (0,1)$. Mathematically, this statistical constraint is expressed as

$$P(\bar{T} < \xi, X_{\tau} = 1) \geq \gamma$$

(4.27)

We now show how the SMDP transition probabilities given by equations (4.18) - (4.23) can be used to evaluate this statistical constraint. The probability $P(\bar{T} < \xi, X_{\tau} = 1)$ corresponds precisely to the case that the SMDP starts in state 1 and enters state $L+2$ before it enters other states $L+1$ or $L+3$. To compute this probability, we denote $r_{i,L+2}$ as the probability that the SMDP starts in state $l$ and ends in state $L+2$.

Then for $l \in K_1$, $r_{i,L+2}$ satisfies the following system of linear equations

$$r_{i,L+2} = \sum_{m \in K_1} p_{lm} \cdot r_{m,L+2} + p_{l,L+2}$$

(4.28)

Let $\mathcal{S}_\gamma = \{(\Pi, \Delta) : r_{i,L+2} \geq \gamma\}$ be the set of decision parameters that satisfy the statistical constraint. Thus, the maximum average availability $g(\Pi', \Delta') = \sup_{(\Pi, \Delta) \in \mathcal{S}_\gamma} g(\Pi, \Delta)$ subject to the statistical
constraint can be computed as follows. For a given combination of the control limit $\Pi$ and sampling interval $\Delta$, we solve for $r_{i,L+2}$ using the system of linear equations (4.28). If the statistical constraint $r_{i,L+2} \geq \gamma$ is satisfied, i.e. $(\Pi,\Delta) \in S_\gamma$, then the long-run average availability $g(\Pi,\Delta)$ is computed as the solution of the linear equations (4.8) and (4.9). The maximum average availability $g(\Pi^*,\Delta^*)$ can be obtained by solving the system of equations (4.8) and (4.9) over different combinations of $\Pi$ and $\Delta$ satisfying $(\Pi,\Delta) \in S_\gamma$. In the next section, we illustrate this procedure with a numerical example.

4.1.4 Numerical example

4.1.4.1 Simulated data

In this numerical example, we assume that the model parameters are already known and proceed with the optimization under this assumption. However, in practice, the model parameters need to be estimated first using the techniques in Chapter 2, before applying the Bayesian control methodologies. We assume the system deterioration follows a continuous-time homogeneous Markov chain $(X_t : t \geq 0)$, with state space $S = \{0,1,2\}$. States 0 and 1 are unobservable, representing the healthy and unhealthy operational states respectively, and state 2 corresponds to the observable failure state. The transition rates $\lambda_{ij}$ of the state process are given by

$$
\lambda_{01} = 0.15, \lambda_{02} = 0.02, \lambda_{10} = 0.01, \lambda_{12} = 0.2
$$

The residual observation process is calculated using equation (2.6) and is assumed to follow $N_2(\mu_0, \Sigma_0)$ when the system is in the healthy state and follow $N_2(\mu_1, \Sigma_1)$ when the system is in the unhealthy state where

$$
\mu_0 = \begin{pmatrix} 0.5 \\ 1 \end{pmatrix}, \Sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}
$$

$$
\mu_1 = \begin{pmatrix} 3 \\ 1.5 \end{pmatrix}, \Sigma_1 = \begin{pmatrix} 9 & 0 \\ 0 & 9 \end{pmatrix}
$$

64
The time required for maintenance actions are given by

\[ T_0 = 1, T_1 = 3, T_2 = 10 \]

We first compute the optimal sampling interval and control limit that maximize the long-run average availability without any statistical constraint. We choose the partition parameter \( L = 30 \), and use equations (4.8) and (4.9) to obtain the optimal results as shown in Table 4.1. We have coded the algorithm in MATLAB (R2008a) on an Intel Corel 2 6420, 2.13 GHz PC with 2 GB RAM.

### Table 4.1 Results for the Optimal Bayesian Control Chart

<table>
<thead>
<tr>
<th>Optimal control limit</th>
<th>Optimal sampling interval</th>
<th>Maximum average availability</th>
<th>( P ) (Cycle ends with true alarm)</th>
<th>Computation time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3481</td>
<td>0.2640</td>
<td>0.5871</td>
<td>0.7549</td>
<td>130.4</td>
</tr>
</tbody>
</table>

Table 4.1 shows that the computation algorithm takes only 130.4 seconds to compute the maximum average availability, which is extremely fast for off-line computations. The optimal values of the control limit and sampling interval are 0.3481 and 0.2640, which give a maximum average availability 0.5871. Thus, over a long-run horizon, the system is operational for 58.71% of time on average.

We now fix the sampling interval \( \Delta = 0.2640 \) and investigate the effect of \( T_0, T_1 \), and \( T_2 \) on the maximum average availability and optimal control limit. The results are shown in Table 4.2.

For all three cases \( i = 0,1,2 \), as \( T_i \) increases, it is clear that the maximum average availability will necessarily decrease. For fixed \( \Delta, T_0, T_1 \), when the system failure parameter \( T_2 \) increases, the optimal control limit \( \Pi' \) will decrease to more effectively prevent lengthy system downtime due to failure. The effect on the control limit is opposite for increasing \( T_0 \) and \( T_1 \). That is, for fixed \( \Delta, T_2 \), as \( T_0 \) or \( T_1 \) increases, the optimal control limit \( \Pi' \) will increase since it is more “costly” to frequently initiate full system inspection while the system is operational.
Table 4.2 Optimal Values of Average Availability and Control Limit with Varying $T_i$

<table>
<thead>
<tr>
<th>$T_0$</th>
<th>$T_i$</th>
<th>$T_2$</th>
<th>Maximum average availability</th>
<th>Optimal control limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>3</td>
<td>10</td>
<td>0.5893</td>
<td>0.2645</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>10</td>
<td>0.5871</td>
<td>0.3481</td>
</tr>
<tr>
<td>1.5</td>
<td>3</td>
<td>10</td>
<td>0.5847</td>
<td>0.4019</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>10</td>
<td>0.5825</td>
<td>0.4754</td>
</tr>
<tr>
<td>2.5</td>
<td>3</td>
<td>10</td>
<td>0.5799</td>
<td>0.5231</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>10</td>
<td>0.6087</td>
<td>0.3279</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>10</td>
<td>0.5871</td>
<td>0.3481</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>10</td>
<td>0.5665</td>
<td>0.3728</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>10</td>
<td>0.5432</td>
<td>0.4021</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>10</td>
<td>0.5224</td>
<td>0.4259</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>6</td>
<td>0.6727</td>
<td>0.5787</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>8</td>
<td>0.6274</td>
<td>0.4539</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>10</td>
<td>0.5871</td>
<td>0.3481</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>12</td>
<td>0.5384</td>
<td>0.2521</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>14</td>
<td>0.4933</td>
<td>0.1465</td>
</tr>
</tbody>
</table>

The illustrative plots are shown below.

![Optimal Control Limit & Maximum Average Availability versus $T_0$](image-url)

Figure 4.2 Optimal Control Limit & Maximum Average Availability versus $T_0$
Figure 4.3 Optimal Control Limit & Maximum Average Availability versus $T_1$

Figure 4.4 Optimal Control Limit & Maximum Average Availability versus $T_2$
We next compare the Bayesian control policy with the well-known age-based replacement policy which does not take condition monitoring information into account. Barlow and Hunter (1960) proved that under the age-based replacement policy, the optimal preventive replacement time \( \tau \) satisfies

\[
\int_0^\tau R(t \mid 0) dt - (1 - R(\tau \mid 0)) = \frac{T_e - T_s}{T_e - T_s}
\]

(4.29)

where \( T_s = T_1 \) and \( T_e = T_2 \), the conditional reliability function \( R(t \mid 0) \) is given by

\[
R(t \mid 0) = 1 - P_{02}(t) = e^{-\beta t} + \frac{\lambda_{10} + \lambda_{12} + \lambda_{01} - \alpha}{\beta - \alpha} (e^{-\alpha t} - e^{-\beta t})
\]

(4.30)

and the hazard rate function \( h(t) \) is given by

\[
h(t) = \frac{1}{R(t \mid 0)} \left( -\frac{dR(t \mid 0)}{dt} \right) = \frac{\beta^2 e^{-\beta t} - \alpha^2 e^{-\alpha t} - (\lambda_{10} + \lambda_{12} + \lambda_{01})(\beta e^{-\beta t} - \alpha e^{-\alpha t})}{\beta e^{-\beta t} - \alpha e^{-\alpha t} - (\lambda_{10} + \lambda_{12} + \lambda_{01})(e^{-\beta t} - e^{-\alpha t})}
\]

(4.31)

with constants \( \alpha, \beta \) defined in equation (4.3).

Solving equation (4.29) gives \( \tau = 18.9981 \). The long-run expected average availability per unit time for the age-based replacement policy is calculated as 0.5136, which is less than the Bayesian maximum average availability of 0.5871. Although this increase of 7.35% does not seem dramatic, the system cycle length is often very long in practical applications, which converts to a huge increase in actual system uptime. For example, in the mining industry, the system cycle length is typically well over 25,000 operational hours, so that any small increases in system average availability convert to significant increases in system operational time.

We now consider the effect the statistical constraint has on the maximum average availability. In Table 4.3, we compute different optimal combinations of sampling interval and control limit for different choices of the statistical constraint parameter \( \gamma \).
Table 4.3 Results for the Optimal Bayesian Control Chart with Statistical Constraint

<table>
<thead>
<tr>
<th>Statistical constraint ( \gamma )</th>
<th>Optimal control limit</th>
<th>Optimal sampling interval</th>
<th>Maximum average availability</th>
<th>( P ) (Cycle ends with true alarm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.3481</td>
<td>0.2640</td>
<td>0.5871</td>
<td>0.7549</td>
</tr>
<tr>
<td>0.1</td>
<td>0.3481</td>
<td>0.2640</td>
<td>0.5871</td>
<td>0.7549</td>
</tr>
<tr>
<td>0.2</td>
<td>0.3481</td>
<td>0.2640</td>
<td>0.5871</td>
<td>0.7549</td>
</tr>
<tr>
<td>0.3</td>
<td>0.3481</td>
<td>0.2640</td>
<td>0.5871</td>
<td>0.7549</td>
</tr>
<tr>
<td>0.4</td>
<td>0.3481</td>
<td>0.2640</td>
<td>0.5871</td>
<td>0.7549</td>
</tr>
<tr>
<td>0.5</td>
<td>0.3481</td>
<td>0.2640</td>
<td>0.5871</td>
<td>0.7549</td>
</tr>
<tr>
<td>0.6</td>
<td>0.3481</td>
<td>0.2640</td>
<td>0.5871</td>
<td>0.7549</td>
</tr>
<tr>
<td>0.7</td>
<td>0.3481</td>
<td>0.2640</td>
<td>0.5871</td>
<td>0.7549</td>
</tr>
<tr>
<td>0.8</td>
<td>0.5529</td>
<td>0.2485</td>
<td>0.5840</td>
<td>0.8000</td>
</tr>
<tr>
<td>0.85</td>
<td>0.8328</td>
<td>0.1779</td>
<td>0.5614</td>
<td>0.8500</td>
</tr>
<tr>
<td>0.9</td>
<td>0.9475</td>
<td>0.0852</td>
<td>0.5013</td>
<td>0.9000</td>
</tr>
</tbody>
</table>

Table 4.3 shows that the statistical constraint only takes effect when the constraint parameter \( \gamma > 0.7549 \). To meet the constraint for higher values of \( \gamma \), the control limit must increase and the sampling interval must decrease to reduce the likelihood of false alarm. Naturally, the maximum average availability decreases when the constraint \( \gamma \) increases. Illustrative graphs are provided below.

Figure 4.5 Maximum Average Availability versus Statistical Constraint
Figure 4.6 Optimal Control Limit versus Statistical Constraint

Figure 4.7 Optimal Sampling Interval versus Statistical Constraint
4.1.4.2 Spectrometric oil data

We next apply the Bayesian control scheme to the spectrometric oil application, which is summarized by the following process map.

![Figure 4.8 Process Map of the Bayesian Control Chart for Spectrometric Oil Application](image)

**Figure 4.8 Process Map of the Bayesian Control Chart for Spectrometric Oil Application**

In Chapter 2, we have obtained the residuals from the spectrometric oil data histories, and the estimated parameters of the residual observation process and state process are also given in Table 2.4. The system inspection and replacement time parameters are given by

\[ T_0 = 5hrs, T_1 = 23hrs, T_2 = 75hrs \]

Using Tijms’ approach, i.e. equations (4.8) and (4.9), the optimal control limit is found to be \( \Pi^* = 0.5007 \). An example of the Bayesian control chart is plotted below for a failure history, which failed at \( 600hrs \times 17.6 = 10560hrs \). The control chart suggests that the system should be stopped for inspection at the 15th sampling epoch, which can effectively prevent the occurrence of system failure.
Once the posterior probability that system is in the unhealthy state exceeds $\Pi^* = 0.5007$ on the Bayesian control chart, full system inspection is initiated to investigate system’s deterioration. We apply the Bayesian control chart to each of the 36 data histories and obtain the following failure prediction results.

**Table 4.4 Results for the Bayesian Control Policy in Spectrometric Oil Application**

<table>
<thead>
<tr>
<th></th>
<th>No. of preventive maintenance</th>
<th>No. of failure replacement</th>
<th>Average availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayesian control policy</td>
<td>31</td>
<td>5</td>
<td>0.7913</td>
</tr>
</tbody>
</table>

Table 4.4 shows that the Bayesian control chart can effectively predict 8 out of 13 failures and the average availability could achieve 0.7913. The 5 unpredicted failures are all from very short data histories, i.e. failure occurred shortly after system started. The healthy data of such histories is not sufficient enough to fit the VAR model. Therefore, the Bayesian control chart based on the calculated residuals did not predict the failure accurately.
4.2 Bayesian control chart design using ARL criterion

Besides the availability maximization criterion, the run length and average run length statistics are also traditional measures of a statistical control scheme's performance. Determining the run length distribution and its average is frequently a difficult and tedious task. In this section, we apply a simple unified method based on a finite Markov chain approach for finding the ARL of a Bayesian control chart.

4.2.1 Model formulation

Continue with the model formulation in Section 4.1 and the SMDP formulation in Section 4.1.3. For \( i, j \in K_1 \), denote

\[
p_y(0) = T_0 \left( \frac{j-1}{L}, \frac{j}{L}, \xi > n\Delta, \Pi_{(n-1)} \right)
\]

\[
p_y(1) = T_1 \left( \frac{j-1}{L}, \frac{j}{L}, \xi > n\Delta, \Pi_{(n-1)} \right)
\]

(4.32)

where \( p_y(0) \) represents the transition probability that the SMDP transits from state \( i \) to \( j \) given \( X_t = 0 \), \( p_y(1) \) represents the transition probability that the SMDP transits from state \( i \) to \( j \) given \( X_t = 1 \), \( T_0(\theta_1, \theta_2 | \xi > n\Delta, \Pi_{(n-1)}) \) and \( T_1(\theta_1, \theta_2 | \xi > n\Delta, \Pi_{(n-1)}) \) are defined in equation (4.19). Then the SMDP transition matrices when system is in state 0 and 1 are given by \( P_0 \) and \( P_1 \), where

\[
P_0 = \left( p_y(0) \right)_{i,j \in K_1}
\]

\[
P_1 = \left( p_y(1) \right)_{i,j \in K_1}
\]

(4.33)

From Fu et al. (2002), we know

\[
\text{ARL} = (I - P)^{-1} \cdot 1
\]

(4.34)
where \( P = (p_{lm})_{l,m \in K_1} \) is the \( L \times L \) matrix, and the \( L \) - dimensional vector \( \mathbf{1} \) has all ones for each element.

When the behavior of the ARL is studied, it is usually assumed that the SMDP starts in state \( 1 \in K_1 \) and the state process \( (X_t : t \geq 0) \) starts in state 0. Specially, we use \( ARL_0 \) to denote the average run length when the system is in state 0, and use \( ARL_1 \) to denote the average run length when the system is in state 1. Similarly, we have

\[
ARL_0 = (I - P_0)^{-1} \cdot \mathbf{1} \\
ARL_1 = (I - P_1)^{-1} \cdot \mathbf{1}
\] (4.35)

In the next section, we use \( ARL_0 = 200 \) as the criterion to design the Bayesian control chart for fault prediction.

### 4.2.2 Numerical example

#### 4.2.2.1 Simulated data

We assume the system deterioration follows a continuous-time homogeneous Markov chain \( (X_t : t \geq 0) \), with state space \( S = \{0, 1, 2\} \). States 0 and 1 are unobservable, representing the healthy and unhealthy operational states respectively, and state 2 corresponds to the observable failure state. The transition rates of the state process are given by

\[
\lambda_{01} = 0.1, \lambda_{02} = 0.01, \lambda_{10} = 0, \lambda_{12} = 0.4
\]

For convenience, assume the sampling interval \( \Delta = 1 \), and parameters for the residual observation process \( (Y_n : n \in \mathbb{N}) \) are given by

\[
\mu_0 = \begin{pmatrix} 0 \\ 0.5 \end{pmatrix}, \mu_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \\
\Sigma_0 = \Sigma_1 = \Sigma = \begin{pmatrix} 2 & 3 \\ 3 & 6 \end{pmatrix}
\]
Using the criterion $\text{ARL}_0 = 200$, the optimal control limit for the Bayesian control chart is calculated as $\bar{\Pi} = 0.695$. Based on the given state and residual observation parameters, we generate a bivariate residual observation process $(Y_n \in \mathbb{R}^2 : n \in \mathbb{N})$ with two types of data histories, one ended with failure and the other ended with suspension. Denote $N = 20$ as the number of failure histories and $M = 10$ as the number of suspension histories. We plot the posterior probability statistic on the Bayesian control chart for all the generated 30 histories.

We use the following rules to differentiate true alarms and false alarms on the control chart. For a failure history, if the posterior probability statistic exceeds the control limit at the end of the history, a preventive maintenance action will be triggered prior to the impending failure. This exceed is therefore regarded as a true alarm. For a suspension history, only when the posterior probability statistic exceeds the control limit at the very last sampling epoch of the history, a preventive maintenance action will be triggered to avoid future failure. This exceed is regarded as a true alarm, otherwise the exceed is treated as a false alarm. For example, the following two plots are the Bayesian control charts for a failure history and a suspension history.

![Figure 4.10 The Bayesian Control Chart for Simulated Failure History using ARL Criterion](image)
Figure 4.11 The Bayesian Control Chart for Simulated Suspension History using ARL Criterion

Figure 4.10 shows that the posterior probability statistic exceeds the control limit at the 9th sampling epoch. This exceed predicts the failure (occurred at 9.66) and is treated as a true alarm. In Figure 4.11, there is no exceed on the Bayesian control chart, which means there is no false alarm or true alarm for this history.

In the statistical process control literature, Chi-square control chart, MEWMA control chart and MCUSUM control chart are three widely used multivariate control charts. Using the same criterion $ARL_0 = 200$, we calculate the control limits for Chi-square chart (calculated using $\chi^2$ distribution), MEWMA chart (Lowry and Woodall 1992), and MCUSUM chart (Lee and Khoo 2006).

<table>
<thead>
<tr>
<th></th>
<th>Chi-square</th>
<th>MEWMA</th>
<th>MCUSUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control limit</td>
<td>10.597</td>
<td>8.660</td>
<td>7.054</td>
</tr>
</tbody>
</table>

For Chi-square chart, the control statistic $\chi_a^2$ is given by
\[ z_n^2 = (Y_n - \mu_0)'\Sigma^{-1}(Y_n - \mu_0) \]  

(4.36)

For MCUSUM chart, the control statistic \( S_n \) is given by

\[ S_n = \max \{ S_{(n-1)} + a(Y_n - \mu_0) - 0.5l, 0 \} \]  

(4.37)

where \( a = \frac{(\mu_i - \mu_0)'\Sigma^{-1}}{\sqrt{(\mu_i - \mu_0)'\Sigma^{-1}(\mu_i - \mu_0)}} \), \( l = (\mu_i - \mu_0)'\Sigma^{-1}(\mu_i - \mu_0) \) and \( S_0 = 0 \).

For MEWMA chart, the control statistic \( T_n \) is given by

\[ T_n = r^{-2}U_n'\Sigma^{-1}U_n \]  

(4.38)

where

\[ U_n = R\cdot Y_n + (I - R)\cdot U_{(n-1)} \]

\[ R = \text{diag}(r,r) \]  

(4.39)

and \( U_0 = 0 \), \( r = 0.1 \).

We next plot the control statistics given by equations (4.36) - (4.38) on the three control charts, along with the control limits given in Table 4.5. Comparisons of the fault prediction capabilities between the Bayesian control chart and these three control charts are summarized in Table 4.6.

<table>
<thead>
<tr>
<th>Table 4.6 Fault Prediction Capability Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted failures (Out of ( N = 20 ) failure histories)</td>
</tr>
<tr>
<td>---------------------------------------------------------------</td>
</tr>
<tr>
<td>Chi-square</td>
</tr>
<tr>
<td>MCUSUM</td>
</tr>
<tr>
<td>MEWMA</td>
</tr>
<tr>
<td>Bayesian</td>
</tr>
</tbody>
</table>

Table 4.6 shows that the numbers of false alarms for these four control schemes are all low and acceptable, but the Bayesian control chart can predict more failures compared with other three charts. In practice, it usually takes more time durations and expenses to complete the failure
replacement when the system has failed. Therefore, the Bayesian control scheme will have the lowest maintenance cost and highest average availability for practical fault prediction and control.

4.2.2.2 Vibration data

In this section, we consider the Bayesian control chart design for the vibration data collected from accelerometers located on the transmission housing from Syncrude Test Rig. It is functionally a motor-drive train-generator test stand. The transmission diagram of the gearbox is shown in Figure 4.12.

![Figure 4.12 Transmission Diagram of the Gearbox from Syncrude Ltd.](image)

The gearbox contains three stages and all the gears are straight-tooth gears. The number of teeth, speed, and the reduction ratio of each stage are listed in Table 4.7.

<table>
<thead>
<tr>
<th>Stage</th>
<th>Input speed (RPM)</th>
<th>Reduction ratio</th>
<th>Carrier speed (RPM)</th>
<th>No. of teeth of sun/shaft</th>
<th>No. of teeth of planet/bevel</th>
<th>No. of teeth of ring</th>
<th>Gear meshing frequency (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1&lt;sup&gt;st&lt;/sup&gt; stage</td>
<td>1200</td>
<td>4.0</td>
<td>-----</td>
<td>18</td>
<td>72</td>
<td>-----</td>
<td>360</td>
</tr>
<tr>
<td>2&lt;sup&gt;nd&lt;/sup&gt; stage</td>
<td>300.0</td>
<td>6.429</td>
<td>46.667</td>
<td>28</td>
<td>62</td>
<td>152</td>
<td>118.2222</td>
</tr>
<tr>
<td>3&lt;sup&gt;rd&lt;/sup&gt; stage</td>
<td>46.667</td>
<td>5.263</td>
<td>8.867</td>
<td>19</td>
<td>31</td>
<td>81</td>
<td>11.9700</td>
</tr>
</tbody>
</table>
The vibration data are collected in 300 seconds which covers 3,000,000 sampling points in total. The sampling frequency is 10,000 Hz for files from rtf_1st_run to rtf_3rd_run, and the sampling frequency is 5,000 Hz for files from rtf_4th_run to rtf_13th_run. In this research, we apply the discrete wavelet transform (DWT) technique to the time synchronously averaged (TSA) signals from different stages in the gearbox application. Details can be found in Yu and Makis (2011). The variance, kurtosis, RMS and crest factor signals are obtained and used as fault indicators.

<table>
<thead>
<tr>
<th></th>
<th>Variance</th>
<th>Kurtosis</th>
<th>RMS</th>
<th>Crest factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rtf_1st</td>
<td>1.0835×10^{-5}</td>
<td>707.8083</td>
<td>0.4857</td>
<td>0.0939</td>
</tr>
<tr>
<td>Rtf_2nd</td>
<td>7.3706×10^{-5}</td>
<td>3817.7</td>
<td>1.6937</td>
<td>0.2422</td>
</tr>
<tr>
<td>Rtf_3rd</td>
<td>7.8197×10^{-6}</td>
<td>11887</td>
<td>0.8386</td>
<td>0.0451</td>
</tr>
<tr>
<td>Rtf_4th</td>
<td>7.4748×10^{-5}</td>
<td>4652.4</td>
<td>0.7021</td>
<td>0.4474</td>
</tr>
<tr>
<td>Rtf_5th</td>
<td>9.0243×10^{-5}</td>
<td>2166.3</td>
<td>0.8384</td>
<td>0.3765</td>
</tr>
<tr>
<td>Rtf_6th</td>
<td>3.7882×10^{-4}</td>
<td>927.0498</td>
<td>1.6576</td>
<td>0.7159</td>
</tr>
<tr>
<td>Rtf_7th</td>
<td>3.8035×10^{-4}</td>
<td>1021.4</td>
<td>1.8446</td>
<td>0.8080</td>
</tr>
<tr>
<td>Rtf_8th</td>
<td>4.2011×10^{-4}</td>
<td>753.1540</td>
<td>1.4313</td>
<td>0.7831</td>
</tr>
<tr>
<td>Rtf_9th</td>
<td>8.7496×10^{-4}</td>
<td>758.8632</td>
<td>2.8564</td>
<td>0.2392</td>
</tr>
<tr>
<td>Rtf_10th</td>
<td>9.8800×10^{-4}</td>
<td>2051.8</td>
<td>3.6058</td>
<td>0.9373</td>
</tr>
<tr>
<td>Rtf_11th</td>
<td>12.0000×10^{-4}</td>
<td>549.9466</td>
<td>4.1973</td>
<td>0.9531</td>
</tr>
<tr>
<td>Rtf_12th</td>
<td>18.0000×10^{-4}</td>
<td>1911.2</td>
<td>4.9814</td>
<td>1.3510</td>
</tr>
<tr>
<td>Rtf_13th</td>
<td>83.0000×10^{-4}</td>
<td>4525.1</td>
<td>10.7413</td>
<td>3.7663</td>
</tr>
</tbody>
</table>

We graph below these four signals (variance, kurtosis, RMS and crest factor).
Figure 4.13 Variance Signal after DWT Applied to TSA Signals

Figure 4.14 Kurtosis Signal after DWT Applied to TSA Signals
Figure 4.15 RMS Signal after DWT Applied to TSA Signals

Figure 4.16 Crest Factor Signal after DWT Applied to TSA Signals
Figures 4.13 - 4.16 and past experiences suggest that system is in the run-in stage from sampling epoch 1 to 3. Take the variance and RMS signals for example, the values have a steady trend from sampling epoch 4 to 8, and there is a sudden increase at sampling epoch 9, indicating the occurrence of deterioration. After that there is a dramatic increase at sampling epoch 13, which is caused by failure. We next use the variance and RMS signals to design a Bayesian fault prediction and control scheme using the ARL criterion.

We first test the independence and normality of the bivariate signal (variance and RMS). The p-values for the Portmanteau Independence Test and the Henze-Zirkler Multivariate Normality Test are calculated below.

| Table 4.9 p-values of the Independence and Normality Tests for Variance and RMS Signals |
|---------------------------------|---------------------------|---------------------------|
|                                | Healthy data set         | Unhealthy data set        |
|                                | (From sampling epoch 4 to 8) | (From sampling epoch 9 to 12) |
| Independence (Portmanteau)    | 0.6350                   | 0.5338                    |
| Normality (Henze-Zirkler)     | 0.3325                   | 0.3469                    |

Table 4.9 shows that there is no statistical evidence to reject the null hypotheses that the bivariate signals are independent and have multivariate normal distribution.

We assume the deterioration of the gearbox system follows a continuous-time homogeneous Markov chain \( (X_t : t \geq 0) \), with state space \( S = \{0,1,2\} \). States 0 and 1 are unobservable, representing the healthy and unhealthy operational states respectively, and state 2 corresponds to the observable failure state. While the system is operational, the bivariate observation (variance and RMS) has a state-dependent multivariate normal distribution. The sampling interval is \( \Delta = 5\text{min} \).

Using the EM algorithm, the system observation parameters are estimated as

\[
\begin{align*}
\mu_0 &= \begin{pmatrix} 2.6979 \times 10^{-4} \\ 1.2995 \end{pmatrix}, \\
\Sigma_0 &= \begin{pmatrix} 2.9392 \times 10^{-8} & 7.9769 \times 10^{-5} \\ 7.9769 \times 10^{-5} & 0.25244 \end{pmatrix}, \\
\mu_1 &= \begin{pmatrix} 12.1730 \times 10^{-4} \\ 3.9166 \end{pmatrix}, \\
\Sigma_1 &= \begin{pmatrix} 16.9310 \times 10^{-8} & 3.4930 \times 10^{-4} \\ 3.4930 \times 10^{-4} & 0.81020 \end{pmatrix}
\end{align*}
\]

and the system state parameters are estimated as
\[ \lambda_{01} = 0.036264, \lambda_{12} = 0.043687, \lambda_{02} = 0 \]

Details can be found in Jiang, Yu and Makis (2011).

Considering the criterion \( ARL_0 = 200 \), the Bayesian control limit is calculated as

\[ \Pi = 0.7403 \]

The Bayesian control chart for the bivariate observation (variance and RMS) is plotted below.

Figure 4.17 shows that the posterior probability statistic exceeds the control limit \( \Pi = 0.7403 \) at the 10th sampling epoch, indicating that the gearbox system should be stopped for full inspection immediately. This would effectively prevent the impending failure (occurred after the 12th sampling epoch).
4.3 Residual life estimation

4.3.1 Problem description and derivation of estimation formulae

In this section, we are interested in deriving explicit formulae of several important quantities for system residual life estimation, such as the conditional reliability function, the hazard rate function and the mean residual life function. We will show that each of these quantities can be expressed in terms of the posterior probability statistic. Furthermore, we will show that the mean residual life function can be computed via a recursive procedure, which makes our approach computationally attractive for practical applications.

We assume the system deterioration follows a continuous-time homogeneous Markov chain $(X_t : t \geq 0)$, with state space $S = \{0,1,2\}$. States 0 and 1 are unobservable, representing the healthy and unhealthy operational states respectively, and state 2 corresponds to the observable failure state. While the system is operational, the residual observation process $(Y_n : n \in \mathbb{N})$, which is obtained from equation (2.6) through condition monitoring, has a state-dependent multivariate normal distribution.

It follows that the conditional distribution function, density function, hazard rate function and mean residual life function are all functions of the posterior probability statistic $\Pi_n$, which is defined in equation (3.10). In particular, for any $t \geq 0$, the conditional distribution function is given by

$$F(t \mid \Pi_n) = P(\xi \leq n\Delta + t \mid \xi > n\Delta, Y_1, \ldots, Y_n)$$
$$= 1 - R(t \mid \Pi_n)$$ (4.40)

where the condition reliability function is provided in equation (4.17) as

$$R(t \mid \Pi_n) = P(\xi > n\Delta + t \mid \xi > n\Delta, Y_1, \ldots, Y_n, \Pi_n)$$
$$= (1 - \Pi_n)(1 - P_{02}(t)) + \Pi_n (1 - P_{12}(t))$$
$$= 1 - P_{02}(t) + \Pi_n (P_{02}(t) - P_{12}(t))$$
$$= \frac{(\lambda_{01} + \lambda_{02})e^{-(\lambda_{01} + \lambda_{02})t} - \lambda_{12}e^{-\lambda_{12}t} + (\lambda_{12} + \lambda_{01} + \Pi_n (\lambda_{02} - \lambda_{12}))(e^{-\lambda_{12}t} - e^{-(\lambda_{01} + \lambda_{02})t})}{\lambda_{01} + \lambda_{02} - \lambda_{12}}$$
The density function is given by

\[
f(t \mid \Pi_n) = \frac{dF(t \mid \Pi_n)}{dt} = \left(\lambda_{01} + \lambda_{02}\right)^2 e^{-(\lambda_{01} + \lambda_{02})t} - \lambda_{12}^2 e^{-\lambda_{12}t} - \left(\lambda_{12} + \lambda_{01} + \Pi_n(\lambda_{02} - \lambda_{12})\right)\left(\lambda_{01} + \lambda_{02}\right)e^{-(\lambda_{01} + \lambda_{02})t} - \lambda_{12}e^{-\lambda_{12}t} \over \lambda_{01} + \lambda_{02} - \lambda_{12}\right)
\]

(4.41)

and the hazard rate function is given by

\[
h(t \mid \Pi_n) = \frac{f(t \mid \Pi_n)}{R(t \mid \Pi_n)} = \left(\lambda_{01} + \lambda_{02}\right)^2 e^{-(\lambda_{01} + \lambda_{02})t} - \lambda_{12}^2 e^{-\lambda_{12}t} - \left(\lambda_{12} + \lambda_{01} + \Pi_n(\lambda_{02} - \lambda_{12})\right)\left(\lambda_{01} + \lambda_{02}\right)e^{-(\lambda_{01} + \lambda_{02})t} - \lambda_{12}e^{-\lambda_{12}t} \over \left(\lambda_{01} + \lambda_{02}\right)e^{-(\lambda_{01} + \lambda_{02})t} - \lambda_{12}e^{-\lambda_{12}t} + \left(\lambda_{12} + \lambda_{01} + \Pi_n(\lambda_{02} - \lambda_{12})\right)(e^{-\lambda_{12}t} - e^{-(\lambda_{01} + \lambda_{02})t})
\]

(4.42)

Then, the mean residual life function is given by

\[
\mu_{n\Delta} = E(\xi - n\Delta \mid \xi > n\Delta,Y_1,...,Y_n,\Pi_n) = \int_0^\infty R(t \mid \Pi_n)dt
\]

(4.43)

\[
= \frac{\lambda_{12} + \lambda_{01} + \Pi_n(\lambda_{02} - \lambda_{12})}{\lambda_{12}(\lambda_{01} + \lambda_{02})}
\]

It is interesting to note that equations (4.11) and (4.43) imply that the mean residual life function \( \mu_{n\Delta} \) can be computed recursively using its previous value \( \mu_{(n-1)\Delta} \) as

\[
\mu_{n\Delta} = \frac{1}{\lambda_{12}} + \frac{(\lambda_{02} - \lambda_{12})}{(\lambda_{01} + \lambda_{12})\lambda_{12}} + fratio \cdot P_{00}(\Delta)\left(\lambda_{01} + \lambda_{02}\right)^{-\lambda_{12}}(\lambda_{01} + \lambda_{02})^{-\mu_{(n-1)\Delta}}\right) - P_{01}(\Delta)\left(\lambda_{01} + \lambda_{12}\right) - \lambda_{12}(\lambda_{01} + \lambda_{02})^{-\mu_{(n-1)\Delta}}\right)
\]

(4.44)

where \( fratio = \frac{f_{Y|X}(y \mid 0)}{f_{Y|X}(y \mid 1)} \).
Although the MRL statistic describes the average remaining life of the system, it is also important for practical applications to consider the variance of the system residual life. Using the probability density function in equation (4.41), the variance of the system residual life is given by

\[
\sigma^2_{n\Delta} = \text{var}(\xi - n\Delta | \xi > n\Delta, Y_1, ..., Y_n, \Pi_n) \\
= E\left((\xi - n\Delta)^2 | \xi > n\Delta, Y_1, ..., Y_n, \Pi_n\right) - \mu^2_{n\Delta} \\
= \int_0^\infty t^2 f(t | \Pi_n) dt - \mu^2_{n\Delta}
\]

\[\sigma^2_{n\Delta} = \left(\lambda_{12} + \lambda_{01} + \Pi_n(\lambda_{02} - \lambda_{12})\right)\left(\lambda_{12} + \lambda_{01} + 2\lambda_{02} - \Pi_n(\lambda_{02} - \lambda_{12})\right) - 2\lambda_{12}(\lambda_{01} + \lambda_{02})
\]

\[\lambda_{12}^2(\lambda_{01} + \lambda_{02})^2
\]

In the next section, we present a numerical example to illustrate the parameter estimation and residual life prediction procedure.

### 4.3.2 Numerical example

We assume the system deterioration follows a continuous-time homogeneous Markov chain \(X_t: t \geq 0\), with state space \(S = \{0, 1, 2\}\). States 0 and 1 are unobservable, representing the healthy and unhealthy operational states respectively, and state 2 corresponds to the observable failure state. The transition rate matrix of the state process is given by

\[
\Lambda = \begin{pmatrix}
-0.62 & 0.6 & 0.02 \\
0 & -2 & 2 \\
0 & 0 & 0
\end{pmatrix}
\]

Assume the sampling interval \(\Delta = 1\) and the residual observation vectors \(Y_1, Y_2, \ldots\), follow \(N_2(\mu_0, \Sigma_0)\) when \(X_n = 0\), and follow \(N_2(\mu_1, \Sigma_1)\) when \(X_n = 1\), where
\[ \mu_0 = \begin{pmatrix} 1 \\ 1.5 \end{pmatrix}, \Sigma_0 = \begin{pmatrix} 1 & 1.5 \\ 1.5 & 4 \end{pmatrix} \]

\[ \mu_i = \begin{pmatrix} 6 \\ 8 \end{pmatrix}, \Sigma_i = \begin{pmatrix} 1.5 & 3 \\ 3 & 8 \end{pmatrix} \]

We first simulate \( N = 30 \) failure histories and \( M = 10 \) suspension histories using the above true model parameters. Using the EM algorithm estimates given by equations (2.29) and (2.32), we obtain the optimal estimates for the state and residual observation parameters.

**Table 4.10 Parameter Estimates using the EM Algorithm**

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Initial values</th>
<th>Final estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\lambda}_{\nu_1} )</td>
<td>0.57</td>
<td>0.59</td>
</tr>
<tr>
<td>( \hat{\lambda}_{\nu_2} )</td>
<td>0.04</td>
<td>0.03</td>
</tr>
<tr>
<td>( \hat{\lambda}_{1,2} )</td>
<td>2.2</td>
<td>2.05</td>
</tr>
<tr>
<td>( \hat{\mu}_0 )</td>
<td>( \begin{pmatrix} 0 \ 2 \end{pmatrix} )</td>
<td>( \begin{pmatrix} 0.92 \ 1.37 \end{pmatrix} )</td>
</tr>
<tr>
<td>( \hat{\mu}_1 )</td>
<td>( \begin{pmatrix} 5 \ 10 \end{pmatrix} )</td>
<td>( \begin{pmatrix} 6.25 \ 7.86 \end{pmatrix} )</td>
</tr>
<tr>
<td>( \hat{\Sigma}_0 )</td>
<td>( \begin{pmatrix} 1 &amp; 2 \ 2 &amp; 5 \end{pmatrix} )</td>
<td>( \begin{pmatrix} 0.87 &amp; 1.38 \ 1.38 &amp; 3.51 \end{pmatrix} )</td>
</tr>
<tr>
<td>( \hat{\Sigma}_1 )</td>
<td>( \begin{pmatrix} 2 &amp; 2.5 \ 2.5 &amp; 7 \end{pmatrix} )</td>
<td>( \begin{pmatrix} 1.64 &amp; 3.01 \ 3.01 &amp; 8.25 \end{pmatrix} )</td>
</tr>
<tr>
<td>( Q )</td>
<td>(-1.9967\times10^5)</td>
<td>(-1.1969\times10^5)</td>
</tr>
</tbody>
</table>

Now we simulate a new group of 100 failure histories using the true model parameters. For each failure history, the MRL statistic is computed using equation (4.43) and the optimal parameter estimates in Table 4.10 at the last sampling epoch before failure. For example, Figure 4.18 plots the bivariate residual observations of one of the 100 failure histories that ended in failure at time 7.82. The solid line denotes one variable of the residual observation vector, and the dotted line denotes the other variable of the residual observation vector.
Figure 4.18 Simulated Residual Observation Process for Failure History

Figure 4.19 plots the system state process for the simulated failure history in Figure 4.18. It shows that the system started in the healthy state 0 at time 0, ran under the healthy condition till sometime between 6 and 7, ran for another short time period under the unhealthy condition, and finally failed at time 7.82.
The MRL statistic is computed at sampling time 7, and is found to be equal to 0.51. This result suggests that on average system failure would occur before the next sampling epoch. We find that 81 out of the 100 failure histories have an MRL value less than 1, when computed at the last sampling epoch before failure. This result suggests that a high percentage of failures can be predicted using the Bayesian MRL statistic. We compare this result with the MRL statistic given by a standard 2-parameter Weibull model, which does not take into account the condition monitoring data. Using the 100 failure histories, the Weibull scale and shape parameters are estimated as 2.318 and 1.154, respectively. The Weibull MRL value is equal to 2.732 for the failure history given in Figure 4.18. We find that only 34 out of the 100 failures have a Weibull MRL value less than 1, when computed at the last sampling epoch before failure. This suggests that a large percentage of failures would not be predicted using the MRL statistic given by a standard 2-parameter Weibull model as compared to the Bayesian MRL statistic which predicts 81 out of 100 failures.

We next simulate a new suspension history of length 12 as below.
We graph the conditional density function of the system residual life at the 12th sampling epoch, as shown in Figure 4.21. Such plot can provide useful insights for maintenance and reliability engineers. For example, Figure 4.21 shows that the system residual life has a small variance and most of the probability mass is concentrated between 0 and 5 time units. The MRL statistic is calculated as 2.10, and the mode of the density function is equal to 0.8, which suggests that immediate preventive maintenance should be carefully considered since there is a high probability that the system will fail before the next sampling epoch.
Figure 4.21 The Probability Density Function of Residual Life for Suspension History
In this research, we have presented a new framework for predicting failures of a partially observable deteriorating system, using Bayesian estimation and control methodologies. A time series model is fitted to a vector observation process representing partial information about the system state. Residuals are then calculated using the fitted model, which are indicative of system deterioration. The system deterioration process is modeled as a 3-state continuous-time homogeneous Markov chain with two unobservable operational states and one observable failure state. The form of the optimal control policy that maximizes the long-run expected average availability per unit time has been investigated, and it has been proved that a control limit policy is optimal for decision making. The model parameters have been estimated using the EM algorithm. The optimal Bayesian fault prediction and control scheme, considering long-run average availability maximization along with a practical statistical constraint, has been proposed and compared with the age-based replacement policy. The optimal control limit and sampling interval are calculated in the SMDP framework. Another Bayesian fault prediction and control scheme has been developed based on the ARL criterion. Comparisons with traditional control charts are provided. Formulae for the mean residual life and the distribution function of system residual life have been derived in explicit forms as functions of the posterior probability statistic. The advantage of the Bayesian model over the well-known 2-parameter Weibull model in system residual life prediction is shown. The methodologies are illustrated using simulated data, real data obtained from the spectrometric analysis of oil samples collected from transmission units of heavy hauler trucks in the mining industry, and vibration data from a planetary gearbox machinery application. The purpose of this chapter is to summarize the conclusions of this research and to propose future possible research.

5.1 Conclusions

Results and conclusions are summarized by chapters.
In Chapter 2, an autoregressive model was fitted to a vector observation process representing partial information about the system unobservable state. Residuals were calculated using the fitted model, which are indicative of system deterioration. The residuals of the fitted model were chosen as the observation process in the HMM framework. The state process was modeled as a 3-state continuous-time homogeneous Markov chain with state space \{0,1,2\}. States 0 and 1 are not observable, representing the healthy and unhealthy operational system conditions, respectively. Only the failure state 2 is assumed to be observable. While the system is operational, the residual observation process has a state-dependent multivariate normal distribution. The maximum likelihood estimates of the model parameters were obtained using the EM algorithm. It was shown that the update formulae for estimates have explicit forms and the procedure is computationally efficient and converges rapidly to reasonably close parameter estimates. We also extended the 3-state HMM to a simple HSMM. The optimal estimates for model parameters were also explicitly obtained by using the EM algorithm.

In Chapter 3, the form of the optimal control policy, which maximizes the long-run expected average availability per unit time for partially observable deteriorating system subject to random failure, was investigated. We considered the system deterioration process and residual observation process discussed in Chapter 2. The value function and dynamic optimality equation were established. Important properties of the value function were derived. We analyzed the dynamic optimality equation to show that a simple control limit policy is optimal for the long-run average availability maximization objective. A computational algorithm for calculating the optimal control limit and maximum average availability was presented.

In Chapter 4, the multivariate Bayesian control chart was introduced and formulated for the objective of long-run average availability maximization. The posterior probability process was described and its stochastic evolution was characterized. A computational algorithm was developed in the SMDP framework to compute the optimal control limit and sampling interval, which could achieve the maximum long-run average availability. The incorporation of a practical statistical constraint that bounds the probability of true alarms was also considered, and effect of the statistical constraint on the average availability objective function was discussed. We compared the Bayesian control policy with the well-known age-based replacement policy and showed the advantage of the Bayesian control policy in long-run average availability
maximization. Another Bayesian control scheme was also designed based on the ARL criterion. Comparisons with traditional control charts, such as Chi-square chart, MEWMA chart and MCUSUM chart, were provided. After that, several important quantities for system residual life estimation, such as the distribution function, reliability function, probability density function and hazard rate function of the system residual life, were derived. The system’s mean residual life function was also computed. We used these quantities to illustrate the advantage of the Bayesian model over the well-known 2-parameter Weibull model, in which no condition monitoring data is taken into account in system residual life prediction. The methodologies were illustrated using simulated data, real data obtained from the spectrometric analysis of oil samples collected from transmission units of heavy hauler trucks in the mining industry, and vibration data from a planetary gearbox machinery application.

5.2 Future research

We suggest some possible directions for future research as follows.

- The 3-state hidden Markov model considered in this research can be extended to a more general $N$-state hidden Markov model. If more states are considered in the model building, it will approximate the real situation more closely. Meanwhile, such an extension would lead to interesting theoretical and practical challenges for both the estimation and control procedures.

- A simple hidden semi-Markov model has been considered in this research as an extension to the 3-state HMM. A phase-type distribution is used to describe the sojourn time when system is in the healthy state. Typically, it is more difficult to analyze the HSMM due to the loss of the exponential memoryless property, even though the HSMM has more precise practical representations compared with the HMM. More realistic distributions, such as Weibull or Gamma distributions, could be considered to describe the state sojourn time distribution for the HSMM.

- In this research, the optimal control policy has been investigated for partially observable deteriorating systems subject to random failure. It has been proved that the control limit
policy is optimal for decision making for the long-run average availability maximization objective. However, no such works exist in the literature about determining the optimal control policy for finite horizon availability maximization problems, which is worth a thorough study in future works.

- Another possible future research topic would be to test the effectiveness of our Bayesian estimation and control methodologies on some other real-world data sets such as vibration data, performance or quality monitoring data using the cost minimization criterion, which is sometimes more preferable in practice compared to the availability maximization criterion. This should also lead to a further refinement of both the model and the estimation and control algorithms.
Bibliography


98


Appendix 1

Proof of Lemma 2.3

Let $S_0 = X_{\tau_0}$ be the system state at time $\tau_0$. For all $t \geq 0$,

$$P(\xi \leq t) = p_{01}^t P(\xi \leq t \mid S_0 = 1) + p_{02}^t P(\xi \leq t \mid S_0 = 2)$$

$$= p_{01}^t \int_{u=0}^t \left(1 - e^{-\nu_1(t-u)}\right) \cdot g_0(u)du + p_{02}^t \int_{u=0}^t g_0(u)du$$

$$= p_{01}^t \int_{u=0}^t \left(1 - e^{-\nu_1(t-u)}\right) \cdot v_0^2ue^{-v_0u}du + p_{02}^t \int_{u=0}^t v_0^2ue^{-v_0u}du$$

$$= 1 - e^{-\nu_1t} - e^{-\nu_1v_0t} - \frac{p_{01}^t \nu_0^2 t}{\nu_0 - \nu_1} e^{-v_0t} - \frac{p_{02}^t \nu_0^2}{(\nu_0 - \nu_1)^2} (e^{-v_0t} - e^{-\nu_1t})$$

which is differentiable in $t$, so that the density function of $\xi$ is given by

$$f_\xi(t) := \frac{dP(\xi \leq t)}{dt}$$

$$= tv_0^2 e^{-v_0t} + \frac{p_{01}^t \nu_0^2}{\nu_0 - \nu_1} (e^{-v_0t} - v_0te^{-\nu_1t}) - \frac{p_{02}^t \nu_0^2}{(\nu_0 - \nu_1)^2} (v_0e^{-v_0t} - \nu_1e^{-\nu_1t})$$

for all $t \geq 0$, and zero otherwise.

For all non-negative $s < t$,

$$P(\tau_0 \leq s, \xi \leq t) = p_{01}^s P(\tau_0 \leq s, \xi \leq t \mid S_0 = 1) + p_{02}^s P(\tau_0 \leq s, \xi \leq t \mid S_0 = 2)$$

$$= p_{01}^s \int_{u=0}^s \left(1 - e^{-\nu_1(t-u)}\right) \cdot g_0(u)du + p_{02}^s \int_{u=0}^s g_0(u)du$$

$$= p_{01}^s \int_{u=0}^s \left(1 - e^{-\nu_1(t-u)}\right) \cdot v_0^2ue^{-v_0u}du + p_{02}^s \int_{u=0}^s v_0^2ue^{-v_0u}du$$

$$= 1 - e^{-\nu_1t} - e^{-\nu_1v_0t} - \frac{p_{01}^s \nu_0^2 t}{\nu_0 - \nu_1} e^{-v_0t} - \frac{p_{02}^s \nu_0^2}{(\nu_0 - \nu_1)^2} (e^{-v_0t} - e^{-\nu_1t})$$

for all $s \geq 0$. 

103
\[
\begin{align*}
&= p_{01} \int_{u=0}^{s} (1-e^{-\nu_1(t-u)}) \cdot g_0(u)du + p_{02} \int_{u=0}^{s} g_0(u)du \\
&= p_{01} \int_{u=0}^{s} (1-e^{-\nu_1(t-u)}) \cdot v_0^2 u e^{-\nu_1u} du + p_{02} \int_{u=0}^{s} v_0^2 u e^{-\nu_1u} du \\
&= 1 - e^{-\nu_0^2} - e^{-\nu_0^2} v_0 s - \frac{p_{01} v_0 s}{\nu_1 - \nu_0} e^{(\nu_0 - \nu_1)s} - \frac{p_{02} v_0^2 e^{-\nu_0 t}}{(\nu_1 - \nu_0)^2} (1 - e^{(\nu_0 - \nu_1)s}) \\
\end{align*}
\]

which is differentiable in both variables \(s, t\), so that the joint density function of \(\tau_0\) and \(\xi\) for all non-negative \(s < t\) is given by

\[
f_{\tau_0 \xi}(s,t) := \frac{\partial^2 P(\tau_0 \leq s, \xi \leq t)}{\partial s \partial t} = p_{01} v_0^2 v_1 s e^{-\nu_1 t} e^{-\nu_0 x}
\]

and for all non-negative \(s < t\), the density function is

\[
f_{\tau_0 \xi}(s \mid t) := \frac{f_{\tau_0 \xi}(s,t)}{f_\xi(t)} = \frac{p_{01} v_1 s e^{-\nu_1 t} e^{-\nu_0 x}}{te^{-\nu_1 d} + \frac{p_{01}}{(\nu_0 - \nu_1)^2} (v_1 e^{-\nu_1 t} - v_0 e^{-\nu_0 t} + v_0 v_1 t e^{-\nu_1 d} - v_0^2 t e^{-\nu_0 t})}
\]

For \(s = t\), the mass function is

\[
m_{\tau_0 \xi}(t \mid t) := P(\tau_0 = t \mid \xi = t)
\]

\[
= 1 - P(\tau_0 < t \mid \xi = t)
\]

\[
= 1 - \int_{s=0}^{t} f_{\tau_0 \xi}(s \mid t)ds
\]

\[
= \frac{tp_{02} e^{-\nu_0 t}}{te^{-\nu_1 d} + \frac{p_{01}}{(\nu_0 - \nu_1)^2} (v_1 e^{-\nu_1 t} - v_0 e^{-\nu_0 t} + v_0 v_1 t e^{-\nu_1 d} - v_0^2 t e^{-\nu_0 t})}
\]

which completes the proof. \(\square\)
Appendix 2

Proof of Theorem 2.3

Using equations (2.33) – (2.35) of Lemma 2.3 and the formula for the likelihood function $L_F(\lambda, \theta)$ given by equation (2.36), we have

\[
Q_F(\lambda, \theta | \hat{\lambda}, \hat{\theta}) = E_{\hat{\lambda}, \hat{\theta}}\left( \ln L_F(\lambda, \theta) \right) = E_{\hat{\lambda}, \hat{\theta}}\left( \ln L_F(\lambda, \theta) | \bar{Y} = \bar{Y}, \xi = t \right) = \int E_{\hat{\lambda}, \hat{\theta}}\left( \ln L_F(\lambda, \theta) | \bar{Y} = \bar{Y}, \xi = t, \tau_0 = s \right) \hat{h}_{\tau_0 \bar{Y}, \xi}(s | \bar{Y}, t) ds
\]

\[
= \int \left[ \ln \left( \tilde{g}_{\bar{y}, r_0} (\bar{y} | t, s) f_{t \xi}(s | t) f_{\xi}(t) \right) \tilde{g}_{\bar{y}, r_0} (\bar{y} | t, s) f_{t \xi}(s | t) f_{\xi}(t) ds
\]

\[
+ \ln \left( \tilde{g}_{\bar{y}, r_0} (\bar{y} | t, t) m_{t \xi}(t | t) f_{\xi}(t) \right) \tilde{g}_{\bar{y}, r_0} (\bar{y} | t, t) m_{t \xi}(t | t) f_{\xi}(t) ds \right) \right] ds
\]

\[
= \int \tilde{g}_{\bar{y}, r_0} (\bar{y} | t, u) f_{t \xi}(u | t) f_{\xi}(t) du + \hat{g}_{\bar{y}, r_0} (\bar{y} | t, t) m_{t \xi}(t | t) f_{\xi}(t)
\]

\[
= \int \tilde{g}_{\bar{y}, r_0} (\bar{y} | t, u) f_{t \xi}(u | t) f_{\xi}(t) du + \hat{g}_{\bar{y}, r_0} (\bar{y} | t, t) m_{t \xi}(t | t) f_{\xi}(t)
\]

where the notations $\tilde{g}_{\bar{y}, r_0}$, $f_{t \xi}$ and $m_{t \xi}$ are used to signify that the functions $g_{\bar{y}, r_0}$, $f_{t \xi}$ and $m_{t \xi}$ are parameterized by fixed estimates $\hat{\lambda}$ and $\hat{\theta}$. Since $g_{\bar{y}, r_0}$ defined in equations (2.16) and (2.17) depends only on the residual observation parameter $\theta = (\mu_0, \mu_1, \Sigma_0, \Sigma_1)$, and $f_{t \xi}$, $f_{t \xi}$ and $m_{t \xi}$ depend only on the state parameter $\lambda = (\lambda_0, \lambda_{12})$, the equation above can be decomposed into two terms,

\[
Q_F(\lambda, \theta | \hat{\lambda}, \hat{\theta}) = Q_F^{stat}(\lambda | \hat{\lambda}, \hat{\theta}) + Q_F^{obs}(\theta | \hat{\lambda}, \hat{\theta})
\]

105
Substituting equations (2.33) – (2.35) of Lemma 2.3, the first term $Q^\text{state}_F$ simplifies to

$$
Q^\text{state}_F(\lambda | \hat{\lambda}, \hat{\theta}) = \frac{\int_{s < t} \ln \left( f_{\tau_s \theta} (s \mid t) f_{\xi_s} (t) \right) \hat{g}_{\xi_s, r_0} (\tilde{y} \mid t, s) \hat{g}_{\tau_s \theta} (s \mid t) \, ds}{\int_{u < t} \hat{g}_{\xi_s, r_0} (\tilde{y} \mid t, u) f_{\tau_s \theta} (u \mid t) \, du + \hat{g}_{\xi_s, r_0} (\tilde{y} \mid t, t) \hat{m}_{\tau_s \theta} (t \mid t)}
$$

where constants that depend only on fixed parameter estimates $\hat{\lambda}$ and $\hat{\theta}$ are given by

$$
\begin{align*}
\hat{a}_{01} &= \hat{a}_{02} = -\hat{p}_{01} \hat{v}_1 e^{-\hat{v}_1 t} \langle \hat{e}_3, \hat{g} \rangle + t^2 \hat{p}_{02} e^{-\hat{v}_2 t} \hat{g}_{\xi_s, r_0} (\tilde{y} \mid t, t) \\
\hat{a}_{12} &= \hat{a}_{12} = \hat{p}_{01} \hat{v}_1 e^{-\hat{v}_1 t} \langle \hat{e}_3, \hat{g} \rangle \\
\hat{b}_{01} &= \hat{b}_{12} = \hat{p}_{01} \hat{v}_1 e^{-\hat{v}_1 t} \langle \hat{e}_3, \hat{g} \rangle \\
\hat{b}_{02} &= \hat{b}_{02} = \hat{p}_{02} e^{-\hat{v}_2 t} \hat{g}_{\xi_s, r_0} (\tilde{y} \mid t, t) \\
\hat{d} &= \hat{d} = \hat{p}_{01} \hat{v}_1 e^{-\hat{v}_1 t} \langle \hat{e}_3, \hat{g} \rangle + \hat{p}_{02} e^{-\hat{v}_2 t} \hat{g}_{\xi_s, r_0} (\tilde{y} \mid t, t) \\
\hat{d}_1 &= \hat{d}_1 = \hat{p}_{01} \hat{v}_1 e^{-\hat{v}_1 t} \int_{s < t} \left( s e^{(\hat{\theta}_0 - \hat{\theta}_1) s} \ln s \right) \hat{g}_{\xi_s, r_0} (\tilde{y} \mid t, s) \, ds + \hat{p}_{02} e^{-\hat{v}_2 t} \ln t \hat{g}_{\xi_s, r_0} (\tilde{y} \mid t, t)
\end{align*}
$$

and vectors $\hat{e}_1 = (\hat{e}_1^1, \ldots, \hat{e}_1^T, \hat{e}_1^T)'$, $\hat{e}_2 = (\hat{e}_2^1, \ldots, \hat{e}_2^T, \hat{e}_2^T)'$ are defined by

$$
\begin{align*}
\hat{e}_1^k &= \int_{(k-1)\Delta}^k e^{-(\hat{\theta}_0 - \hat{\theta}_1) u} \, du = \frac{e^{-(\hat{\theta}_0 - \hat{\theta}_1) k \Delta} - e^{-(\hat{\theta}_0 - \hat{\theta}_1) (k-1) \Delta}}{\hat{v}_0 - \hat{v}_1} \\
\hat{e}_1^t &= \int_{(k-1)\Delta}^t e^{-(\hat{\theta}_0 - \hat{\theta}_1) u} \, du = \frac{e^{-(\hat{\theta}_0 - \hat{\theta}_1) \Delta} - e^{-(\hat{\theta}_0 - \hat{\theta}_1) u}}{\hat{v}_0 - \hat{v}_1} \\
\hat{e}_2^k &= \int_{(k-1)\Delta}^k u e^{-(\hat{\theta}_0 - \hat{\theta}_1) u} \, du = \frac{\hat{e}_2^k - k \Delta e^{-(\hat{\theta}_0 - \hat{\theta}_1) k \Delta} + (k-1) \Delta e^{-(\hat{\theta}_0 - \hat{\theta}_1) (k-1) \Delta}}{\hat{v}_0 - \hat{v}_1} \\
\hat{e}_2^t &= \int_{(k-1)\Delta}^t u e^{-(\hat{\theta}_0 - \hat{\theta}_1) u} \, du = \frac{\hat{e}_2^t - t e^{-(\hat{\theta}_0 - \hat{\theta}_1) u} + T \Delta e^{-(\hat{\theta}_0 - \hat{\theta}_1) t}}{\hat{v}_0 - \hat{v}_1}
\end{align*}
$$
and \( \hat{e}_3 = (\hat{e}_3^1, \ldots, \hat{e}_3^r, \hat{e}_3^r)' \) is defined by

\[
\hat{e}_3^k = \int_{(k-1)\Delta}^{k\Delta} u^2 e^{-(\hat{t}_0 - t)u} \, du = \frac{2\hat{e}_2^k - k^2 \Delta^2 e^{-(\hat{t}_0 - t_k)k\Delta} + (k-1)^2 \Delta^2 e^{-(\hat{t}_0 - t_k)(k-1)\Delta}}{\hat{t}_0 - \hat{t}_1}
\]

\[
\hat{e}_3' = \int_{\tau\Delta}^t u^2 e^{-(\tau - \hat{t})u} \, du = \frac{2\hat{e}_2' - t^2 e^{-(\tau - \hat{t})t} + t^2 \Delta^2 e^{-(\tau - \hat{t})(\tau - \hat{t})}}{\hat{t}_0 - \hat{t}_1}
\]

Similarly, the second term \( Q_{\theta}^{obs} \), which is a function only of the residual observation parameter \( \theta \), simplifies to

\[
Q_{\theta}^{obs}(\theta | \hat{\lambda}, \hat{\theta}) = \frac{\int_{s < t} \ln \left( g_{\hat{y}|\hat{\tau}, \tau_0}(\hat{y} | t, s) \hat{g}_{\hat{y}|\hat{\tau}, \tau_0}(\hat{y} | t, s) \hat{f}_{\tau|\hat{\tau}}(s | t) ds \right) + \ln \left( g_{\hat{y}|\hat{\tau}, \tau_0}(\hat{y} | t, t) \hat{g}_{\hat{y}|\hat{\tau}, \tau_0}(\hat{y} | t, t) \hat{m}_{\tau|\hat{\tau}}(t | t) \right)}{\int_{u < t} \hat{g}_{\hat{y}|\hat{\tau}, \tau_0}(\hat{y} | t, u) \hat{f}_{\tau|\hat{\tau}}(u | t) du + \hat{g}_{\hat{y}|\hat{\tau}, \tau_0}(\hat{y} | t, t) \hat{m}_{\tau|\hat{\tau}}(t | t)}
\]

\[
= \sum_{k=1}^{r} \hat{c}_k \ln \left( g_{\hat{y}|\hat{\tau}, \tau_0}(\hat{y} | t, k\Delta) \right) + \hat{c}_t \ln \left( g_{\hat{y}|\hat{\tau}, \tau_0}(\hat{y} | t, t) \right)
\]

where constants that depend only on \( \hat{\lambda} \) and \( \hat{\theta} \) are given by

\[
\hat{c}_k = \frac{\hat{p}_0 \hat{p}_0 e^{-\hat{t}_k} \hat{e}_2^k}{\hat{d}} \hat{g}_{\hat{y}|\hat{\tau}, \tau_0}(\hat{y} | t, k\Delta), \quad k = 1, \ldots, T
\]

\[
\hat{c}_t = \left( \frac{\hat{p}_0 \hat{p}_0 e^{-\hat{t}_0} \hat{e}_2' + \hat{p}_0 e^{-\hat{t}_0}}{\hat{d}} \right) \hat{g}_{\hat{y}|\hat{\tau}, \tau_0}(\hat{y} | t, t)
\]

To complete the proof we put \( \hat{a} = (\hat{a}_01, \hat{a}_02, \hat{a}_12)' \), \( \hat{b} = (\hat{b}_01, \hat{b}_02, \hat{b}_12)' \) and \( \hat{c} = (\hat{c}_1, \ldots, \hat{c}_r, \hat{c}_t)' \). \( \square \)
Appendix 3

Proof of Theorem 2.4

Using equations (2.37) and (2.38) of Lemma 2.4 and the formula for the likelihood function $L_S(\lambda, \theta)$ given by equation (2.39), we have

$$Q_S(\lambda, \theta | \hat{\lambda}, \hat{\theta}) = E_{\hat{\lambda}, \hat{\theta}}\left[ \ln L_S(\lambda, \theta) | S \right]$$

$$= E_{\hat{\lambda}, \hat{\theta}}\left[ \ln L_S(\lambda, \theta) | \bar{\mathbf{Y}} = \bar{\mathbf{y}}, \xi > t \right]$$

$$= \int \mathbb{E}_{\hat{\lambda}, \hat{\theta}}\left[ \ln L_S(\lambda, \theta) | \bar{\mathbf{Y}} = \bar{\mathbf{y}}, \xi > t, \tau_0 = s \right] \hat{p}_{\tau_0 | \bar{\mathbf{Y}}, \xi}(s | \bar{\mathbf{y}}, t) ds$$

$$= \int \mathbb{E}_{\hat{\lambda}, \hat{\theta}}\left[ \ln \left( g_{\bar{\mathbf{Y}}, \xi, \tau_0}(\bar{\mathbf{y}} | t, s) h(t | s) f_{\tau_0}(s) \right) \hat{g}_{\bar{\mathbf{Y}}, \xi, \tau_0}(\bar{\mathbf{y}} | t, s) \hat{h}(t | s) \hat{f}_{\tau_0}(s) ds \right]$$

$$= \int \mathbb{E}_{\hat{\lambda}, \hat{\theta}}\left[ \frac{\hat{g}_{\bar{\mathbf{Y}}, \xi, \tau_0}(\bar{\mathbf{y}} | t, u) \hat{h}(t | u) \hat{f}_{\tau_0}(u) du}{\int \hat{g}_{\bar{\mathbf{Y}}, \xi, \tau_0}(\bar{\mathbf{y}} | t, u) \hat{h}(t | u) \hat{f}_{\tau_0}(u) du} \right]$$

Since $h$ and $f_{\tau_0}$ defined in equations (2.37) and (2.38) depend only on the state parameter $\lambda = (\lambda_{01}, \lambda_{02}, \lambda_{12})$, the equation above can be decomposed into two terms,

$$Q_S(\lambda, \theta | \hat{\lambda}, \hat{\theta}) = Q_{S_{\text{state}}}^r(\lambda | \hat{\lambda}, \hat{\theta}) + Q_{S_{\text{obs}}}^r(\theta | \hat{\lambda}, \hat{\theta})$$

where the first term $Q_{S_{\text{state}}}^r$ depends only on $\lambda$ and the second term $Q_{S_{\text{obs}}}^r$ depends only on $\theta$.

Substituting equations (2.37) and (2.38), the first term $Q_{S_{\text{state}}}^r$ simplifies to

$$Q_{S_{\text{state}}}^r(\lambda | \hat{\lambda}, \hat{\theta}) = \int \mathbb{E}_{\hat{\lambda}, \hat{\theta}}\left[ \ln \left( h(t | s) f_{\tau_0}(s) \right) \hat{g}_{\bar{\mathbf{Y}}, \xi, \tau_0}(\bar{\mathbf{y}} | t, s) \hat{h}(t | s) \hat{f}_{\tau_0}(s) ds \right]$$

$$= \hat{\alpha}_{01} \lambda_{01} + \hat{\alpha}_{02} \lambda_{02} + \hat{\alpha}_{12} \lambda_{12} + \hat{\gamma}_1 \ln(\lambda_{01}) + \hat{\gamma}_2 \ln(\lambda_{01} + \lambda_{02}) + \hat{d}_2$$

where constants that depend only on fixed parameter estimates $\hat{\lambda}$ and $\hat{\theta}$ are given by
\[ \hat{a}_{01} = \hat{a}_{02} = -\frac{\hat{\lambda}_{01}(\hat{\lambda}_{01} + \hat{\lambda}_{02})e^{-i\gamma t}(\hat{e}_3, \hat{g}) + \hat{g}_{\mathcal{V}^{[6]r_0}}(\mathbf{y} | t, t)e^{-i\gamma t}(2t + \hat{r}_1(\hat{\lambda}_{01} + \hat{\lambda}_{02}) + 2(\hat{\lambda}_{01} + \hat{\lambda}_{02})^{-1})}{\hat{\delta}} \]

\[ \hat{a}_{12} = \frac{\hat{\lambda}_{01}(\hat{\lambda}_{01} + \hat{\lambda}_{02})e^{-i\gamma t}}{\hat{\delta}}(\langle \hat{e}_3, \hat{g} \rangle - t \langle \hat{e}_2, \hat{g} \rangle) \]

\[ \hat{\gamma} = \frac{e^{-i\gamma t}(1 + (\hat{\lambda}_{01} + \hat{\lambda}_{02})t)}{\hat{\delta}} \hat{g}_{\mathcal{V}^{[6]r_0}}(\mathbf{y} | t, t) + 1 \]

\[ \hat{\delta} = \hat{\lambda}_{01}(\hat{\lambda}_{01} + \hat{\lambda}_{02})e^{-i\gamma t} \int_{x,t} (se^{i\theta_k - i\gamma t}) \ln s \hat{g}_{\mathcal{V}^{[6]r_0}}(\mathbf{y} | t, s)ds + (\hat{\lambda}_{01} + \hat{\lambda}_{02})^2 \hat{g}_{\mathcal{V}^{[6]r_0}}(\mathbf{y} | t, t) \int_{x,t} se^{-i\gamma t} \ln s ds \]

\[ \hat{d}_2(\hat{\lambda}, \hat{\theta}) = \frac{\hat{\lambda}_{01}(\hat{\lambda}_{01} + \hat{\lambda}_{02})e^{-i\gamma t} \int_{x,t} (se^{i\theta_k - i\gamma t}) \ln s \hat{g}_{\mathcal{V}^{[6]r_0}}(\mathbf{y} | t, s)ds + (\hat{\lambda}_{01} + \hat{\lambda}_{02})^2 \hat{g}_{\mathcal{V}^{[6]r_0}}(\mathbf{y} | t, t) \int_{x,t} se^{-i\gamma t} \ln s ds}{\hat{\delta}} \]

and vectors \( \hat{e}_1, \hat{e}_2 \) and \( \hat{e}_3 \) are defined in Appendix 2. Similarly, the second term \( Q_S^{obs} \), which is a function only of the residual observation parameter \( \theta \), simplifies to

\[ Q_S^{obs}(\theta | \hat{\lambda}, \hat{\theta}) = \left[ \ln \left( \hat{g}_{\mathcal{V}^{[6]r_0}}(\mathbf{y} | t, s) \hat{g}_{\mathcal{V}^{[6]r_0}}(\mathbf{y} | t, s) \hat{h}(t | s) \hat{f}_{r_0}(s) ds \right) \right] \int \hat{g}_{\mathcal{V}^{[6]r_0}}(\mathbf{y} | t, u) \hat{h}(t | u) \hat{f}_{r_0}(u) du \]

\[ = \sum_{k=1}^{\hat{r}} \hat{\beta}_k \ln \left( \hat{g}_{\mathcal{V}^{[6]r_0}}(\mathbf{y} | t, k\Delta) \right) + \hat{\beta}_1 \ln \left( \hat{g}_{\mathcal{V}^{[6]r_0}}(\mathbf{y} | t, t) \right) \]

where constants that depend only on \( \hat{\lambda} \) and \( \hat{\theta} \) are given by

\[ \hat{\beta}_k = \frac{\hat{\lambda}_{01}(\hat{\lambda}_{01} + \hat{\lambda}_{02})e^{-i\gamma t} \hat{e}_2^k \hat{g}_{\mathcal{V}^{[6]r_0}}(\mathbf{y} | t, k\Delta)}{\hat{\delta}}, \quad k = 1, \ldots, T \]

\[ \hat{\beta}_1 = \frac{e^{-i\gamma t}(1 + (\hat{\lambda}_{01} + \hat{\lambda}_{02})t)}{\hat{\delta}} \hat{g}_{\mathcal{V}^{[6]r_0}}(\mathbf{y} | t, t) \]

To complete the proof we put \( \hat{a} = (\hat{a}_{01}, \hat{a}_{02}, \hat{a}_{12})' \) and \( \hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_T)' \). \( \square \)
Appendix 4

Derivation of the Probability Transition Matrix in (4.2) using Laplace Transform

The Kolmogorov backward differential equations are given by

\[
\begin{align*}
P_{00}(t) &= \lambda_{01} P_{10}(t) - (\lambda_{01} + \lambda_{02}) P_{00}(t) \\
P_{01}(t) &= \lambda_{01} P_{11}(t) - (\lambda_{01} + \lambda_{02}) P_{01}(t) \\
P_{02}(t) &= \lambda_{01} P_{12}(t) + \lambda_{02} P_{22}(t) - (\lambda_{01} + \lambda_{02}) P_{02}(t) \\
P_{10}(t) &= \lambda_{10} P_{00}(t) - (\lambda_{10} + \lambda_{12}) P_{10}(t) \\
P_{11}(t) &= \lambda_{10} P_{01}(t) - (\lambda_{10} + \lambda_{12}) P_{11}(t) \\
P_{12}(t) &= \lambda_{10} P_{02}(t) + \lambda_{12} P_{22}(t) - (\lambda_{10} + \lambda_{12}) P_{12}(t)
\end{align*}
\]

Taking Laplace transform of the above equations gives

\[
\begin{align*}
s \tilde{P}_{00}(s) - 1 &= \lambda_{01} \tilde{P}_{10}(s) - (\lambda_{01} + \lambda_{02}) \tilde{P}_{00}(s) \\
s \tilde{P}_{01}(s) &= \lambda_{01} \tilde{P}_{11}(s) - (\lambda_{01} + \lambda_{02}) \tilde{P}_{01}(s) \\
s \tilde{P}_{02}(s) &= \lambda_{01} \tilde{P}_{12}(s) + \frac{\lambda_{02}}{s} - (\lambda_{01} + \lambda_{02}) \tilde{P}_{02}(s) \\
s \tilde{P}_{10}(s) &= \lambda_{10} \tilde{P}_{00}(s) - (\lambda_{10} + \lambda_{12}) \tilde{P}_{10}(s) \\
s \tilde{P}_{11}(s) - 1 &= \lambda_{10} \tilde{P}_{01}(s) - (\lambda_{10} + \lambda_{12}) \tilde{P}_{11}(s) \\
s \tilde{P}_{12}(s) &= \lambda_{10} \tilde{P}_{02}(s) + \frac{\lambda_{12}}{s} - (\lambda_{10} + \lambda_{12}) \tilde{P}_{12}(s)
\end{align*}
\]

where \( \tilde{P}_y(s) := \int_0^\infty e^{-st} P_y(t) dt \) denotes the Laplace transform of \( P_y(t) \). Solving for \( \tilde{P}_y(s) \) gives

\[
\begin{align*}
\tilde{P}_{00}(s) &= \frac{s + \lambda_{10} + \lambda_{12}}{(s + \alpha)(s + \beta)}, & \tilde{P}_{01}(s) &= \frac{\lambda_{01}}{(s + \alpha)(s + \beta)} \\
\tilde{P}_{10}(s) &= \frac{\lambda_{10}}{(s + \alpha)(s + \beta)}, & \tilde{P}_{11}(s) &= \frac{s + \lambda_{01} + \lambda_{02}}{(s + \alpha)(s + \beta)}
\end{align*}
\]

where
\[\alpha = \frac{(\lambda_{01} + \lambda_{02} + \lambda_{10} + \lambda_{12}) - \kappa^{1/2}}{2}\]

\[\beta = \frac{(\lambda_{01} + \lambda_{02} + \lambda_{10} + \lambda_{12}) + \kappa^{1/2}}{2}\]

\[\kappa = \lambda_{01}^2 + \lambda_{02}^2 + \lambda_{10}^2 + \lambda_{12}^2 + 2\lambda_{01}\lambda_{02} + 2\lambda_{10}\lambda_{12} + 2\lambda_{10}\lambda_{01} - 2\lambda_{02}\lambda_{10} - 2\lambda_{01}\lambda_{12} - 2\lambda_{02}\lambda_{12}\]

Taking the inverse Laplace transform of above equations gives

\[P_{00}(t) = e^{-\beta t} + \frac{\lambda_{10} + \lambda_{12} - \alpha}{\beta - \alpha} (e^{-\alpha t} - e^{-\beta t})\]

\[P_{01}(t) = \frac{\lambda_{01}}{\beta - \alpha} (e^{-\alpha t} - e^{-\beta t})\]

\[P_{10}(t) = \frac{\lambda_{10}}{\beta - \alpha} (e^{-\alpha t} - e^{-\beta t})\]

\[P_{11}(t) = e^{-\beta t} + \frac{\lambda_{01} + \lambda_{02} - \alpha}{\beta - \alpha} (e^{-\alpha t} - e^{-\beta t})\]

Finally, using the fact that \(\sum_{j=0}^{\infty} P_{ij}(t) = 1, i = 0, 1\), the probability transition matrix is explicitly given by (4.3).
Appendix 5

Provost and Rudiuk (1996) showed that any indefinite quadratic form in normal vectors $Q = U^T A U$, where $U \sim N_d(\mu, \Sigma)$, can be expressed as the difference of two linear combinations of independent non-central Chi-square variables. In particular, let $\Sigma = L L^T$, and $\lambda_1, \ldots, \lambda_d$ denote the eigenvalues of $L^T A L$. Then

$$Q = \sum_{i=1}^{d} \lambda_i \chi_i^2(\delta_i^2) = W - V$$

where $W = \sum_{i=1}^{p} \lambda_i \chi_i^2(\delta_i^2)$ and $V = \sum_{i=p+1}^{k} (-\lambda_i) \chi_i^2(\delta_i^2)$, and eigenvalues $\lambda_i > 0$ for $i = 1, \ldots, p$, $\lambda_i < 0$ for $i = p+1, \ldots, k$, and $\lambda_i = 0$ for $i = k+1, \ldots, d$. The random variables $\chi_i^2(\delta_i^2)$ have i.i.d. Chi-square distribution with one degree of freedom and non-centrality parameter $\delta_i^2$. Using this fact, the authors were able to derive an explicit formula for the exact distribution function of $Q = U^T A U$, which is stated below.

**Theorem 3.1 of Provost and Rudiuk (1996)**

Let $Q = W - V = \sum_{j=t}^{l} l_j T_j - \sum_{j=t+1}^{l+w} l_j T_j$, where $l_j$ are positive real numbers and $T_j$ are independent non-central Chi-square variables with $\alpha_j$ degrees of freedom and non-centrality parameter $d_j$, $j = 1, \ldots, t+w$. Let $\alpha = \sum_{j=1}^{l} \alpha_j / 2$, $\alpha' = \sum_{j=t+1}^{l+w} \alpha_j / 2$ and $b = (\beta^{-1} + \beta'^{-1} / 2$, where $\beta$ and $\beta'$ satisfy that $|1 - \beta / l_j| < 1$, $j = 1, \ldots, t$, and $|1 - \beta' / l_j| < 1$, $j = t+1, \ldots, t+w$. Then provided $\alpha$ and $\alpha'$ are not both nonnegative integers plus $1/2$, for $q \leq 0$, the distribution function of $Q$ is given by
F(q) = \sum_{k=0}^{\infty} \sum_{v=0}^{\infty} \sum_{l=0}^{\infty} \frac{\theta_k \theta'_v}{\Gamma(\alpha' + \nu)} b^{-\nu/2} \left( \frac{\Gamma(-2\mu), \mu - l + 1/2, b^{\nu+l+2+i}}{\Gamma(1/2 - \mu - l) \Gamma(1/2 + \mu - l)} \right) I_{1v} \\
+ \frac{\Gamma(2\mu), -\mu - l + 1/2, b^{\nu+l+2+i}}{\Gamma(1/2 - \mu - l) \Gamma(1/2 + \mu - l)} I_{2v}

and for \( q > 0 \), the distribution function of \( Q \) is given by

\[
F(q) = F(0) + \sum_{k=0}^{\infty} \sum_{v=0}^{\infty} \sum_{l=0}^{\infty} \frac{\theta_k \theta'_v}{\Gamma(\alpha' + \nu)} b^{-\nu/2} \left( \frac{\Gamma(-2\mu), \mu - l + 1/2, b^{\nu+l+2+i}}{\Gamma(1/2 - \mu - l) \Gamma(1/2 + \mu - l)} \right) I_{3v} \\
+ \frac{\Gamma(2\mu), -\mu - l + 1/2, b^{\nu+l+2+i}}{\Gamma(1/2 - \mu - l) \Gamma(1/2 + \mu - l)} I_{4v}
\]

where constants are given by

\[
a_k = (2k)^{\nu/2} \sum_{r=0}^{k-1} (k - r) \beta \sum_{j=1}^{r} (d_j/l_j)(1 - \beta/l_j)^{k-r} + \sum_{j=1}^{r} \alpha_j(1 - \beta/l_j)^{k-r} a_r
\]

\[
a'_v = (2v)^{\nu/2} \sum_{r=0}^{v-1} (v - r) \beta' \sum_{j=1}^{r} (d_j/l_j)(1 - \beta'/l_j)^{v-r} + \sum_{j=1}^{r} \alpha_j(1 - \beta'/l_j)^{v-r} a'_r
\]

\[
a_0 = e^{-\sum_{j=1}^{t} d_j/l_j} \prod_{j=1}^{t} (\beta/l_j)^{\alpha_j/2}
\]

\[
a'_0 = e^{-\sum_{j=1}^{t+w} d_j/l_j} \prod_{j=1}^{t+w} (\beta'/l_j)^{\alpha'/2}
\]

\[
\theta_k = a_k/(2\beta)^{\alpha+k}, \quad \theta'_v = a'_v/(2\beta')^{\alpha'+v}
\]

l = -(\alpha + k - \alpha' - \nu)/2

l_2 = (\alpha + k - \alpha' - \nu)/2

\[
\mu = (1 - \alpha - k - \alpha' - \nu)/2
\]

\[
I_{1v} = \Gamma(i+1, -q/(2\beta')) (2\beta')^{i+1}
\]

\[
I_{2v} = \Gamma(-2\mu + i + 1, -q/(2\beta')) (2\beta')^{-2\mu+i+1}
\]

\[
I_{3v} = \gamma(i+1, q/(2\beta')) (2\beta')^{i+1}
\]

\[
I_{4v} = \gamma(-2\mu + i + 1, q/(2\beta')) (2\beta')^{-2\mu+i+1}
\]

\[
\gamma(a, x) = \int_0^x e^{-t} t^{a-1} dt
\]

\[
\Gamma(a) = \int_0^\infty e^{-t} t^{a-1} dt
\]

113